Report on Artificial Intelligence Models for Sensitive Content Detection in European School Environments

Contenido

1. Introduction 9
   1.1 Context and Relevance 9
   1.2 Research Objectives 9
   1.3 Methodology 10
2. European School Environment and Applicable Regulatory Framework 11
   2.1 Data Protection in the School Context 11
   2.2 General Data Protection Regulation (GDPR) 11
   2.3 European Union Artificial Intelligence Act (EU AI Act) 12
   2.4 Other Regulatory and Ethical Considerations 13
   2.5 Practical Implications for the Project 13
3. Problem Statement 14
   3.1 Needs for Sensitive Content Detection in European School Environments 14
   3.2 Data Volume and Processing Requirements 14
   3.3 Technical and Cultural Complexity 15
   3.4 Multilingualism Challenges in Sensitive Content Detection 15
   3.5 Problem Summary 16
4. Analysis of Advanced AI Model Capabilities for Sensitive Content Detection 17
   4.1. Limitations of Conventional Computer Vision Toolkit-Based Approaches 17
   4.2. Advanced Requirements for Contextual, Ethical, and Efficient Detection 18
5. Analysis of Proprietary AI Models for Image Recognition 19
   5.1 Analysis of the Google Cloud Vision API Model 19
      5.1.1. Model Identification 19
      5.1.2. General Model Description 19
      5.1.3. Technical Capabilities 20
      5.1.4. Licensing and Terms of Use 21

5.1.5. Legal and Ethical Analysis 21
5.1.6. Infrastructure Requirements 22
5.1.7. Advantages and Limitations 23
5.1.8. Conclusions and Recommendations 23
5.1.10. Visual Summary 24
5.2 Analysis of the Amazon Rekognition Model 24
5.2.1. Model Identification 24
5.2.2. General Model Description 24
5.2.3. Technical Capabilities 26
5.2.4. Licensing and Terms of Use 27
5.2.5. Legal and Ethical Analysis 28
5.2.6. Infrastructure Requirements 29
5.2.7. Advantages and Limitations 30
5.2.8. Conclusions and Recommendations 31
5.2.9. Visual Summary 31
6. Analysis of Open Source AI Models for Image Recognition 32
6.1 Analysis of the GluonCV Model 32
6.1.1. Model Identification 32
6.1.2. General Model Description 32
6.1.3. Technical Capabilities 34
6.1.4. Licensing and Terms of Use 36
6.1.5. Legal and Ethical Analysis 37
6.1.6. Infrastructure Requirements 41
6.1.7. Advantages and Limitations 43
6.1.8. Conclusions and Recommendations 44
6.1.9. Visual Summary 46
6.2 Analysis of the RTMDet Model 46
6.2.1. Model Identification 46
6.2.2. General Model Description 47
6.2.3. Technical Capabilities 49
6.2.4. Licensing and Terms of Use 52
6.2.5. Legal and Ethical Analysis 54

6.2.6. Infrastructure Requirements 58
6.2.7. Advantages and Limitations 61
6.2.8. Conclusions and Recommendations 62
6.2.9. Visual Summary 63
6.3 Combinations of Open Source AI Models for Image Recognition 64
6.3.1 Introduction: Why Combine Open Source Models? 64
6.3.2 Common Applicable Combination Strategies 66
6.3.3 Key Practical Considerations for Implementing Combinations 68
6.4 Conclusions and Recommendations on Open Source Models 71
6.4.1 Synthesis of Open Source Model Evaluation 71
6.4.2 Strategic Recommendations for the Project 74
7. Analysis of Proprietary AI Models for Text Recognition 76
7.1 Analysis of the Google Cloud Natural Language API Model 76
7.1.1. Model Identification 76
7.1.2. General Model Description 76
7.1.3. Technical Capabilities 79
7.1.4. Licensing and Terms of Use 83
7.1.5. Legal and Ethical Analysis 84
7.1.6. Infrastructure Requirements 87
7.1.7. Advantages and Limitations 88
7.1.8. Conclusions and Recommendations 89
7.1.9. Visual Summary 90
7.2 Analysis of the Amazon Comprehend Model 91
7.2.1. Model Identification 91
7.2.2. General Model Description 91
7.2.3. Technical Capabilities 94
7.2.4. Licensing and Terms of Use 98
7.2.5. Legal and Ethical Analysis 99
7.2.6. Infrastructure Requirements 102
7.2.7. Advantages and Limitations 103
7.2.8. Conclusions and Recommendations 104
7.2.10. Visual Summary 106

7.3 Analysis of IBM Watson Natural Language Understanding 106
7.3.1. Model Identification 106
7.3.2. General Model Description 106
7.3.3. Technical Capabilities 108
7.3.4. Licensing and Terms of Use 111
7.3.5. Legal and Ethical Analysis 112
7.3.6. Infrastructure Requirements 115
7.3.7. Advantages and Limitations 116
7.3.8. Conclusions and Recommendations 117
7.3.9. Visual Summary 118
8. Analysis of Open Source AI Models for Text Recognition 118
8.1 Analysis of the Multilingual Toxic-BERT Model (Fine-tuned variants from XLM-RoBERTa Base/Large) 118
8.1.1. Model Identification 118
8.1.2. General Model Description 118
8.1.3. Technical Capabilities 120
8.1.4. Licensing and Terms of Use 122
8.1.5. Legal and Ethical Analysis 123
8.1.6. Infrastructure Requirements 124
8.1.7. Advantages and Limitations 125
8.1.8. Conclusions and Recommendations 127
8.1.9. Visual Summary 128
9. Analysis of Proprietary Multimodal AI Models 128
9.1 Analysis of the Microsoft Azure Content Safety Model 128
9.1.1. Model Identification 128
9.1.2. General Model Description 128
9.1.3. Technical Capabilities 131
9.1.4. Licensing and Terms of Use 133
9.1.5. Legal and Ethical Analysis 134
9.1.6. Infrastructure Requirements 136
9.1.7. Advantages and Limitations 137
9.1.8. Conclusions and Recommendations 138

9.1.9. Visual Summary 140
10. Case Studies and Pilot Projects of AI Implementation in Educational Institutions 140
10.1 Introduction 140
10.2. Government Strategies and National AI Initiatives in Education: A Global Mosaic 141
10.2.1 Analysis of governmental approaches and policy frameworks in different regions 141
10.2.2 Asia-Pacific 142
10.2.3 North America 142
10.2.4 Europe 143
10.2.5 Latin America and the Caribbean 143
10.2.6 Africa 144
10.2.3 Fostering pilot projects and public-private collaboration 144
10.3. Case Studies and Pilot Projects Highlighted by Region 145
10.3.1 North America (USA, Canada) 145
10.3.2 Europe (EU and individual countries) 147
10.3.3 Asia-Pacific (China, South Korea, Singapore, Australia, India, Pakistan, Japan, Indonesia) 149
10.3.4 Latin America and the Caribbean (Regional and national initiatives) 152
10.3.5 Africa (Regional and national initiatives) 154
10.4. Specific AI Applications for Sensitive Content Detection in School Environments 156
10.4.1 Contextualization: The growing need for safe learning environments 156
10.4.2 Detection of School Bullying and Cyberbullying 157
10.4.3 Violence Prevention and School Safety Improvement 158
10.4.4 Identification of Risks to Student Well-being (Self-Harm, Suicide, Eating Disorders, Addictions) 158
10.4.5 Detection of Content Related to Radicalization and Extremism 159
10.4.6 Cybersecurity and Protection against Digital Threats 159
10.4.7 Plagiarism Detection and Promotion of Academic Integrity 160
11. Psychological and Pedagogical Impact of Sensitive Content Detection on Students and Educators 160

11.1 Introduction 160
11.2 Psychological Impact on Students 161
11.2.1 Anxiety and Stress from the Perception of Continuous Surveillance 162
11.2.2 "Chilling Effect" on Expression and Exploration 162
11.2.3 Impact on Trust, Relationship with the Institution, and Student Perception 163
11.2.4 False Positives and Their Emotional and Academic Consequences. 164
11.2.5 Risk of Stigmatization, Labeling, and Algorithmic Bias 165
11.2.6 Impact on Identity Development and Student Digital Autonomy..... 165
11.3 Pedagogical Impact and Impact on Educators 166
11.3.1 Additional Workload and "Digital Overload" for Educators 166
11.3.3 Displacement of the Educator's Role and Pedagogical Agency 168
11.3.4 Ethical Dilemmas and Decision-Making in Teaching Practice 168
11.3.5 Impact on Classroom Dynamics and Teacher-Student Relationship 169
11.4 Fostering a Balanced, Ethical, and Supportive Approach: Strategies and Recommendations 169
11.4.1 Strict Adherence to the European Regulatory Framework: The EU AI Act and GDPR 169
11.4.2 Radical Transparency and Proactive Communication with the Educational Community 171
11.4.3 Student Participation and Co-design of Digital Safety Policies 171
11.4.4 Qualified and Indispensable Human Oversight at All Stages 172
11.4.5 Prioritization of Preventive Education: Digital Citizenship and Critical Media Literacy 172
11.4.6 Establishment of Clear, Fair, and Robust Appeal Processes 172
11.4.7 Privacy Protection and Data Governance by Design and by Default . 173
11.4.8 Continuous Evaluation, Research, and Adaptation of Strategies 173
11.4.9 Fostering Explainability (XAI) and Auditability of AI Systems 173
11.5 Towards a Responsible and Human-Centered Implementation 176
12. Practical Implementation Strategies and Change Management 177
12.1 Introduction 177
12.2 Phase 1: Strategic Planning and Specific Needs Assessment 179

12.2.1 Clear Definition of Objectives, Scope, and Types of Sensitive Content to Detect 180
12.2.2 Analysis of the Specific School Context and Institutional Preparation 181
12.2.3 Exhaustive Risk Assessment and Regulatory Compliance (DPIA/EIPD) 182
12.2.4 Stakeholder Engagement and Co-design..... 186
12.3 Phase 2: Selecting the Right Technological Solution 187
12.3.1 Establishing Detailed Selection Criteria 188
12.3.2 Evaluation of Alternatives: Proprietary vs. Open Source vs. Hybrid .. 190
12.3.3 Proofs of Concept (PoC) and Thorough Vendor Evaluation 192
12.4 Phase 3: Detailed Implementation Design and Data Governance 193
12.4.1 Technical Implementation Plan and System Architecture 194
12.4.2 Development of Clear and Detailed Policies and Protocols 196
12.4.3 Designing the Workflow with Essential Human Oversight 198
12.5 Phase 4: Change Management, Strategic Communication, and Training 199
12.5.2 Comprehensive Training and Continuous Professional Development 201
12.5.3 Proactively Addressing Resistances and Concerns 203
12.6 Phase 5: Pilot Implementation and Gradual Rollout..... 204
12.6.1 Critical Importance of Pilot Testing 204
12.6.2 Systematic Feedback Collection and Iterative Adjustments..... 205
12.6.3 Progressive and Planned Scaling 206
12.7 Phase 6: Monitoring, Continuous Evaluation, Maintenance, and Adaptation 207
12.7.2 Periodic Comprehensive Impact Evaluation 209
12.7.3 Technical Maintenance, Model Updates, and Continuous Security . 210
12.7.4 Continuous Adaptation to a Changing Environment 211
12.8 Conclusion: Towards Responsible and Sustainable AI Implementation for School Safety 212
13. General Conclusions and Final Recommendations 214
13.1 Recapitulation of Challenges and Opportunities 214

13.2 Synthesis of AI Model Evaluation 215
13.3 Ethical, Legal, and Pedagogical Imperatives 215
13.4 Towards Responsible Implementation: Key Recommendations 216
13.5 Future Perspectives and Final Considerations 217

1. Introduction

1.1 Context and Relevance

The current integration of digital platforms and internet access in European educational institutions has profoundly transformed teaching and learning methods. Students now interact with a variety of online content, which has expanded their educational opportunities but also introduced new risks associated with exposure to inappropriate or dangerous material. These risks include, among others, violence, explicit sexuality, cyberbullying, drug use, and radicalization.

The educational environment faces the challenge of offering a rich learning experience while simultaneously ensuring students are protected from content that could compromise their well-being and emotional development. This balance becomes more complex with the increasing reliance on technological tools, making the creation of safe and regulated digital environments imperative.

Artificial intelligence (AI) presents itself as a promising tool to address this issue, providing effective solutions for the automatic detection of sensitive content in real time. However, implementing AI models in schools is not without its challenges, such as concerns about data privacy, algorithmic accuracy, algorithmic biases, and the legal and ethical complexities arising from the use of these technologies in the educational sphere. Furthermore, AI solutions must align with current data protection regulations in Europe, such as the General Data Protection Regulation (GDPR), and with regulations protecting minors in school environments.

1.2 Research Objectives

The primary purpose of this research is to provide a comprehensive analysis of the most suitable artificial intelligence models for detecting sensitive content in images and texts within the European school context. The specific objectives of the research include:

Identify and evaluate the main AI models that can be used for the recognition and classification of sensitive content in images and texts, including those requiring fine-tuning and those that can be used directly.

Analyze the legal and ethical framework applicable to the implementation of AI technologies in school environments, ensuring that proposed solutions comply with European legislation, such as the GDPR and specific regulations for the protection of minors.

Establish a clear categorization of the types of sensitive content that need to be detected in school environments, covering topics such as violence, bullying, drug consumption, radicalization, and other harmful content.

Compare the benefits and limitations of proprietary models versus open-source solutions, based on their costs, licenses, and capabilities.

Propose practical recommendations for integrating these technologies into commercial products intended for use in the educational field, with an emphasis on implementation feasibility and necessary infrastructure requirements.

1.3 Methodology

The research will be conducted using a combination of qualitative and quantitative methods, which include:

Literature Review: An exhaustive review of technical documentation from AI providers, academic studies, and current regulations will be carried out to gain a comprehensive understanding of available AI models and their applicability in the educational context.

AI Model Evaluation: The most relevant artificial intelligence models, both proprietary and open source, will be analyzed to detect and classify sensitive content in images and texts, evaluating their performance, accuracy, and applicability in school environments.

Infrastructure Study: The technological infrastructure required to process between 2000 and 3000 images daily per school will be investigated, considering the option of dedicated servers versus global servers and the possibility of cloud or local implementation in schools.

Specialized Document Review: Documentation and reference frameworks in ethics, law, pedagogy, and technology have been consulted to ensure that the proposed solutions are viable, secure, and respectful of students' rights.

Comparative Analysis: Finally, a comparative analysis of the evaluated AI models will be performed, considering aspects such as costs, licensing terms, advantages, disadvantages, and the adaptability of the models to the specific needs of each school.

The resulting report will provide a clear guide for selecting the most appropriate AI model for developing technological solutions that protect students in educational environments, while promoting a learning experience free from risks related to harmful content.

2. European School Environment and Applicable Regulatory Framework

2.1 Data Protection in the School Context

The use of artificial intelligence (AI) technologies in European school environments must operate within a strict legal framework that protects students' fundamental rights, especially the right to privacy and the protection of personal data. This protection is even more rigorous when it concerns minors, who are considered a particularly vulnerable category under European law.

Two fundamental pillars form the legal framework governing the implementation of AI systems for sensitive content detection in schools: the General Data Protection Regulation (GDPR) and the European Union Artificial Intelligence Act (EU AI Act). Their implications are explained in detail below.

2.2 General Data Protection Regulation (GDPR)

The GDPR (Regulation (EU) 2016/679) is the primary regulation governing the protection of personal data in Europe. In the school context, its compliance is mandatory and of crucial importance, given that minor students are considered especially protected "data subjects."

Some fundamental principles of the GDPR that directly impact the use of AI in schools are:

- Lawfulness, fairness, and transparency: Data processing must be based on a valid legal basis (such as explicit consent from parents or guardians) and must be transparent; that is, students and their families must be clearly informed about how and why their data is collected and used.
- Purpose limitation: Personal data can only be collected for specified, explicit, and legitimate purposes and cannot be further processed in a manner incompatible with those purposes.
- Data minimization: Only data strictly necessary to fulfill the defined purpose should be collected. There cannot be a massive and indiscriminate collection of data.
- Storage limitation: Data should not be stored for longer than necessary to fulfill the purpose for which it was collected.
- Integrity and confidentiality: Appropriate technical and organizational measures must be implemented to ensure the security of personal data, including protection against unauthorized access or unlawful processing.

- Data subject rights: Students and their guardians have rights that must be respected, such as the right to access, rectification, erasure ("right to be forgotten"), and objection to the processing of their data.

In particular, the use of AI to analyze content created or consumed by minors requires additional protection measures, such as Data Protection Impact Assessments (DPIAs) and strict control over access to processed information.

2.3 European Union Artificial Intelligence Act (EU AI Act)

The EU AI Act is the world's first specific regulatory framework for artificial intelligence. Although still in the process of full implementation in some Member States, this act introduces a risk-based approach to regulating AI systems, classifying them into four categories:

- Unacceptable risk (prohibited)
- High risk (heavily regulated)
- Limited risk (transparency requirements)
- Minimal risk (freedom of use)

AI systems implemented in school environments, particularly those affecting access to education, student assessment, or the safety of minors, are classified as "high-risk" systems. This has several implications:

- Prior assessment obligations: Before deployment, high-risk AI systems must undergo a thorough assessment to ensure they comply with all legal requirements for safety, fairness, reliability, and transparency.
- Documentation and record-keeping: Detailed documentation of the AI system's design, operation, and purpose must be maintained to allow for audits and compliance evaluations by regulatory authorities.
- Transparency requirements: Schools and provider companies must clearly inform users (in this case, students and parents) about the use of AI, explaining in an understandable way how the system works and what decisions it can make or influence.
- Mandatory human oversight: The system cannot operate entirely autonomously. There must be active human oversight that can intervene, correct, or stop the system in case of error or undesirable behavior.
- Risk and bias mitigation: Measures must be implemented to prevent algorithmic biases, which could unfairly discriminate against students based on their origin, gender, disability, or other protected characteristics.

Furthermore, any serious incident related to the operation of a high-risk AI system must be reported to the competent authorities.

2.4 Other Regulatory and Ethical Considerations

In addition to the GDPR and the AI Act, other regulations and ethical principles are relevant to the use of AI in school environments:

- Convention on the Rights of the Child: Recognizes minors' right to privacy protection and to the development of their personality in a safe environment.
- Digital Rights Charters (such as the European initiative for a Digital Rights Declaration): Promote safe and equitable access to digital technologies, especially for minors.
- National regulations: Each European country may have additional data protection laws or specific educational regulations that reinforce or complement European standards.

From an ethical point of view, it is fundamental to ensure that AI solutions:

- Respect the dignity and rights of students.
- Are transparent and understandable for all stakeholders.
- Do not perpetuate stereotypes or discrimination.
- Maintain an appropriate balance between safety, privacy, and access to education.

2.5 Practical Implications for the Project

For an AI-based solution to be implemented in European schools legally and ethically, it must:

- Ensure strict compliance with GDPR and the AI Act.
- Secure the collection of valid consents from minors' legal guardians.
- Conduct Data Protection Impact Assessments (DPIAs) before implementation.
- Design systems that allow for effective human oversight.
- Document all decision-making, training, and operational processes of the AI models.
- Establish clear information security policies and incident response.

Complying with this legal and ethical framework is not only an obligation but also strengthens the trust of families, teachers, and educational authorities in using new technologies to improve students' school experience safely and responsibly.

3. Problem Statement

3.1 Needs for Sensitive Content Detection in European School Environments

Educational institutions today face the challenge of maintaining safe and respectful digital environments for their students. As the use of digital platforms expands in schools, so does the exposure to risks arising from the circulation of inappropriate or harmful content.

Among the main categories of sensitive content requiring attention are:

- Violence and bullying: Images, texts, or symbols that promote physical, psychological, or verbal violence, including threats and bullying.
- Inappropriate sexual content: Material of a sexual nature that is not suitable for minors, including explicit or suggestive content.
- References to drugs and alcohol: Promotion, consumption, or trivialization of illicit or addictive substances.
- Images of weapons: Content related to firearms, bladed weapons, or other dangerous instruments.
- Hate speech, homophobia, and racism: Expressions of discrimination, exclusion, or incitement to hatred based on race, religion, gender, sexual orientation, or other personal characteristics.
- Content related to eating disorders: Material that encourages dangerous practices such as anorexia, bulimia, or unhealthy extreme diets.
- Self-harm and suicide material: Content that incites or normalizes self-harming or suicidal behaviors.
- Radicalization and extremism: Messages or images that promote political, religious, or ideological violence.

Each of these categories requires analysis systems capable of identifying not only keywords or explicit patterns but also contextual nuances and cultural variations specific to European diversity. Furthermore, they must distinguish between genuinely problematic content and legitimate cases of academic or artistic discussion to avoid undue censorship.

3.2 Data Volume and Processing Requirements

The need to process between 2,000 and 3,000 images daily per school adds a significant technical dimension to the problem. It's not merely about detecting sensitive content but doing so at a considerable scale and under conditions that allow for a rapid and effective response.

The main requirements stemming from this processing volume include:

- Real-time or near real-time processing capability: Systems must be able to analyze large quantities of images quickly, allowing for intervention before students are exposed to harmful content.
- Scalability: Solutions need to adapt to variations in workload, especially during school hours when the most digital activity occurs.
- Data management and privacy: Data must be stored and managed with strict privacy controls, in compliance with GDPR and other applicable regulations.
- Reduced latency: For alerts about problematic content to be effective, processing time must be short enough to allow for timely intervention by school officials.

This scenario implies that the technological solution must not only be accurate in detection but also efficient, secure, and scalable.

3.3 Technical and Cultural Complexity

Another significant challenge lies in the complexity of content interpretation. Not all instances of violence, sexuality, or controversial discourse should necessarily be censored in an educational setting; some may have pedagogical value when discussed in appropriate contexts (e.g., in history, social studies, or health education classes).

Therefore, AI systems must:

- Be capable of interpreting the context in which content appears.
- Adapt to cultural differences and local regulations within the diverse European environment.
- Minimize both false positives (legitimate content flagged as inappropriate) and false negatives (harmful content that goes undetected).

This level of sophistication adds complexity to the design, training, and fine-tuning of the AI models selected for the project.

3.4 Multilingualism Challenges in Sensitive Content Detection

The European educational environment is intrinsically multilingual. Schools may use different official languages depending on the country, region, and even the type of educational program (e.g., bilingual or international programs). Furthermore, within the same institution, students speaking different languages may coexist in their digital interactions.

This introduces several specific challenges for sensitive content detection systems:

- Accurate detection in multiple languages: AI models must be capable of analyzing texts in various European languages such as English, French, German, Italian, Spanish, Dutch, Polish, among others, without losing accuracy in sensitive content detection.
- Analysis of slang and local expressions: Often, problematic language doesn't use formal terms but rather youth slang, abbreviations, or colloquialisms that vary from one language to another and from one region to another.
- Multilingual contextual interpretation: The meaning of certain words or phrases can depend on the cultural or linguistic context. A term might be neutral in one language and offensive in another.
- Consistency in moderation: The system must apply homogeneous detection and moderation criteria, regardless of the language used, to ensure equitable protection for all students.

Therefore, the selected AI models must have:

- Native multilingual capabilities or be adapted through specific training techniques.
- Databases of examples in different European languages and dialects.
- Strategies for constant updates to include new emerging expressions among young people.

Effective management of multilingualism is, therefore, an essential requirement for the success of any sensitive content detection solution in the European school context.

3.5 Problem Summary

In summary, the problem posed is:

- Accurately and rapidly detect multiple types of sensitive content in images and texts generated or shared in European school environments.
- Comply with strict legal regulations (especially GDPR and the EU AI Act) for the protection of minors' data.
- Manage large daily data volumes for each school, ensuring efficiency, privacy, and responsiveness.
- Reduce classification errors through AI models that understand contextual and cultural nuances.

- Effectively manage multilingual analysis, ensuring reliable detection of sensitive content in the different languages and slang used in European school environments.

The solution must, therefore, be technically and legally robust, but also respectful of the educational environment, adapting to the specific realities and needs of European schools.

4. Analysis of Advanced AI Model Capabilities for Sensitive Content Detection

4.1. Limitations of Conventional Computer Vision Toolkit-Based Approaches

Traditional computer vision (CV) toolkits, such as GluonCV, have been fundamental in advancing and democratizing deep learning techniques in this field. They provide robust and optimized implementations of standard algorithms for tasks like image classification, object detection, semantic and instance segmentation, pose estimation, and action recognition. These toolkits typically offer a wide range of pre-trained models (e.g., ResNet, MobileNet, YOLOv3, SSD, Faster R-CNN, FCN), training scripts that reproduce state-of-the-art results, and APIs designed to facilitate rapid prototyping and reproducible research. Furthermore, many modern toolkits, including GluonCV, support multiple deep learning backends like Apache MXNet and PyTorch, offering flexibility to developers. Export and deployment capabilities, sometimes optimized for specific hardware (Intel CPU with MKL-DNN, NVIDIA GPUs with CUDA) or even for quantized inference (INT8) to improve performance, are also common features.

However, for nuanced and high-risk applications like sensitive content detection, these traditional toolkits present significant limitations. Their understanding of content is primarily based on the recognition of objects or visual patterns defined during training, lacking the deep contextual understanding often necessary to discern the sensitive nature of an image or video. For example, differentiating between an image of real violence and a scene from a movie, or between satire and genuine hate speech, requires an understanding that goes beyond simple object identification.

Moreover, these toolkits generally require large labeled datasets to train or fine-tune models for new categories. The dynamic and ever-evolving nature of sensitive content (new memes, emerging hate symbols, etc.) makes the continuous creation of large labeled datasets impractical and time-consuming. Standard transfer

learning capabilities, while useful, may not be sufficient to rapidly adapt to very specific or rare categories with few available examples.

Finally, toolkits like GluonCV do not typically integrate explicit and dedicated mechanisms for explainability (eXplainable AI - XAI) or for the evaluation and mitigation of fairness. While it is possible to apply XAI or Fairness techniques using external libraries, the lack of integration complicates a deep analysis of the model's internal behavior or the systematic application of bias mitigation strategies during training or post-processing.

4.2. Advanced Requirements for Contextual, Ethical, and Efficient Detection

Effective sensitive content detection demands capabilities that overcome the limitations of traditional approaches. First, enhanced contextual understanding is required. It's not enough to merely identify the objects present; it's crucial to understand the relationships between them, the environment, possible cultural or social connotations, and the implicit intent, which is often conveyed through subtle visual cues or the interaction between visual and textual elements (e.g., superimposed text on an image). Multimodal models, which jointly process information from different sources like images and text, are promising for addressing this challenge.

Second, the changing nature of sensitive content necessitates models with few-shot learning (FSL) or zero-shot learning (ZSL) capabilities. FSL allows a model to learn to recognize new categories from very few labelled examples, while ZSL aims to classify categories never seen during training, based solely on auxiliary information such as textual descriptions. These capabilities would drastically reduce the reliance on large labeled datasets for each new emerging sensitive category.

Third, the implementation of sensitive content detection systems carries significant ethical and social risks. The decisions of these systems (e.g., flagging, removing, or restricting content) can have significant consequences for freedom of expression, access to information, and fairness. Therefore, transparency through XAI is fundamental to understanding why a model classifies content as sensitive, allowing for auditing, debugging, and justification of decisions.

Finally, addressing fairness is indispensable. AI models can learn and amplify biases present in training data, which could lead to disproportionate or unfair content moderation for certain demographic groups. Tools and methodologies are needed to detect these biases and apply mitigation techniques to ensure that systems operate equitably for all users.

5. Analysis of Proprietary AI Models for Image Recognition

5.1 Analysis of the Google Cloud Vision API Model

5.1.1. Model Identification

- Model Name: Google Cloud Vision API
- Model Type: Vision
- Provider: Google

5.1.2. General Model Description

What tasks does the model currently perform?

The Google Cloud Vision API offers image analysis functions including image labeling, object detection, detection of places, actions, geographic landmarks, and logos. It also performs Optical Character Recognition (OCR) and explicit content detection through its SafeSearch feature. The Vertex AI Vision platform, which complements the API, includes a Content Moderator with broader predefined labels and the ability to use custom labels.

What tasks is it not specialized in and would require adaptation?

The API in its current state is not specialized in the detection of all specified sensitive content categories, such as "Bullying," "Addiction," "Eating Disorders," "Gambling," "Sexual Harassment," "Homophobia," "Racism," "Radicalization," "Self-Harm and Suicide," and "Cybersecurity." The SafeSearch feature primarily focuses on explicit content like adult material and violence.

Even with fine-tuning, what can it not do or what are its limitations?

Even with fine-tuning or the creation of custom labels in Vertex AI Vision, achieving perfect detection across all sensitive content categories may be impossible. The contextual understanding needed to detect some types of sensitive content, such as bullying or radicalization, presents inherent challenges for any automated system based solely on images.

Foreseen use cases (future specialization in sensitive content detection). The API's specialization, especially through custom labels and fine-tuning in Vertex AI Vision, would allow its use to detect a wider range of sensitive content in images, addressing categories such as bullying, hate symbols, or the presence of weapons. Prompting with Gemini Pro Vision could also be explored for complex sensitive content detection.

5.1.3. Technical Capabilities

Accuracy and recall (if available).

There are no formal comparative benchmarks for accuracy and recall across all desired sensitive content categories. The SafeSearch feature has proven useful with few false positives for explicit content. The Content Moderator in Vertex AI Vision returns confidence scores, allowing sensitivity adjustment. An accuracy of up to 98% in harmful content detection is claimed but has not been formally demonstrated.

Processing speed (images/second, words/second, etc.).

The API supports asynchronous offline batch processing. Fast decision times (under 0.1 seconds) are claimed.

Fine-Tuning Capability:

Does it allow Fine-Tuning?

Yes, Vertex AI Vision allows fine-tuning of pre-trained models.

Vertex AI supports different fine-tuning methods:

- Parameter-efficient tuning: Only a small portion of the model's parameters are adjusted, reducing training cost and time.
- Full fine-tuning: All model parameters are adjusted, which can offer better results for complex tasks but requires more computational resources.

How difficult or costly is it?

Fine-tuning requires a higher level of technical expertise and can be more costly in terms of computational resources and development time.

Workload:

Can it process between 2000 and 3000 units (images/texts) daily?

Yes, processing 2000 to 3000 images daily seems feasible using the API's asynchronous batch processing, which supports up to 2000 images per request.

Does it require high CPU, GPU, RAM resources?

As a cloud-based service, the main processing is performed on Google's infrastructure, so substantial local CPU, GPU, or RAM resources are not required for basic API use. Using Vertex AI for fine-tuning would require additional resources within Google Cloud.

5.1.4. Licensing and Terms of Use

Is it open source or proprietary?

Proprietary

License type (Apache 2.0, MIT, commercial license, etc.).

Documentation content typically has a Creative Commons Attribution 4.0 license, and code examples are Apache 2.0. The API's use is governed by the Google API Terms of Service.

Does it allow commercial use?

Yes, commercial use is permitted, but resale of the API service is prohibited.

Approximate cost:

Annual or monthly license.

There is no fixed annual or monthly license.

Cost per use (if applicable, e.g., pay-per-API, etc.).

The pricing model is pay-per-use. The SafeSearch feature is free with label detection, or $1.50 per 1000 units for the next 5,000,000 units per month. For 60,000 to 90,000 images per month, the estimated cost for SafeSearch would be between $88.50 and $133.50. Use of other features or Vertex AI will incur additional costs.

5.1.5. Legal and Ethical Analysis

GDPR compliance and European regulations.

Google Cloud is committed to GDPR compliance. The Google Cloud Vision API complies with the Cloud Data Processing Amendment. It is the user's responsibility to ensure the overall compliance of their application with GDPR and obtain appropriate consent for data processing, especially for minors.

Bias handling (racial, gender, cultural...).

There is a possibility of bias in AI models based on training data. It is important to be aware of these potential biases and take steps to monitor and mitigate their impact.

Explainability capacity ("explainability") of decisions.

Google Cloud has specific tools to explain decisions made by its Artificial Intelligence (AI) models. While the standard API may not offer detailed explanations of how a decision was reached, when custom models are created, Google offers advanced features to analyze and understand these decisions.

These Explainable AI (XAI) tools, primarily available in Vertex AI and BigQuery ML, provide detailed insights into which factors influenced the model's results.

5.1.6. Infrastructure Requirements

Can it be deployed locally or does it require the cloud?

It requires the cloud, as it is a Google Cloud service.

Is it feasible to have a server per school or a global server instead?

A global cloud server managed by Google Cloud is the most viable architecture, as the API is accessed over the Internet.

Estimated infrastructure consumption (CPU, GPU, RAM).

The primary infrastructure consumption falls on Google Cloud. The user would need an Internet connection and storage for images.

5.1.7. Advantages and Limitations

Aspect	Advantages	Limitations
Accuracy	High for explicit content (adult, violence). Confidence scores for sensitivity adjustment in Vertex AI.	Limited coverage of all specified sensitive content categories. Possibility of false positives and negatives.
Cost	Pay-per-use model, scalable. Free tier for the first 1000 units/month.	Costs can increase with the use of multiple features or Vertex AI for specialization.
Ease of Integration	Easy-to-integrate REST and RPC API.	Requires technical expertise for advanced specialization (fine-tuning).
Explainability	Explainable AI tools available in Vertex AI and BigQuery ML.	Direct explainability for the basic API's SafeSearch feature may be limited.
Flexibility for Fine-Tuning	High flexibility for specialization with custom labels in Vertex AI. Ability to fine-tune models with Vertex AI Vision.	Effective custom label creation and fine-tuning require quality training data and technical expertise.
Regulatory Compliance	Google Cloud is committed to GDPR compliance. COPPA-compliant services for minor protection.	The user is responsible for correct implementation and data handling to ensure full compliance. Adequate consent is required for processing minors' data.

5.1.8. Conclusions and Recommendations

Is this model suitable for our use case?

Google Cloud Vision API is suitable for detecting explicit content such as adult material and violence, thanks to its SafeSearch feature. However, in its standard form, it is not entirely suitable for comprehensive detection of sensitive content such as bullying, addictions, eating disorders, gambling, sexual harassment, homophobia, racism, radicalization, self-harm, suicide, and cybersecurity.

What conditions or adjustments would be necessary to use it?

To fully adapt the Google Cloud Vision API to the specific use case, a comprehensive approach combining advanced techniques and complementary measures would be necessary. The creation and use of custom labels through Vertex AI Vision would significantly extend the model's scope to broader categories of sensitive content. Additionally, performing fine-tuning with specific datasets would substantially improve accuracy in particularly complex areas such as bullying, radicalization, and other sensitive topics. It is also essential to implement robust additional processes to ensure explicit consent is obtained, thus guaranteeing strict GDPR compliance, especially when handling sensitive data of minors. Finally, to address the model's inherent limitations and reduce the risks of false positives and negatives, it is recommended to integrate effective human review mechanisms that act as an additional layer of validation and quality control over the results generated by the API.

5.1.10. Visual Summary

Category	**Result**
Model Type	Vision
Accuracy	Medium / High
Cost	$/month or cost per API
License	Proprietary (commercial use allowed)
Fine-Tuning	Yes (medium difficulty)
GDPR Compliance	Partial
Final Recommendation	Requires adaptation

5.2 Analysis of the Amazon Rekognition Model

5.2.1. Model Identification

- Model Name: Amazon Rekognition
- Model Type: Vision
- Provider: Amazon Web Services (AWS)

5.2.2. General Model Description

What tasks does the model currently perform?

Amazon Rekognition can currently perform various tasks related to visual content analysis. It is capable of detecting a wide variety of objects, scenes, and concepts in images and videos, providing descriptive labels based on their visual content. For example, it can identify common objects like "tree" or scenes like "beach." Additionally, the service offers facial analysis functionalities, detecting faces in images and videos and analyzing facial attributes such as the presence of glasses,

emotions (e.g., happiness), and the estimation of the age range of detected individuals. It also includes celebrity recognition, identifying known public figures in images and videos. Another important capability is the detection and extraction of text from images and videos, which allows for the conversion of visual text into machine-readable format. One aspect is the content moderation capability offered by Amazon Rekognition. The service can identify images and and videos containing inappropriate, offensive, or unwanted content across various categories, such as nudity, violence, and drugs. Finally, Amazon Rekognition allows for the creation of custom models through its Custom Labels feature. This enables users to train the model to detect specific objects, scenes, and concepts relevant to their particular needs by uploading and labeling a set of training images.

What tasks is it not specialized in and would require adaptation?

While the model can generally detect "violence" or "adult content," the precise identification of "bullying" or content related to "eating disorders" might require more specific training. Detecting abstract concepts or complex situations that lack a direct and clear visual representation, such as "bullying" (which often involves complex social interactions) or "addiction" (which is a behavioral state), would necessitate very specific training and could be difficult to achieve with high accuracy solely through image analysis. Therefore, the use of Amazon Rekognition Custom Labels or the Custom Moderation feature would be required to train models that can recognize these types of sensitive content based on relevant visual examples. Custom Moderation, in particular, allows for improving the accuracy of the base model for specific moderation labels by training with user-provided images. Adaptation through Custom Labels or Custom Moderation is essential for handling the different types of specified sensitive content.

Even with fine-tuning, what can it not do or what are its limitations?

Even with fine-tuning through Custom Labels or Custom Moderation, there are inherent limitations to computer vision. Detecting internal mental states, such as "addiction" or "suicidal intentions," based solely on images can be impossible or highly unreliable. While images of drug-related objects or representations of self-harm could be identified, inferring the underlying mental state is a complex task. Similarly, detecting complex social dynamics like "bullying" is a challenge. Although images of physical aggression could be identified, more subtle forms of harassment (like cyberbullying through screenshots of messages) might not be directly detectable through image analysis or would require processing multiple data types (e.g., text). Furthermore, the effectiveness of fine-tuning critically depends on the availability and quality of labeled training data for each sensitive content category. If training data is insufficient, biased, or not representative of the

actual content to be encountered in European schools, the performance of the customized model will be limited.

5.2.3. Technical Capabilities

Accuracy and recall (if available).

Amazon Rekognition Custom Labels provides detailed precision and recall metrics for each user-defined label after training a custom model. Precision measures the proportion of positive predictions that were correct, while recall measures the proportion of actual positive instances that were correctly identified by the model. For pre-existing content moderation, the DetectModerationLabels API returns confidence scores for each detected label. The MinConfidence parameter allows users to set a threshold for the labels to be returned, which can directly influence precision and recall.

Processing speed (images/second, words/second, etc.).

The image processing speed used by Amazon Rekognition Image depends on a variety of factors, including image size, the complexity of analysis operations (such as detecting multiple labels or facial analysis), and how the image is provided. For near real-time processing, submitting images as blocks of digital information can be fast for smaller images, as it avoids the latency associated with uploading. In Amazon Rekognition Custom Labels, inference performance, i.e., the speed at which the model can analyze new images to detect custom labels, can be scaled by adjusting the number of provisioned inference units for the model. By increasing inference units, processing speed can be significantly improved, allowing for handling a higher volume of images within a given time period.

Fine-Tuning Capability:

Does it allow Fine-Tuning?

Yes, Amazon Rekognition does allow fine-tuning through Amazon Rekognition Custom Labels and the Custom Moderation feature.

How difficult or costly is it?

The process of using Custom Labels is designed to be relatively straightforward and does not require deep machine learning expertise. It involves collecting and labeling a training dataset of images, which can be done through the AWS console or by using Amazon SageMaker Ground Truth for large-scale labeling of larger datasets. The cost associated with Custom Labels is based on model training hours and inference hours, meaning the time during which the model is active and available to analyze images. Similarly, Custom Moderation involves creating a project and training an adapter using user-annotated images.

Workload:

Can it process between 2000 and 3000 units (images/texts) daily?

Yes, Amazon Rekognition is designed to be highly scalable and can process large volumes of images and videos daily. The ability to process between 2000 and 3000 images daily will depend on the processing speed per image and the infrastructure used. Since Rekognition is a managed cloud service, it can automatically scale to handle this workload, especially if batch inference is used or sufficient inference units are provisioned for custom models.

Does it require high CPU, GPU, RAM resources?

For the use of Rekognition's pre-trained APIs and for inference with already trained custom models, the underlying infrastructure, including CPU, GPU, and RAM, is fully managed by AWS and is transparent to the user. For training custom models with Custom Labels or creating adapters with Custom Moderation, AWS uses its own computational resources, which may include instances with GPUs to accelerate the training process. Users do not need to directly provision or manage these resources to use the service.

5.2.4. Licensing and Terms of Use

Is it open source or proprietary?

Proprietary

License type (Apache 2.0, MIT, commercial license, etc.).

Documentation content typically has a Creative Commons Attribution 4.0 license, and code examples are Apache 2.0. The API's use is governed by the Google API Terms of Service.

Does it allow commercial use?

Yes, commercial use is permitted.

Approximate cost:

Annual or monthly license.

There is no fixed annual or monthly license.

Cost per use (if applicable, e.g., pay-per-API, etc.).

The cost is based entirely on the use of the different services and APIs. The pricing model is pay-per-use. Costs are incurred for the quantity of images and videos analyzed, the duration of the analysis, training and inference hours for custom models, storage of facial metadata, and the use of other features.

5.2.5. Legal and Ethical Analysis

GDPR compliance and European regulations.

AWS is committed to helping its customers comply with the European Union's General Data Protection Regulation (GDPR). This includes the incorporation of Standard Contractual Clauses (SCCs) for transfers of personal data outside the European Economic Area (EEA). AWS customers maintain control over the physical location where their data is stored and have the ability to choose the security status of their data, including encryption both in transit and at rest. In the specific context of processing personal data of minors in schools, it is imperative to comply with the particular GDPR regulations designed for child protection. These regulations demand special consideration for their privacy and the need to obtain verifiable parental or guardian consent for certain online services offered. The use of image analysis technologies for minors will require a clear legal basis for processing, which could be explicit consent or the school's legitimate interest in ensuring a safe environment.

Bias handling (racial, gender, cultural...).

Concerns have been raised about the potential for bias in facial recognition algorithms, including Amazon Rekognition, with some studies indicating higher error rates for individuals with darker skin tones. While these studies have primarily focused on facial recognition for law enforcement purposes, it's important to recognize that bias could affect accuracy in other facial analysis tasks, such as emotion detection or age estimation, which could be relevant in the context of sensitive content detection (e.g., related to harassment or radicalization). Amazon states that its models, including Face Liveness, are trained and tested using datasets that represent a wide diversity of facial characteristics and skin tones, with the aim of mitigating bias. However, it is crucial to be aware of these potential biases and take proactive measures to mitigate them when using Rekognition. This could include evaluating model performance on datasets that are representative of the school population, carefully adjusting confidence thresholds for model predictions, and using Amazon Rekognition Custom Labels with specific and balanced training data for the sensitive content categories to be detected. The Custom Moderation feature also offers the possibility of improving accuracy for very specific use cases by training with proprietary data.

Explainability capacity ("explainability") of decisions.

Amazon Rekognition provides confidence scores for each of its predictions, indicating the model's level of certainty about its outcome. These scores can be useful for users to understand the reliability of the detections made by the system. For custom models created using Amazon Rekognition Custom Labels, detailed

performance metrics are provided after the training process. These metrics offer valuable information about how the model performs across different categories and help identify areas where improvements could be made. Additionally, Amazon Augmented AI (A2I) integrates with Rekognition, allowing for the creation of workflows for human review of predictions that have a low confidence level. This adds an additional layer of explainability and allows human experts to validate or correct decisions made by the automated model. In the case of the Custom Moderation feature, the AWS console allows users to verify predictions made by the customized model on a test dataset. This functionality is crucial for understanding how the model is making its decisions and for identifying potential errors or false positives that may arise.

5.2.6. Infrastructure Requirements

Can it be deployed locally or does it require the cloud?

It requires the cloud, as it is an AWS service.

Is it feasible to have a server per school or a global server instead?

A global cloud server managed by AWS is the most viable architecture, as the API is accessed over the Internet.

Estimated infrastructure consumption (CPU, GPU, RAM).

The primary infrastructure consumption falls on AWS. The user would need an Internet connection and storage for images.

5.2.7. Advantages and Limitations

Aspect	Advantages	Limitations
Accuracy	High with fine-tuning and quality training data. The base model already offers moderation for certain categories.	Heavily dependent on the quality and quantity of training data for specialization in the specific types of sensitive content.
Cost	Pay-per-use model that can be cost-effective for moderate volumes. Scalability to handle large volumes.	Costs can accumulate for large volumes of images or intensive use of custom models.
Ease of Integration	Easily integrates with applications through APIs and SDKs.	Requires programming knowledge for integration into existing school applications.
Explainability	Provides confidence scores for predictions. Performance metrics available for custom models. Integration with Amazon A2I for human review.	Full explainability of deep learning models is an inherent challenge.
Flexibility for Fine-Tuning	Amazon Rekognition Custom Labels and Custom Moderation allow for high flexibility to adapt the model to specific needs.	Fine-tuning requires the collection and labeling of relevant training data, which can be a laborious and costly process.
Regulatory Compliance	AWS provides tools and documentation to assist with GDPR compliance.	The user (the school or organization) is responsible for ensuring full GDPR compliance, especially regarding the processing of minors' data.

5.2.8. Conclusions and Recommendations

Is this model suitable for our use case?

Amazon Rekognition presents a solution with potential for sensitive content detection in images within European school environments, but its suitability depends on several critical factors. The base model offers content moderation capabilities that can serve as a starting point. However, the specificity of the sensitive content types identified in the prompt requires significant adaptation through Amazon Rekognition Custom Labels or the Custom Moderation feature.

What conditions or adjustments would be necessary to use it?

To effectively use Rekognition in this context, several adjustments and conditions would be necessary. Firstly, a substantial effort will be required in the collection, labeling, and curation of high-quality training data that is representative of the specific sensitive content relevant to European schools. The model's accuracy will directly depend on the quality of this data. Secondly, the confidence thresholds for the model's predictions must be carefully configured, seeking a balance between precision (minimizing false positives) and recall (minimizing false negatives). Thirdly, it is fundamental to implement robust measures to ensure GDPR compliance and the protection of minors' privacy, including obtaining appropriate consent when necessary and transparency in data processing. Additionally, the potential for bias in the model must be actively addressed by continuously evaluating its performance across diverse demographic groups and adopting strategies to mitigate any identified biases. Finally, integration with Amazon Augmented AI (A2I) for human review of low-confidence predictions might be necessary to ensure accuracy and provide a layer of explainability in the system's decisions.

5.2.9. Visual Summary

Category	**Result**
Model Type	Vision
Accuracy	Medium / High
Cost	$/month or cost per API
License	Proprietary (commercial use allowed)
Fine-Tuning	Yes (medium difficulty)
GDPR Compliance	Partial
Final Recommendation	Requires adaptation

6. Analysis of Open Source AI Models for Image Recognition

6.1 Analysis of the GluonCV Model

6.1.1. Model Identification

- Model Name: GluonCV (Toolkit/Library). It's important to note that GluonCV is not a single model, but a set of tools and a collection of pre-trained models (known as a "model zoo"). For a specific task, a concrete model from this zoo would be selected and adapted, such as ResNet50_v1d for classification or ssd_512_resnet50_v1_coco / yolo3_darknet53_coco for object detection. The specific version will depend on the chosen backend (MXNet or PyTorch) and the time of installation. The installable library name is gluoncv. A relevant consideration is that GluonCV is an open-source project maintained by the community, which influences the support and pace of updates compared to commercial models. Furthermore, the official documentation suggests exploring AutoGluon for image classification or object detection needs, indicating it might be a more actively developed successor within the ecosystem. This possible reduction in GluonCV's maintenance activity should be taken into account for long-term projects.
- Model Type: Vision
- Provider: DMLC (Distributed Machine Learning Community) and open-source community contributors. Originally closely tied to Apache MXNet, it later added support for PyTorch. Scientists from Amazon Web Services (AWS) have been involved in its creation and promotion.

6.1.2. General Model Description

What tasks does the model currently perform?

- GluonCV provides implementations of state-of-the-art (SOTA) models for a variety of fundamental computer vision tasks. These include:
- Image Classification: Recognizing the main object in an image (e.g., ResNet, MobileNet, VGG).
- Object Detection: Locating multiple objects in an image using bounding boxes (e.g., SSD, YOLOv3, Faster R-CNN).
- Semantic Segmentation: Assigning a class label to each pixel in an image (e.g., FCN, PSP, DeepLabV3).
- Instance Segmentation: Similar to semantic, but differentiating individual instances of the same class (e.g., Mask RCNN).

- Pose Estimation: Detecting the pose of human figures in images (e.g., Simple Pose).
- Video Action Recognition: Identifying human actions in video sequences (e.g., TSN, I3D).

Other tasks such as Depth Prediction, Generative Adversarial Networks (GANs), and Person Re-identification are also supported. GluonCV offers the building blocks and pre-trained models for these generic vision tasks.

What tasks is it not specialized in and would require adaptation?

The main task of interest, the explicit detection of a wide range of sensitive content (Violence, Sexuality, Drugs, Weapons, Bullying, Addiction, Adult content, Cybersecurity, Eating disorders, Gambling, Sexual harassment, Homophobia, Racism, Radicalization, Self-harm and Suicide, etc.), is not a native functionality of GluonCV. Existing models are trained on standard datasets (like COCO, ImageNet, Pascal VOC) that identify common objects (dogs, bicycles, cars), people, or general scenes, but not these specific and complex semantic categories.

Therefore, significant adaptation would be required, primarily through fine-tuning. This would involve defining new classes corresponding to each sensitive content category and re-training or fine-tuning a base model (likely an object detection or classification model) using a specific and labeled dataset for this domain. This process is analogous to adapting vision models for medical or industrial uses, but with much greater ethical and data complexity.

Even with fine-tuning, what can it not do or what are its limitations?

Even with successful fine-tuning, the models available in GluonCV would present significant inherent limitations for this task:

- Lack of Deep Contextual Understanding: Classification and detection models learn to recognize visual patterns. They can learn to detect a "knife," but lack the intrinsic ability to discern whether that knife is presented in an innocuous culinary context or a threatening one (violence). Fine-tuning can improve the detection of objects or scenes correlated with sensitive content, but it does not equip the model with a real semantic understanding of context or intent.
- Subjectivity and Nuances: Many of the defined sensitive categories (e.g., "Bullying," "Radicalization," "Adult content") are inherently subjective, depend heavily on cultural and social context, and often do not have unique or unambiguous visual correlates. Visual cues can be ambiguous or insufficient on their own. A fine-tuned model might identify certain visual

elements, but correctly interpreting the situation remains a fundamental challenge.

- Dependence on Labeled Data: Standard models in GluonCV generally require a considerable amount of labeled data to effectively learn new categories. They lack the advanced zero-shot or few-shot learning capabilities that could allow the detection of new sensitive categories with few or no specific examples, unlike more recent multimodal architectures (which are not the focus of GluonCV).
- Limited Explainability: Models like Convolutional Neural Networks are often considered "black boxes." GluonCV does not incorporate native Explainable AI (XAI) tools. This means that, by default, the system would not be able to explain why it classified an image as sensitive, which is a critical limitation for high-risk applications requiring transparency, accountability, and the ability to audit decisions, especially those affecting minors.

6.1.3. Technical Capabilities

Accuracy and recall (if available).

GluonCV provides detailed performance metrics for its models, but these refer to standard benchmarking tasks and datasets, such as ImageNet for classification, and COCO or Pascal VOC for object detection. There are no published metrics for the specific task of sensitive content detection as defined in the objective, as this would require a fine-tuned model and a specific evaluation dataset that are not part of the standard toolkit.

Processing speed (images/second, words/second, etc.).

Exact FPS (Frames Per Second) figures are not provided. However, several conclusions can be drawn:

- Model and Hardware Dependence: Speed varies greatly depending on model complexity (e.g., MobileNet is much faster than ResNet or heavier architectures) and the hardware used (CPU vs. GPU, specific GPU model).
- GPU Benefits: GPU usage is recommended for optimal inference performance. Training benchmarks on V100 GPUs show throughputs of thousands of images per second.
- INT8 Optimization for CPU: Quantization to INT8 allows for very noticeable accelerations (up to 7x reported in specific benchmarks with Intel Xeon CPUs supporting VNNI, such as those in AWS C5 instances). This opens the door to more economical CPU-based deployments if accuracy remains acceptable.
- General Estimation: Although exact figures require specific benchmarking, efficient models like MobileNet on a modern GPU can process hundreds or

thousands of images per second. Even with an optimized CPU, speeds of tens or hundreds of images per second can be expected for lightweight models.

Fine-Tuning Capability:

Does it allow Fine-Tuning?

Yes, GluonCV is explicitly designed to facilitate fine-tuning. It provides detailed tutorials for fine-tuning pre-trained classification, detection, and action recognition models on custom datasets. It allows control over which model layers are trained and which are frozen.

How difficult or costly is it?

Difficulty: High. It requires solid Machine Learning expertise. The process involves understanding model architectures, properly preparing and augmenting data, selecting appropriate loss functions, implementing custom training loops, tuning hyperparameters, and performing rigorous validation. While GluonCV provides tools and examples, adapting to a new and sensitive domain like this adds a considerable layer of technical and conceptual complexity.

Cost: Significant, dominated by factors beyond the software:

- Data Acquisition and Labeling: This is likely the biggest cost and challenge. Obtaining a large, diverse, representative, and, crucially, ethically and legally obtained (GDPR-compliant) dataset for the multiple sensitive content categories, especially if it involves images of minors, is extremely costly and complex.
- ML Expertise: Highly qualified personnel are needed to design the fine-tuning strategy, implement it, validate it, and maintain it.
- Computational Resources: Training and fine-tuning deep learning models requires considerable access to high-end GPU resources, which implies hardware costs or cloud service usage.
- Maintenance Costs: Ongoing effort to monitor model performance, retrain as needed (model drift), update dependencies, and manage infrastructure.
- Compliance Costs: Resources dedicated to ensuring strict GDPR compliance, including legal advice, conducting DPIAs, implementing security and privacy measures, and managing data subject rights.
- It is crucial to understand that the "zero" cost of the GluonCV license does not reflect the actual cost of implementing and operating a functional and compliant solution for this complex use case.

Workload:

Can it process between 2000 and 3000 units (images/texts) daily?

Yes, easily.

Does it require high CPU, GPU, RAM resources?

For Training/Fine-tuning: Yes, it requires high resources. Powerful GPUs are needed, significant amounts of RAM (both system and GPU, >12GB VRAM), and fast storage (SSD recommended). Multi-GPU configurations are beneficial.

For Inference: Requirements depend on the chosen model, desired latency, and processing volume.

In summary, processing capacity for the average daily workload is not an issue. The technical challenges lie in the complexity and cost of fine-tuning (especially data acquisition) and in adequately sizing the inference infrastructure to handle peak loads with the required latency, weighing the cost of GPU versus optimized CPU.

6.1.4. Licensing and Terms of Use

Is it open source or proprietary?

GluonCV is an Open Source project. The source code is publicly available.

License type (Apache 2.0, MIT, commercial license, etc.).

It is distributed under the Apache 2.0 License. This is a permissive free software license. Its main features include:

- Allows the use, modification, and distribution of the software.
- Allows the distribution of derivative or larger works under different terms and without the need to release the source code.
- Explicitly grants patent rights by contributors.
- Requires the preservation of copyright and license notices, and indication of changes made to the code.
- Provides no warranties and limits the licensor's liability.

Does it allow commercial use?

Yes, the Apache 2.0 license explicitly permits commercial use of the software.

Approximate cost:

- Annual or monthly license: €0. There are no direct licensing costs associated with using GluonCV due to its open-source nature under Apache 2.0.

- Cost per use (if applicable, e.g., pay-per-API, etc.): Not directly applicable. GluonCV is a library, not a pay-per-use API service.
- Indirect Costs (Total Cost of Ownership - TCO): Despite the free license, implementing a GluonCV-based solution for this use case will incur very significant indirect costs:
    - Development Costs: Requires considerable investment in highly experienced ML personnel to select models, perform fine-tuning, develop the processing pipeline, validate results, and integrate the solution.
    - Data Costs: Potentially the highest cost and challenge. Includes the acquisition (if possible and ethical), cleaning, and manual or semi-manual labeling of a large volume of images to cover all sensitive categories in the specific context.
    - Infrastructure Costs: Deployment and maintenance of servers (on-premise or cloud), including powerful CPUs, GPUs (if chosen for inference/training), RAM, storage, and networking. GPU costs can be substantial.
    - Maintenance Costs: Ongoing effort to monitor model performance, retrain as needed (model drift), update dependencies, and manage infrastructure.
    - Compliance Costs: Dedicated resources to ensure strict GDPR compliance, including legal advice, conducting DPIAs, implementing security and privacy measures, and managing data subject rights.

It is crucial to understand that the "zero" cost of the GluonCV license does not reflect the true cost of implementing and operating a functional and compliant solution for this complex use case.

6.1.5. Legal and Ethical Analysis

GDPR compliance and European regulations.

This is possibly the most critical and challenging aspect of the project. GluonCV, as a software library, does not process data itself and therefore is not inherently GDPR compliant or non-compliant. Compliance depends entirely on how the final solution based on GluonCV is implemented and used.

Implementing an AI system that processes images of students in European schools to detect sensitive content is a high-risk activity under GDPR, especially because it involves sensitive data (potentially revealing information about health, sexual orientation, beliefs, etc.) and minors (considered vulnerable data subjects). Key challenges include:

- Legal Basis for Processing (Art. 6 and Art. 9 GDPR): A robust legal basis is required. "Legitimate interest" (Art. 6(1)(f)) would need a very careful balancing test demonstrating that the interest in protecting minors outweighs their fundamental rights and freedoms, considering the intrusive nature of monitoring. Processing special categories of data (Art. 9) is prohibited by default and requires specific additional conditions (e.g., explicit consent, substantial public interest defined by law). Consent (Art. 6(1)(a), Art. 9(2)(a)) must be informed, specific, freely given, and unambiguous, and may be difficult to obtain validly from minors or could be withdrawn. Other bases like legal obligation or public task might apply depending on the specific country/region's legal framework.
- Purpose Limitation (Art. 5(1)(b)): The purpose (detection of specific sensitive content for safety) must be explicit, legitimate, and defined from the outset. Using data collected for other purposes (e.g., school photos) for this new purpose requires a compatibility justification or a new legal basis.
- Data Minimization (Art. 5(1)(c)): Only strictly necessary data should be processed. Is it necessary to store images? Can the analysis be performed transiently? Can effective pseudonymization or anonymization techniques be applied? Minimization also applies to data used to train the model. There is an inherent tension between the need to minimize data and the need for large, diverse datasets to train robust models and mitigate biases.
- Accuracy (Art. 5(1)(d)): Errors (false positives/negatives) can have serious consequences. High accuracy and continuous monitoring are required.
- Storage Limitation (Art. 5(1)(e)): Clear retention and deletion policies must be established for images and generated metadata.
- Integrity and Confidentiality (Art. 5(1)(f), Art. 32): Robust technical and organizational security measures are required to protect data against unauthorized access, loss, or breaches.
- Data Subject Rights (Chapter III GDPR): Mechanisms must be implemented to facilitate the exercise of rights of access, rectification, erasure ("right to be forgotten"), restriction of processing, data portability, and objection. The right not to be subject to automated individual decisions (Art. 22) is crucial, which likely requires human intervention in the alert review process.
- Data Protection Impact Assessment (DPIA) (Art. 35): Given the nature of the processing (novel technology, sensitive data, minors, large-scale monitoring), a DPIA is almost certainly mandatory before commencing processing.
- Transparency (Art. 13, 14): Clear and comprehensive information must be provided to students and/or their legal guardians about how their data is processed.

- Data Protection by Design and by Default (Art. 25): Data protection principles must be integrated into the system's design from the outset.

GluonCV provides no functionality to address these requirements. All responsibility rests with the organization implementing the solution.

Bias handling (racial, gender, cultural...).

Computer vision models are prone to learning and amplifying biases present in training data. These biases can be related to protected attributes such as race, gender, age, or cultural origin.

Risk with GluonCV: Pre-trained models in GluonCV, based on standard datasets, likely contain biases. When fine-tuning for sensitive content, there is a high risk that these biases will be perpetuated or even amplified. This could lead to discriminatory results, for example, the system incorrectly flagging content as sensitive more frequently for certain demographic groups or failing to detect sensitive content predominantly affecting minority groups. Some architectures like Vision Transformers (ViTs) might even be more prone to amplifying biases than traditional CNNs.

Lack of Native Tools: GluonCV does not include specific tools for detecting or mitigating algorithmic biases.

Mitigation Strategies (External): Addressing bias requires proactive effort and the application of techniques external to the toolkit, which can be grouped into:

- Pre-processing (Acting on data): Includes careful curation of datasets to ensure diversity and representativeness, resampling techniques (oversampling/undersampling), reweighing samples, or data augmentation designed to reduce bias.
- In-processing (Modifying the learning algorithm): Includes adding regularization terms to the loss function to penalize dependence on protected attributes or training the model with specially designed examples so that it learns not to make mistakes when attempts are made to deceive it or in difficult-to-interpret cases (although the latter may negatively affect overall performance).
- Post-processing (Adjusting predictions): Includes adjusting decision thresholds differently for different groups to achieve some fairness metric.

Need for Audit: It is fundamental to audit both datasets and trained models using specific fairness metrics (e.g., demographic parity, equality of opportunity) to identify and quantify existing biases.

Bias mitigation is a complex and active field of research. Implementing these techniques requires additional expertise and adds complexity to system development and validation.

Explainability capacity ("explainability") of decisions.

Explainability (XAI) refers to the ability to understand and justify how an AI model arrives at a specific decision. It is crucial in high-risk applications to build trust, debug errors, ensure fairness, and meet regulatory requirements.

GluonCV Limitation: Standard deep learning models available in GluonCV (CNNs, etc.) are inherently complex and often operate as "black boxes," making it difficult to understand their internal logic. GluonCV does not provide integrated XAI functionalities.

External Tools: To obtain explanations, it would be necessary to integrate third-party XAI tools after training the model with GluonCV. The most popular are:

- LIME (Local Interpretable Model-agnostic Explanations): Explains individual predictions by creating a simple interpretable model (e.g., linear) that locally approximates the behavior of the complex model around the instance to be explained. It can be applied to any type of model, but only explains individual cases, and the quality of those explanations can change depending on the situation. It works with tabular, text, and image data.
- SHAP (SHapley Additive exPlanations): Based on game theory (Shapley values), it assigns each feature a contribution to the final prediction. It offers both local and global explanations and has a strong theoretical basis. It can be computationally expensive, especially for non-tree-based models.
- Additional Complexity: The implementation and validation of these XAI techniques add a significant layer of technical complexity and computational burden to the system. Furthermore, the reliability and consistency of explanations generated by these methods is an active area of debate.

In conclusion, the lack of integrated bias mitigation and explainability tools in GluonCV represents a significant disadvantage for this use case, increasing development burden, costs, and risks associated with creating a responsible and reliable system.

6.1.6. Infrastructure Requirements

Can it be deployed locally or does it require the cloud?

GluonCV, being a Python library, offers flexibility in deployment. Models trained with GluonCV can be run in any environment where the necessary dependencies (Python, MXNet/PyTorch, hardware drivers) can be installed. This includes:

- Local Deployment (On-Premise): On servers physically located within the school's or school district's infrastructure.
- Cloud Deployment: Using cloud provider services (AWS, Azure, GCP). AWS SageMaker, for example, has explicit support for GluonCV models in some of its functionalities.
- Edge Deployment: On devices with lower computational capacity located close to where data is generated (e.g., within the school itself on specific devices).

Is it feasible to have a server per school or a global server instead?

The choice between a decentralized deployment model (server per school) and a centralized one (global server, likely in the cloud) involves a complex balance of technical, economic, operational, and, fundamentally, legal (GDPR) factors.

- Server per School (Local/Edge):
    - Advantages: Greater control over data residency (sensitive images might not leave the school's physical perimeter, potentially simplifying some GDPR aspects if managed correctly), lower latency for local processing.
    - Disadvantages: Higher initial hardware cost (one server for each school), very high operational complexity (management, updates, security, maintenance of multiple distributed servers), possible inconsistency in hardware and performance between schools requires local technical capability at each site. Using smaller edge devices could reduce hardware cost but would limit computing capacity.
- Global Server (Centralized/Cloud):
    - Advantages: Economies of scale in infrastructure, simpler centralized management (updates, maintenance), more consistent performance and potentially greater available computing capacity, flexibility to scale.
    - Disadvantages: Significant privacy and GDPR concerns. Transferring sensitive student images outside the school (especially to a cloud provider, potentially outside the EU) requires a very solid legal basis, extremely robust security measures (encryption in transit and at

rest), compliant data processing agreements (DPAs), and probably a detailed DPIA justifying the transfer. Potentially higher latency depending on the network. Dependence on an external provider and connectivity.

The strategic decision on the deployment model must be strongly guided by GDPR requirements and the privacy risk assessment. A local model may seem preferable from a data control perspective, but its operational complexity could make it unfeasible. A centralized model simplifies operation but introduces considerable legal and privacy challenges that must be meticulously addressed. A hybrid model (e.g., local pre-processing or basic inference with centralized analysis of metadata or alerts) could be explored.

Estimated infrastructure consumption (CPU, GPU, RAM).

Requirements vary drastically between training/fine-tuning and inference:

Training/Fine-tuning:

- CPU: Necessary for process management and data loading (fast CPUs are recommended).
- GPU: Essential. High-end GPUs (e.g., Nvidia V100) are required to efficiently train complex models on large datasets. Multi-GPU configurations accelerate the process.
- RAM: A considerable amount of system RAM and, critically, GPU VRAM (>12GB) is needed.
- Storage: Sufficient space is required for datasets (ImageNet is ~300GB) and SSDs are recommended for fast access.

Inference:

- CPU: It is possible to run inference on CPU, especially if lightweight models and optimizations like MKL-DNN and INT8 quantization are used on compatible hardware (Intel Xeon with VNNI). Performance will be lower than GPU but may be sufficient for low loads or if cost is a primary limiting factor.
- GPU: Recommended for optimal performance (low latency, high processing capacity). The power required will depend on the specific model, expected load, and latency requirements.
- RAM: Requirements are generally lower than for training. They primarily depend on the size of the model that needs to be loaded into memory. A few GB might be sufficient for some models, but it should be adequately sized.
- Storage: Minimal for the model itself. Requirements increase if images need to be stored temporarily or permanently during the process.

The infrastructure for inference must be sized considering expected peak loads, not just the daily average, and the choice between CPU and GPU will depend on the balance between cost, performance, and the feasibility of optimizations like INT8 (which requires accuracy validation).

6.1.7. Advantages and Limitations

Aspect	Advantages	Limitations
Accuracy	High theoretical potential after fine-tuning with quality data.	Actual accuracy for the specific task is unknown and not guaranteed. Requires extensive fine-tuning and validation. Inherent difficulty with context, subjectivity, and nuances of sensitive categories.
Cost	Free software license (Apache 2.0). Possibility of reducing inference costs through optimization (INT8 on CPU).	High Total Cost of Ownership: Massive investment needed in development (ML expertise), data (ethical and legal acquisition/labeling), infrastructure (GPU/CPU), maintenance, and, crucially, regulatory and ethical compliance.
Ease of Integration	Flexible framework with documented APIs and tutorials for standard tasks. Support for MXNet and PyTorch backends.	High Complexity: It's a toolkit, not a plug-and-play solution. Requires building, training, and validating the complete solution. Complexity multiplies for this use case (sensitive data, ethics, GDPR).
Explainability	Theoretical possibility of applying external XAI tools (LIME, SHAP) to trained models.	No native XAI capability. Models are intrinsically "black boxes." Implementing and validating XAI adds a significant layer of complexity and cost.
Flexibility for Fine-Tuning	High. The toolkit is designed for adaptation	Requires high ML expertise and a suitable

	and fine-tuning. Extensive model zoo as a starting point.	sensitive dataset. Creating/obtaining this dataset is the main bottleneck (cost, difficulty, ethics, legality).
Regulatory Compliance	Permissive license does not prevent building compliant systems.	No integrated compliance features. All burden of ensuring GDPR compliance, bias mitigation, and ethical management rests solely with the implementer. Very high legal and ethical risk.

6.1.8. Conclusions and Recommendations

Is this model suitable for our use case?

GluonCV is a technically capable computer vision toolkit that provides the basic components (pre-trained models, training tools) that could be adapted through fine-tuning for the task of sensitive content detection. However, it is not a directly suitable or recommended "as is" solution for this specific use case (European school environments, processing images of minors). The risks and challenges associated with its implementation in this context are extremely high, primarily due to non-technical factors such as the ethical and legal acquisition of data, strict GDPR compliance, bias mitigation, and the need for explainability. GluonCV offers a technological basis but building a responsible and effective solution requires very significant investment and risk management.

What conditions or adjustments would be necessary to use it?

To consider using GluonCV in this project, the following conditions and adjustments would be essential, representing a considerable effort:

- Massive and Ethical Data Investment: The most critical condition is the ability to create or acquire a training and validation dataset that is large, diverse, representative of the European school context, and obtained in a completely ethical and legally GDPR-compliant manner, especially concerning minors' data. This is the biggest obstacle.
- Highly Qualified ML Team: A team with deep expertise not only in fine-tuning vision models but also in bias mitigation techniques, rigorous validation in sensitive domains (with emphasis on recall and fairness), and potentially in implementing XAI tools is required.

- Adequate and Secure Infrastructure: Plan and deploy the necessary hardware infrastructure (likely GPU-based for adequate performance), carefully choosing between a local or centralized model after a thorough assessment of GDPR risks and operational capacity. Infrastructure security is paramount.
- Comprehensive Regulatory and Legal Compliance: A "privacy by design and by default" approach. Mandatory conduct of a Data Protection Impact Assessment (DPIA). Establishment of a clear and documented legal basis for processing. Implementation of all technical and organizational measures required by GDPR (security, rights management, retention policies, transparency). Specialized legal advice on GDPR and child protection is indispensable.
- Proactive Bias Mitigation: Implement a robust pipeline that includes bias auditing in data and models, and the application and validation of appropriate mitigation techniques (pre-, in- or post-processing).
- Explainability (XAI) Integration: Select, implement, and validate external XAI tools (such as LIME or SHAP) to provide transparency and allow auditing of model decisions.
- Significant Human Oversight: Establish a clear and efficient process for human review of all (or a statistically significant and high-risk subset) detections made by the automated system. Given the sensitivity of the context and the inherent imperfection of AI models, relying solely on automated decisions is unfeasible and ethically questionable.

6.1.9. Visual Summary

Category	**Result**
Model Type	Vision
Accuracy	Low (without adaptation) / Medium-High (Potential post-fine-tuning if quality data is available and contextual limitations are overcome)
Cost	€0 (License) / High (TCO: Development, Massive and sensitive data, Infrastructure, GDPR/Ethical Compliance)
License	Apache 2.0 (Commercial use allowed)
Fine-Tuning	Yes / Very High difficulty (Requires sensitive data difficult to obtain ethically/legally, high ML expertise, bias/XAI management)
GDPR Compliance	Implementation must ensure compliance; very high risk due to sensitive data of minors in school environment.
Final Recommendation	Requires massive adaptation and high-risk management. Exploring alternatives (commercial APIs, hybrid/non-AI approaches) is highly recommended.

6.2 Analysis of the RTMDet Model

6.2.1. Model Identification

- Model Name: RTMDet (Real-Time Models for Object Detection). The model is available in various scaled versions based on size and computational complexity, namely RTMDet-tiny, RTMDet-s, RTMDet-m, RTMDet-l, and RTMDet-x. Additionally, there are specialized variants derived from the base architecture, such as RTMDet-Ins for instance segmentation and RTMDet-R for rotated object detection. The selection of a specific variant for a particular application will depend on the desired balance between detection accuracy and required inference speed.
- Model Type: Vision
- Provider: The RTMDet model was developed and is primarily maintained by OpenMMLab, distributed through their MMDetection and MMYOLO toolboxes. However, the open nature and performance of the model have led other organizations to offer implementations, optimizations, or pre-trained models based on RTMDet. These include Qualcomm, with optimized versions for deployment on mobile devices; MathWorks, which integrates it into its MATLAB environment; platforms like Roboflow that facilitate its training and deployment; and institutions like Riksarkivet that have adapted it for specific tasks such as text region detection. This

diversity of providers highlights the adaptability and value of the RTMDet architecture. However, it also introduces possible fragmentation: performance metrics, licensing terms (e.g., Qualcomm's specific license for its compiled assets), and ease of fine-tuning can vary considerably between different versions and providers. This complicates direct comparisons and implementation decisions. For custom fine-tuning, the versions offered by OpenMMLab likely represent the most standardized and documented starting point.

6.2.2. General Model Description

What tasks does the model currently perform?

- RTMDet's primary function is real-time generic object detection. It's designed to compete with, and in many cases, outperform the popular YOLO model series in terms of accuracy-inference speed balance. It's typically trained on large-scale datasets like COCO.
- The RTMDet-Ins variant extends its capabilities to instance segmentation, allowing it to not only localize objects but also delineate their exact pixel-level contours.
- The RTMDet-R variant specializes in rotated object detection, a crucial task in domains like aerial or satellite image analysis where objects can appear in any orientation.
- Its efficiency is based on an architecture that employs convolutional building blocks and aims for compatible computational capacity between the backbone (feature extractor) and the neck (feature fusion).

What tasks is it not specialized in and would require adaptation?

Specific sensitive content detection: RTMDet is not pre-trained to identify the specific categories of interest (Violence, Sexuality, Drugs, Weapons, Bullying, etc.). Base models are usually trained on datasets like COCO, which contain generic categories (people, cars, animals). Therefore, for the required task, adaptation is essential, mainly through fine-tuning, using an image dataset specifically annotated with the relevant sensitive content categories.

Deep contextual or semantic analysis: As an object detector, RTMDet identifies the presence and location of defined visual elements. However, it lacks the inherent ability to interpret complex context, underlying intent, or the semantics of a scene. For example, it cannot distinguish on its own between a theatrical representation of violence and a real violent act, or between educational material on drug prevention and content that promotes drug use. This interpretation requires additional layers of logic or human intervention.

Even with fine-tuning, what can it not do or what are its limitations?

Understanding Context and Intent: Despite fine-tuning, RTMDet will remain fundamentally a visual pattern recognizer. Interpreting contextual nuances, intentions (e.g., satire, social commentary), or distinguishing between harmful content and content with artistic or educational value will remain a considerable challenge. The model's effectiveness will heavily depend on the quality, diversity, and representativeness of the dataset used for fine-tuning, and on how sensitive categories are defined and annotated.

Detection of Abstract or Implicit Content: Concepts like 'radicalization', 'homophobia', 'racism', 'cybersecurity', or 'eating disorders' rarely have a direct, unique, and consistent visual representation. These concepts are abstract and manifest through a variety of signals, often subtle or textual. RTMDet, like most current vision models, cannot detect these concepts directly from images if they are not translated into specific, visually identifiable objects, symbols, or texts on which the model can be explicitly trained.

Inherent and Amplified Biases: The fine-tuning process, while necessary, carries the risk of introducing or amplifying biases present in the specific sensitive content dataset. If this dataset is not carefully curated to be diverse and representative of the European school population (in terms of ethnicity, gender, culture, etc.) or the various ways sensitive content manifests in different contexts, the resulting model could exhibit uneven performance, making more errors (false positives or false negatives) for certain demographic groups, cultural groups, or specific types of content.

Novelty and Evolving Content: The model will have difficulty detecting new or emerging forms of sensitive content that were not present or well-represented in its fine-tuning dataset. Tactics to circumvent moderation and forms of online expression are constantly evolving, which will require periodic retraining and updating of the model to maintain its effectiveness.

The nature of the task (sensitive content detection in schools) and the inherent limitations of the model (contextual understanding, biases, abstract content) make exclusive reliance on RTMDet unfeasible and risky. The consequences of errors (blocking legitimate educational content or allowing harmful content) are too high in this environment. Therefore, the most responsible and realistic implementation involves a hybrid approach: RTMDet acts as a supportive tool that flags potential issues, but all detections (or at least positive ones and a sample of negative ones) must be validated by human reviewers who provide the necessary contextual and ethical judgment. This requirement for human intervention has significant implications for workflow design, resource allocation (moderation personnel), and the overall system architecture.

6.2.3. Technical Capabilities

Accuracy and recall (if available).

Performance on Generic Tasks (COCO Dataset): RTMDet demonstrates high accuracy in detecting common objects. Larger variants, RTMDet-L and RTMDet-X, achieve approximately 52.8% Average Precision (AP) on the COCO validation set. Smaller variants also offer competitive performance: RTMDet-tiny achieves around 41.1% AP, RTMDet-s around 44.6% AP, and RTMDet-m close to 50% AP. These metrics position RTMDet favorably against other contemporary real-time detectors like YOLOv5, YOLOX, or YOLOv6, often offering a better compromise between the number of parameters and achieved accuracy.

Performance on Specialized Tasks: RTMDet-Ins (segmentation) and RTMDet-R (rotated detection) variants have also shown state-of-the-art results in relevant benchmarks, such as COCO for segmentation (44.6% Mask AP for RTMDet-Ins-L) and DOTA for rotated detection (up to 81.3% mAP for RTMDet-R).

Performance on the Specific Task (Sensitive Content): Currently, there is no published data on RTMDet's accuracy and recall specifically for detecting the diverse set of required sensitive content categories. This performance will critically depend on factors such as: the quality, size, and representativeness of the dataset used for fine-tuning; the visual similarity between the new sensitive categories and the original generic COCO categories; the inherent complexity and ambiguity of each sensitive category; and the effort invested in the training and hyperparameter tuning process. It is reasonable to expect a decrease in performance compared to COCO metrics, especially for subtle, ambiguous, or explicitly visually underrepresented categories. An exhaustive and rigorous evaluation of the model after the fine-tuning process on a specific test dataset for this use case will be absolutely necessary.

Processing speed (images/second, words/second, etc.).

- On High-End GPUs: RTMDet demonstrates extremely high inference speed. The L/X variants can exceed 300 FPS (Frames Per Second) using optimizations like TensorRT with FP16 precision and batch size (not including NMS post-processing time). Smaller variants are even faster: RTMDet-tiny can exceed 1000 FPS and RTMDet-s reaches about 819 FPS. In terms of latency, RTMDet-tiny can operate below 1 ms and RTMDet-m around 1.22 ms under these conditions.
- On More Modest GPUs: No direct official figures are provided for these GPUs. However, comparative benchmarks of GPUs in other deep learning tasks suggest that an Nvidia 3090 is considerably faster than a V100

(approximately 2-3 times faster in FP16/FP32 inference for models like ResNet/Inception) and substantially faster than a T4 (the T4 is a lower-end data center GPU). An A100 can outperform the 3090 in training or memory-intensive tasks, but not necessarily in pure inference latency if memory is not the bottleneck. Therefore, inference speed on a T4 will be significantly lower than the 300+ FPS reported for the 3090.

- On CPU or Mobile Devices (Edge): Qualcomm has optimized RTMDet (Medium variant) for Snapdragon mobile platforms, reporting inference latencies between 10-30 ms using the NPU (Neural Processing Unit) with FP16 precision. This indicates viability for edge deployments, albeit at a much lower speed than server GPUs. Exclusive standard CPU inference will be very slow and likely unsuitable for real-time or high-volume analysis.
- It's important to note a possible discrepancy between officially reported inference speeds and those obtained in practical implementations. Official figures are usually measured under highly optimized conditions. Some users have reported difficulties replicating these speeds, obtaining much lower performance. Actual speed will depend on factors such as the specific GPU used, the inference framework (native PyTorch vs. ONNX Runtime vs. TensorRT), numerical precision (FP32 vs. FP16), the inference batch size, and whether post-processing time (like NMS) is included. Therefore, official figures should be considered an optimistic upper bound. It is crucial to perform your own benchmarks on the target deployment infrastructure to obtain realistic performance estimates before making final decisions.

Fine-Tuning Capability:

Does it allow Fine-Tuning?

Yes, explicitly. RTMDet has been designed with extensibility in mind. OpenMMLab's frameworks, MMDetection and MMYOLO, on which RTMDet is based, provide robust tools, documentation, and specific tutorials to facilitate the fine-tuning of pre-trained models on custom datasets. There are also example repositories demonstrating the process.

How difficult or costly is it?

Difficulty: It is considered of moderate difficulty. It requires a solid understanding of Python, the PyTorch framework, and familiarity with the MMDetection/MMYOLO ecosystem, including its configuration file system. Correctly preparing and formatting the custom dataset (generally in COCO format) and precisely configuring training parameters are critical steps that can present challenges.

However, the available tutorials and the modular structure of the frameworks help mitigate some of this complexity.

Cost: The direct costs of the base model are zero as it is open source. The main costs are divided into: development cost (engineering time to collect, annotate, and prepare the sensitive content dataset; configure and run fine-tuning experiments; evaluate and validate the resulting model) and computational cost (access to GPUs with sufficient memory and computing power to perform fine-tuning efficiently). There are no software licensing costs for the base RTMDet model.

Workload:

Can it process between 2000 and 3000 units (images/texts) daily?

Yes, with no difficulty in terms of volume. A volume of 3000 images daily averages approximately 0.035 images per second. Even assuming a conservative inference speed on a modest GPU like a T4, such a GPU could theoretically process 864,000 images per day (10 img/s × 3600 s/h × 24 h). Therefore, raw processing capacity for the required daily volume will not be a limiting factor, even with relatively modest hardware. The primary consideration will be latency per image if near real-time response is required.

Does it require high CPU, GPU, RAM resources?

GPU: It is highly recommended for inference, especially if low latency is needed or images are processed in batches. The specific type of GPU will directly impact speed and concurrent processing capacity. For the specified daily volume (2-3k), a low-to-mid-range data center GPU would probably be sufficient in terms of daily throughput, although latency per image will be higher than on high-end GPUs. For fine-tuning, a GPU with a significant amount of VRAM (video memory) is required, the exact amount depending on the size of the chosen RTMDet variant and the training batch size. GPUs with 16GB or more are recommended.

CPU: Necessary for general system orchestration, image pre-processing (loading, decoding, initial resizing), and post-processing of results (NMS, formatting). CPU load during inference will be moderate if primary computation is offloaded to the GPU. Performing full inference on the CPU is technically possible, but will result in very low speeds.

RAM (System Memory): An adequate amount of RAM is required to load the operating system, the deep learning framework, the RTMDet model itself, and input/output data. The exact amount will depend on the size of the model variant (RTMDet-tiny has ~4.8M parameters, RTMDet-s ~8.9M, RTMDet-m ~27.5M, RTMDet-l ~57M, RTMDet-x ~90M) and the inference/training batch size. GPUs have

their own dedicated VRAM, which is the most critical resource for model execution. Several tens of GB of system RAM are recommended for comfortable operation.

6.2.4. Licensing and Terms of Use

Is it open source or proprietary?

RTMDet is an Open Source model. Both its implementation source code and the pre-trained model weights are publicly available, primarily through GitHub repositories managed by OpenMMLab.

License type (Apache 2.0, MIT, commercial license, etc.).

The specific license under which RTMDet is distributed depends on the concrete framework or repository from which it is obtained:

MMDetection: The MMDetection toolbox, which includes an RTMDet implementation, is distributed under the Apache 2.0 license.

MMYOLO: The MMYOLO toolbox, also from OpenMMLab and offering RTMDet, uses the GPL-3.0 license.

Specific Variants: Some derived variants or implementations may have their own licenses or inherit MMDetection's. For example, the mentioned RTMDet-R2 repository uses Apache 2.0.

Fine-Tuning Examples: Repositories demonstrating fine-tuning, such as Makeability Lab's, may use permissive licenses like the MIT License for their example code.

Commercial Providers: Implementations like Qualcomm's have a dual license: Apache 2.0 for the original implementation and a specific Qualcomm license for compiled and optimized assets for their devices.

Does it allow commercial use?

Apache 2.0 (MMDetection): Yes, this license is permissive and allows commercial use, modification, distribution, and sublicensing of the software, subject to certain conditions such as retaining copyright and license notices, indicating changes made, and an explicit patent grant clause. It is generally considered suitable for the development of proprietary commercial products.

GPL-3.0 (MMYOLO): Yes, it allows commercial use, but with an important condition: it is a strong copyleft license. This means that any software that incorporates or derives from code or models obtained under GPL-3.0, and that is distributed, must be licensed in its entirety under the same GPL-3.0 license, and its complete source code must be made public. This obligation can be a significant

restriction for organizations wishing to develop and distribute proprietary commercial solutions based on RTMDet obtained from MMYOLO.

MIT (Fine-tuning example): Yes, it is a very permissive license that allows commercial use with minimal requirements, mainly maintaining the copyright notice and license in copies of the software.

The choice between obtaining RTMDet from MMDetection (Apache 2.0) or MMYOLO (GPL-3.0) has crucial legal and commercial implications. Given that the use case potentially involves creating a solution (e.g., filtering software for schools) that integrates the adapted RTMDet model, if this solution is to be distributed (either as installable software, a cloud service, etc.), the GPL-3.0 license of MMYOLO could force the entire solution to be open source. This may not be desirable or viable for the developer. In contrast, the Apache 2.0 license of MMDetection does not impose this copyleft obligation on software that uses it. Therefore, to develop a commercial or distributable solution that integrates RTMDet, using the implementation available under MMDetection (Apache 2.0) appears to be the preferable and legally less restrictive option. This is a fundamental consideration for development and business strategy.

Approximate cost:

Annual or monthly license: €0 for the open-source versions of RTMDet available under Apache 2.0, GPL-3.0, or MIT licenses. Licensing costs might exist if opting to use specific commercial versions offered by third parties or MLOps platforms that integrate RTMDet as a managed service (though no such offerings were identified).

Cost per use (if applicable, e.g., pay-per-API, etc.): Not directly applicable to the open-source model if deployed independently. Operational costs will be for infrastructure usage (primarily GPUs in the cloud or acquisition and maintenance cost of local hardware). If a third-party service offered RTMDet via an API, there would be usage costs, but this option does not seem to be standard for RTMDet.

Main Costs: The most significant costs associated with using RTMDet for this use case will be related to development (engineering hours for fine-tuning, integration into the final system, validation) and infrastructure (acquisition or rental of GPUs for training and inference, storage, networking). Additionally, as detailed in the following section, costs associated with regulatory compliance (GDPR) can be substantial.

6.2.5. Legal and Ethical Analysis

GDPR compliance and European regulations.

Applicability of GDPR: The EU General Data Protection Regulation (GDPR) is fully applicable to this use case. Images from European school environments will be processed, which are likely to contain personal data (facial images of students and staff, metadata allowing identification) and even special categories of personal data (Art. 9 GDPR) if the detected sensitive content reveals information about health, sexual orientation, ethnic origin, religious beliefs, etc. Furthermore, the processing directly affects minors, who benefit from specific protections under the GDPR.

Legal Basis for Processing (Art. 6 GDPR): It is mandatory to identify and document a valid legal basis for processing this personal data.

Consent (Art. 6(1)(a)) presents significant challenges. For minors, Art. 8 of the GDPR establishes specific requirements: consent is valid if the minor is at least 16 years old (or a lower age, not less than 13, if so provided by Member State law), and below that age, verifiable consent or authorization from the holder of parental responsibility is required. Obtaining and managing this consent granularly, informatively, freely, and revocably for all students potentially captured in images, and verifying parental authority, is logistically complex and may not be practical at scale.

Legitimate interest (Art. 6(1)(f)) of the data controller (e.g., the school or the service provider) to protect minors from harmful content is a potential alternative legal basis. However, its use requires passing a three-step balancing test (LIA - Legitimate Interest Assessment): 1) identify the legitimate interest, 2) demonstrate the necessity of processing to achieve that interest (justifying why less intrusive means, such as synthetic data or models not based on personal data, cannot be used), 3) weigh that interest against the fundamental rights and freedoms of data subjects (primarily, the right to privacy and data protection of minors). Given the sensitive context, this balance is delicate and must be rigorously documented.

Fundamental Principles of GDPR: All principles of Art. 5 GDPR must be respected, including: data minimization (processing only images and data strictly necessary for detection), purpose limitation (using data solely for the defined protection purpose and not for other purposes), accuracy (taking reasonable measures to ensure detections are correct, which is critical given the risks of false positives/negatives), storage limitation (not retaining data longer than necessary), integrity and confidentiality (implementing robust technical and organizational security measures), and proactive accountability (being able to demonstrate compliance).

Anonymization: It is highly unlikely that a model like RTMDet, especially after being fine-tuned with real data (even if superficially anonymized), can be considered truly anonymous according to the strict criteria of the EDPB. For data or the model to be considered anonymous (and thus outside the scope of GDPR), the risk of re-identification of any individual must be insignificant, considering "all means reasonably likely to be used." Extracting personal data from the model or its results must be highly improbable. Therefore, it is prudent to assume that GDPR will continue to apply throughout the model's lifecycle.

Data Protection Impact Assessment (DPIA): Given the nature of the processing (use of novel AI technology, large-scale processing, data of minors, potentially sensitive data, systematic monitoring), it is highly probable that a DPIA will be mandatory under Art. 35 of the GDPR. This assessment must identify and mitigate risks to the rights and freedoms of data subjects.

Transparency (Art. 13 and 14 GDPR): Clear, concise, and accessible information must be provided to data subjects (students, parents, staff) about the processing of their data, including the purpose, legal basis, data types, recipients, retention periods, and their rights.

Legality of Training: If the data used for fine-tuning the model was obtained or initially processed in violation of GDPR, this could compromise the legality of the subsequent use of the model trained with such data.

GDPR compliance emerges as one of the biggest challenges and potential blockers for implementing this system. The legal and administrative requirements (establishing legal basis, performing LIA/DPIA, ensuring security and transparency, managing rights) are complex and burdensome in this high-risk context (minors, sensitive data, monitoring). The costs associated with compliance (legal, consulting, technical, training) could be very significant, possibly exceeding the direct costs of the model and infrastructure. Project viability will largely depend on the organization's ability to navigate these requirements and demonstrate robust compliance.

Bias handling (racial, gender, cultural...).

Risk of Bias: AI models, including RTMDet, are susceptible to learning and perpetuating biases present in the data they are trained with. This applies both to pre-training on general datasets like COCO (which may have their own demographic or representational imbalances) and, more critically, to fine-tuning on the specific sensitive content dataset. If this latter dataset does not adequately reflect the diversity of the European school population (in terms of ethnicity, gender, culture, etc.) or the various ways sensitive content manifests in different contexts, the resulting model could exhibit uneven performance. This could

translate into disproportionately high error rates (false positives or false negatives) for certain groups, leading to unfair consequences such as excessive censorship of legitimate content from certain groups or failure to detect harmful content predominantly affecting others.

Mitigation Strategies: Documentation does not indicate that RTMDet or the MMDetection/MMYOLO frameworks natively incorporate specific mechanisms for active detection or mitigation of algorithmic biases. Therefore, addressing bias will require proactive effort external to the standard fine-tuning process:

Dataset Curation and Auditing: It is crucial to build the fine-tuning dataset carefully, actively seeking diversity and equitable representation of different demographic groups and relevant cultural contexts. Auditing the dataset to identify potential sources of bias before training is a crucial step.

Fairness-Aware ML Techniques: Research and, if possible, implement training techniques that promote fairness during the fine-tuning process. This could involve modifications to the loss function, data resampling, or specific regularizations (these techniques are not mentioned in the provided resources and would require additional research or custom development).

Disaggregated Evaluation: Evaluate the final model's performance not only with global metrics (like overall mAP) but also disaggregated for different demographic subgroups (if such information is available and its use is legally permissible). This allows for identifying and quantifying performance disparities.

Threshold Adjustment: Consider whether it is technically and legally feasible to adjust the model's decision thresholds differently for different groups or contexts, to balance error rates (although this can be complex and controversial).

Continuous Monitoring: Implement a system for monitoring model performance in production to detect drift or the emergence of new biases over time.

Ensuring that the sensitive content detection system is fair and equitable is not an automatic result of simply fine-tuning RTMDet. It requires a conscious and significant investment in data quality, potentially in advanced training techniques, and in rigorous and continuous evaluation. This effort adds technical and ethical complexity to the project.

Explainability capacity ("explainability") of decisions.

Importance of Explainability (XAI): Understanding why the RTMDet model has flagged a specific image as containing sensitive content is crucial for several reasons: user and administrator trust in the system; model debugging to understand and correct errors; auditing to verify its functioning and compliance; and potentially to meet regulatory requirements such as the right to obtain an

explanation in certain cases of automated decisions under the GDPR. XAI seeks to open the model's "black box."

Techniques Applicable to Object Detectors: Gradient-based or perturbation-based attribution methods, such as Grad-CAM, LayerCAM, and variants (GradCAM++, XGradCAM, ScoreCAM, EigenCAM, etc.), originally developed for classification, can be adapted to visualize which regions of an input image were most influential for the detector to localize and classify a specific object. These techniques generate "heatmaps" that highlight relevant areas.

Implementation with RTMDet/MMDetection:

There doesn't appear to be native support or official tutorials for XAI techniques within the MMDetection or MMYOLO frameworks, according to available information.

However, applying these techniques is feasible. Third-party implementations and tutorials demonstrate how to apply Grad-CAM to models within the MMDetection ecosystem.

Popular external libraries like pytorch-grad-cam offer a wide range of CAM methods and provide documentation on how to adapt them to different architectures, including suggestions for target layers for common models and explicit support for tasks like object detection.

Integration may require modifications to the MMDetection code or the inference flow to access intermediate layer activations and gradients needed to calculate CAM maps.

Specific Challenges: Applying XAI to object detectors like RTMDet is inherently more complex than applying it to simple classification models. A detector produces multiple outputs per image (several bounding boxes, each with a class and a confidence score). Therefore, the explanation must be generated for a specific detection (a particular box/class). Interpreting the resulting heatmaps also requires care to understand which aspect of the detection (localization, classification) is being explained.

In summary, although MMDetection/MMYOLO does not offer integrated XAI tools, it is technically possible to achieve explainability for the decisions of the fine-tuned RTMDet model. However, this will require additional development effort to integrate and adapt existing XAI libraries or implementations, adding another layer of technical work to the project.

6.2.6. Infrastructure Requirements

Can it be deployed locally or does it require the cloud?

RTMDet, being an open-source model, offers flexibility regarding the deployment environment:

- Local Deployment (On-Premise): It is entirely feasible to install and run RTMDet on your own servers physically located within the organization's or each school's infrastructure. This option requires the organization to directly manage the hardware (servers with GPUs), the software (operating system, Python/PyTorch/MMDetection dependencies, updates), and the physical and logical security of the infrastructure.
- Cloud Deployment: This is also a viable option. RTMDet can be deployed on virtual machines (VMs) or container services (like Kubernetes) offered by cloud providers (AWS, Azure, Google Cloud, etc.) that provide access to GPU-enabled instances. The cloud facilitates scalability (adjusting the number of GPUs according to demand), management of the underlying infrastructure, and offers additional services (monitoring, load balancing), but implies recurring operational costs based on resource consumption.

Is it feasible to have a server per school or a global server instead?

The choice between a distributed deployment model (one server per school) or a centralized one (a global or regional server) involves considering various technical, economic, management, and, crucially, privacy factors:

Server per School (Local/Edge Deployment):

- Advantages: Greater privacy and data control: Potentially sensitive images are processed locally and do not need to leave the school's physical environment or network, minimizing risks associated with data transfer and facilitating compliance with GDPR data residency requirements. Lower latency: Analysis is performed close to the source, which can be relevant if a quick response is required. Less dependence on Internet connectivity: The system can continue functioning even with temporary external network outages.
- Disadvantages: Higher initial hardware cost: Requires acquiring and installing GPU-capable hardware in each school. Distributed management and maintenance: Updating the model, software, or troubleshooting requires intervention at multiple locations, increasing operational complexity. Possible inconsistency: Maintaining the same model version and configuration across all schools can be a challenge. Local hardware viability: Although a modest GPU might be sufficient for the required daily

volume (2-3k images), the cost and logistics of equipping each school with such hardware and ensuring its proper functioning can be prohibitive.

Global Server (Centralized/Cloud Deployment):

- Advantages: Economies of scale in hardware: Investment in GPUs is concentrated in one or a few data centers, which is usually more economically efficient. Simplified management: Maintenance, model updates, and monitoring are centralized. Consistency: Ensures all requests are processed with the same model version and configuration. Easier scalability: It is easier to add or remove computing capacity (GPUs) in a centralized environment to adapt to changes in load.
- Disadvantages: Greater privacy implications (GDPR): Images must be transferred from schools to the central server. This introduces risks during transfer and requires robust security measures (encryption in transit and at rest). It also raises questions about data processing location (must remain within the EU or in countries with recognized adequacy) and increases the attack surface. Dependence on connectivity: Requires a reliable Internet connection with sufficient bandwidth from each school to the central server. Higher latency: The round-trip time for data to the central server introduces additional latency, although for offline analysis or prioritization, this may not be critical.

Recommendation and Key Considerations: Given the relatively low processing volume (2-3k images/day per school, which does not saturate even a modest GPU) and the high sensitivities related to GDPR (processing of minors' data, sensitive content), the choice of deployment model will be strongly influenced by the privacy risk assessment.

A centralized model (preferably in a private cloud or a secure data center within the EU) appears more operationally manageable and potentially more cost-effective at scale. However, it requires implementing extremely robust data protection measures for centralized transfer, storage, and processing, and a solid justification in the DPIA and LIA.

A local deployment per school is technically feasible (especially using lightweight RTMDet variants) and offers significant advantages from a privacy perspective by minimizing data transfer. However, the management burden and initial hardware cost could make it unfeasible for many institutions.

A hybrid solution, where very light pre-filtering is done locally (perhaps even on CPU or low-power GPU) to discard clearly unproblematic images, and only suspicious images are sent to the central server for deeper analysis, could represent an interesting compromise between privacy and efficiency.

The final decision should not be purely technical or economic but must prioritize the approach that best demonstrates compliance with GDPR principles and minimizes risks to the rights and freedoms of minors, as determined in the DPIA.

Estimated infrastructure consumption (CPU, GPU, RAM).

For Inference (per server instance):

GPU: This is the key component for acceptable performance. A data center GPU like the Nvidia T4 or equivalent is recommended as a minimum for a central server handling the aggregated volume of several schools (or for a local server per school). More powerful GPUs (V100, RTX 3090, A6000, A100) will offer significantly higher speed (lower latency) and capacity to process more images concurrently. The choice will depend on budget, latency requirements, and the total number of images to be processed per server.

VRAM (GPU Memory): The amount needed directly depends on the RTMDet variant used. Larger models like RTMDet-L (~57M parameters) or RTMDet-X (~90M parameters) will require several gigabytes of VRAM just to load the model weights, plus additional memory for intermediate activations during inference and for the image batch. A GPU with 16GB of VRAM (like the T4 or 16GB V100) should be sufficient for inference with reasonable batch sizes, but larger models might benefit from GPUs with more memory (24GB, 32GB, 40GB+).

CPU: A modern CPU with multiple cores is required to manage the operating system, network, data pre/post-processing, and coordination with the GPU. The specific load will depend on the efficiency of the data pipeline.

RAM (System Memory): An adequate amount of RAM is needed for the operating system, software libraries, loading data, and potentially caching. Several tens of gigabytes (e.g., 32GB, 64GB or more) per server are recommended, depending on the load and the number of concurrent processes.

For Fine-Tuning (required initially and for periodic updates):

GPU: One or more powerful GPUs with a large amount of VRAM (e.g., minimum 16GB, ideally 24GB like the RTX 3090/A6000, or 32GB/40GB+ like V100/A100) are needed to efficiently train larger RTMDet variants with appropriate batch sizes. This process can be performed in the cloud (dedicated training instances) or on local hardware if available.

CPU/RAM: Like inference, but the process of loading and pre-processing the entire training dataset may require additional CPU and RAM resources. Storage is also an important consideration for saving large training datasets.

6.2.7. Advantages and Limitations

Aspect	Advantages	Limitations
Accuracy	High accuracy demonstrated in generic object detection (COCO mAP ~52.8% for L/X variants). Offers an excellent balance between accuracy and parameter count compared to models like YOLO.	Specific accuracy for sensitive content detection is unknown and will critically depend on fine-tuning. Will likely struggle with context, ambiguity, and abstract concepts.
Cost	Base model is open-source with a free license (€0). Main costs stem from development (fine-tuning, integration) and necessary infrastructure.	Infrastructure costs (acquisition or rental of GPUs) can be significant. There are potentially high "hidden" costs associated with strict GDPR compliance. MMYOLO's GPL-3.0 license is restrictive.
Ease of Integration	MMDetection and MMYOLO frameworks are designed to facilitate fine-tuning on custom datasets. RTMDet's architecture was conceived to be extensible.	Requires technical expertise in the OpenMMLab ecosystem (PyTorch, MMDetection/MMYOLO) and dataset preparation. Fine-tuning for sensitive content requires considerable effort in data and validation.
Explainability	Standard XAI techniques (like Grad-CAM and variants) are theoretically applicable to visualize the reasons for a detection. Third-party implementations exist for MMDetection.	No native or official XAI support in MMDetection/MMYOLO. Requires custom integration and is more complex for object detectors than for classifiers.
Flexibility for Fine-Tuning	Offers a range of sizes, allowing users to balance accuracy, speed, and computational resources according to their needs. Pre-designed variants	Choosing the optimal variant requires careful analysis of the specific use case, target metrics (accuracy vs. latency), and available

	exist for segmentation and rotated detection.	infrastructure constraints.
Regulatory Compliance	The model itself is not intrinsically incompatible with GDPR. Deployment flexibility (local/cloud) allows for choosing architectures that mitigate privacy risks.	The use case is high-risk under GDPR. Establishing a legal basis is complex. Mandatory DPIA. Effective anonymization is very difficult. Compliance is a major challenge.

6.2.8. Conclusions and Recommendations

Is this model suitable for our use case?

RTMDet has the technical potential to be part of a sensitive content detection solution, thanks to its high speed, good accuracy in generic tasks, open-source nature, and adaptability.

However, it is NOT suitable for this purpose on its own. It requires significant adaptation (fine-tuning) and, more importantly, must be approached with considerable legal and ethical precautions. The biggest challenges for its successful implementation do not lie so much in the intrinsic technical capabilities of the RTMDet model, but rather in aspects related to strict GDPR compliance, the creation of an adequate and bias-free fine-tuning dataset, managing the contextual complexity of sensitive content, and the need for integration into a workflow that includes human oversight.

What conditions or adjustments would be necessary to use it?

Specific Fine-Tuning: It is essential to develop or acquire a large, diverse, representative, and specifically annotated image dataset for the sensitive content categories relevant to the European school context. The acquisition and use of this dataset must rigorously comply with GDPR. Perform the fine-tuning process on a pre-trained RTMDet variant (preferably obtained from MMDetection under an Apache 2.0 license to avoid distribution restrictions).

Rigorous GDPR Compliance: This is a non-negotiable requirement. It implies:

- Conduct a comprehensive DPIA to identify and mitigate risks.
- Establish and document a clear legal basis (likely Legitimate Interest, supported by a robust and defensible LIA).
- Strictly implement the principles of data minimization, purpose limitation, and storage limitation.
- Ensure high levels of technical and organizational security to protect data.
- Provide transparency to data subjects about processing.

- Establish procedures to manage data subject rights (access, rectification, erasure, etc.).
- The choice of deployment model (local vs. centralized) must prioritize privacy risk minimization.

Active Bias Mitigation: Conduct audits of the fine-tuning dataset and the resulting model to detect demographic or cultural biases. Implement mitigation strategies during or after training if significant biases are identified.

Explainability (XAI) Integration: Implement XAI tools (e.g., Grad-CAM based) to allow interpretation of model decisions, facilitating debugging, auditing, and trust-building.

Adequate Infrastructure: Deploy the model on servers equipped with appropriate GPUs. A T4 might be sufficient for the required daily volume, but benchmarks are essential to confirm actual latency and performance. Carefully select the deployment model (centralized seems more operationally viable but presents greater GDPR challenges).

Hybrid Workflow with Human Review: It is absolutely fundamental to design a system where RTMDet's detections are not the final decision. All positive detections (and potentially a random sample of negative ones) must be reviewed and validated by trained human personnel to interpret context, evaluate intent, and apply ethical and regulatory criteria specific to the school environment. RTMDet must act as a supportive tool, not as an autonomous judge.

6.2.9. Visual Summary

Category	**Result**
Model Type	Vision
Accuracy	High (on generic COCO tasks). Medium/Low (Estimated) on sensitive content post-fine-tuning; requires exhaustive validation. Critically dependent on data quality.
Cost	Low (License): €0 (Open Source). Medium/High (Total): Includes significant development (fine-tuning), infrastructure (GPU), and, crucially, GDPR compliance costs.
License	Apache 2.0 (via MMDetection, recommended) / GPL-3.0 (via MMYOLO, restrictive for

	distribution). Allows commercial use (with conditions depending on license).
Fine-Tuning	Yes. Very High difficulty (requires technical expertise in MMDetection/MMYOLO and a high-quality specific dataset).
GDPR Compliance	Partial/Complex. Achieving compliance is a major challenge due to the context (minors, sensitive data). Requires significant legal and technical effort.
Final Recommendation	Requires significant adaptation and rigorous validation. Use only as part of a hybrid system with mandatory human review. Consider alternatives (commercial APIs, non-AI approaches) if GDPR, data, and context challenges are insurmountable.

6.3 Combinations of Open Source AI Models for Image Recognition

6.3.1 Introduction: Why Combine Open Source Models?

The previous sections (6.1 on GluonCV and 6.2 on RTMDet) analyzed individual open-source models and tools for image recognition. While these options offer flexibility and avoid direct licensing costs, it has also become clear that a single model, even after fine-tuning, can have significant limitations for the complex task of detecting sensitive content in school environments. Models like RTMDet excel at fast and accurate object localization, but they are not inherently designed to classify the sensitive nature of a complex situation. On the other hand, models available in toolkits like GluonCV can offer good classification capabilities after adaptation but perhaps won't achieve the speed or detection accuracy of more specialized architectures like RTMDet for localizing specific objects.

Given these limitations, the strategy of combining multiple open-source models emerges. The main objective is to leverage the strengths of each component to build a more robust and accurate overall system than each part could be on its own.

To understand combination strategies, it's helpful to define two fundamental concepts:

- Pipeline: Refers to a chain of models where the output of one model becomes the input for the next. Each model in the sequence performs a specific task. Simple example: A first model could detect all faces in an image, and a second model would take only the regions of those faces to analyze their facial expressions and determine if they indicate distress or aggression. This approach allows for specialization at each step of the analysis.
- Ensemble: Consists of using several models that perform the same task independently on the same input (the same image). The individual results are then combined to obtain a final decision. Simple example: Showing the same image to three different sensitive content detectors and making a decision based on the majority (if two or more detectors agree) or by averaging the confidence scores assigned by each.

Combining models, whether through pipelines or ensembles, offers important potential benefits:

Improved Accuracy and Recall: Ensembles can reduce the probability of errors, as it's less likely that multiple independent models will make the same mistake simultaneously. Pipelines allow the use of highly specialized models for subtasks, improving accuracy at each step. This is crucial given the high sensitivity to errors (both false positives and false negatives) in sensitive content detection, and the need for accuracy imposed by GDPR.

Increased Robustness: A combined system can be less vulnerable to the specific weaknesses of a single model or unexpected variations in input data.

Addressing Complex Tasks: Multifaceted problems, such as detecting bullying which may involve recognizing objects, people, actions, and context, can be broken down into more manageable subtasks addressed by different models in a sequence (pipeline).

However, combining models also introduces significant challenges:

Increased Technical Complexity: Designing, training, deploying, and maintaining multiple interconnected models is considerably more difficult than managing a single model. It requires advanced skills in Machine Learning Engineering.

Higher Computational Needs: Running multiple models, whether in sequence or in parallel, increases the demand for computational resources (CPU, GPU, RAM) and potentially network bandwidth if models communicate with each other. This directly impacts infrastructure costs.

Dependency Management: Ensuring compatibility between models developed with different deep learning frameworks (e.g., PyTorch, MXNet, TensorFlow), library versions, and data formats can be a considerable technical challenge.

Accumulated Latency: In pipelines, the total processing time is the sum of the times of each stage. If any stage is slow, it can compromise the real-time responsiveness of the overall system, even if fast models like RTMDet are used in some part of the sequence.

It is fundamental to understand that the decision to combine models should not be seen merely as an optional optimization. For certain particularly complex and context-dependent sensitive content categories (for example, identifying ambiguous hate symbols, detecting subtle forms of bullying, or interpreting radicalization scenes), the inherent limitations of individual vision models to understand context and intent make it highly probable that a single model will fail. The inability of standard models to go beyond visual pattern recognition and capture deep semantic meaning is a fundamental barrier. In these cases, a combined architecture (such as a pipeline that first detects elements and then classifies them contextually, or an ensemble that integrates diverse perspectives) could be an architectural necessity to achieve a minimally acceptable performance level. The increase in complexity then becomes a necessary trade-off for potentially greater effectiveness in detecting the most difficult and harmful content.

6.3.2 Common Applicable Combination Strategies

There are various ways to combine open-source models to improve sensitive content detection. Below, some of the most relevant and applicable strategies in this context are described:

Specialized Pipeline (Detector + Classifier):

Description: This strategy uses a fast and efficient object detector model, such as an RTMDet variant (e.g., RTMDet-s or RTMDet-m), for a first pass over the image. Its function is to quickly identify regions of interest (ROIs) or the presence of specific objects that could be indicators of sensitive content (e.g., faces, potential weapons, known symbols, groups of people interacting).

Second Step: The ROIs identified by the detector are extracted and sent to one or more classifier models. These classifiers, which could be fine-tuned models from robust architectures available in toolkits like GluonCV or other suitable models, would be specifically trained to determine if the content within that ROI belongs to a specific sensitive category (e.g., a classifier trained to detect violence in interactions, another to identify specific hate symbols).

Advantage: Computational efficiency. More costly analysis (detailed classification) is only applied to the relevant parts of the image identified by the fast detector, saving resources compared to analyzing the entire image with a complex classifier.

Disadvantage: Overall performance critically depends on the initial detector's ability to find all relevant regions (high recall). If the detector fails to identify a region of interest, the classifier will never have the opportunity to analyze it (a false negative from the detector propagates). Also, it requires careful coordination between models (data format of ROIs, etc.).

Voting/Averaging Ensemble:

Description: Instead of a sequence, this strategy uses multiple models that perform the same task in parallel. Several instances of the same base architecture are trained (for example, several RTMDet-l detectors or several ResNet50 classifiers), but with slight variations in their training (e.g., using different data subsets, different random weight initializations, or slightly different hyperparameters).

Combination: All these models process the same input image independently. Their individual predictions are then combined to obtain a more robust final decision. For example, for classification, a majority vote can be used (the class predicted by the majority of models is chosen). If models provide confidence scores, these scores can be averaged for each possible sensitive category.

Advantage: Can significantly improve robustness and reduce prediction variance. It's less likely that all models, having been trained slightly differently, will make exactly the same error on the same image.

Disadvantage: High computational cost. Multiple full models need to be run for each image, which multiplies hardware requirements and inference time. Furthermore, if the model's base architecture has fundamental limitations in capturing certain types of sensitive content, an ensemble of that same architecture might not overcome those limitations.

Hierarchical Filtering:

Description: This approach uses a cascade of models with increasing complexity. It starts with a very fast and lightweight model (could be a highly optimized simple classifier, or even a filter based on basic heuristic rules) whose sole function is to quickly discard images that are clearly not sensitive with high probability.

Second Level: Only those images flagged by the initial filter as potentially sensitive, or on which the filter has low confidence, are passed to a more powerful and computationally expensive model or pipeline (like the Detector + Classifier pipeline described above) for detailed analysis.

Advantage: Optimizes computational resource usage by significantly reducing the load on heavier models. Most computational effort is concentrated on the most difficult or ambiguous cases.

Disadvantage: The success of this strategy critically depends on the accuracy of the initial filter. This filter must have a very low false negative rate (i.e., it must be very good at identifying potentially sensitive content and should rarely mistakenly discard an image that does contain it). An unreliable initial filter would compromise the entire system.

The choice of the appropriate combination strategy has no single answer. The decision will depend on a careful evaluation of project priorities: is latency (response speed) more critical, or the highest possible accuracy? What are the priority sensitive content categories, and what is their nature (concrete objects vs. complex scenes)? What budget and computational resources are available? Pipelines can be more efficient, but they create dependencies and potential bottlenecks. Ensembles can be more robust, but at the cost of higher computational expenditure. Hierarchical filtering seeks a balance, but its effectiveness depends on the first filter. Given the diversity of sensitive content to be detected, it is plausible that different strategies will be optimal for different content types. Therefore, selection requires a detailed analysis of these trade-offs in the specific context of the project's needs and constraints.

6.3.3 Key Practical Considerations for Implementing Combinations

Implementing a solution based on combining open-source models goes beyond selecting an architectural strategy. It requires addressing a series of critical practical considerations that can significantly impact project viability, cost, and risk:

License Compatibility:

Problem: A combined system integrates multiple software components: frameworks, pre-trained models, auxiliary libraries. Each of these components may have its own software license. While major frameworks like GluonCV and OpenMMLab tools for RTMDet are typically distributed under permissive licenses like Apache 2.0, this doesn't always apply to the entire ecosystem. In particular, the pre-trained weights of certain models, or optimized implementations of models like those offered by Qualcomm for RTMDet, might be distributed under different, potentially more restrictive licenses (e.g., for non-commercial use only, or requiring specific agreements).

Required Action: It is absolutely essential to perform a meticulous and documented verification of the licenses of each and every component planned for use in the combination (source code, model weights, dependent libraries). This is

crucial to ensure that the final system can be legally used for the intended commercial purpose in European school environments. Ignoring this step can lead to serious legal risks.

Technical Interoperability:

Challenge: Open-source models may be developed using different deep learning frameworks (for example, a GluonCV model might use MXNet or PyTorch, while RTMDet in MMDetection/MMYOLO uses PyTorch). Ensuring that these models can communicate efficiently with each other within a pipeline is a technical challenge. It is necessary to standardize how data (images, tensors, bounding box coordinates, confidence scores) is passed from one stage to the next.

Possible Solutions: One option is to use neutral model exchange formats like ONNX (Open Neural Network Exchange), which allows converting models between different frameworks. Another is to develop custom data connectors or adapters. A third approach is to try to standardize all development within a single framework (for example, PyTorch, given that both GluonCV and MMDetection support it). Any of these solutions adds an additional development and integration workload.

Resource Management (Amplified):

Impact: As mentioned, running multiple models substantially increases computational resource consumption (CPU, GPU, RAM). This has a direct impact on infrastructure choice (local servers vs. cloud, hardware specifications) and ongoing operational costs. The need for powerful GPUs, recommended for optimal performance of individual models, becomes even more critical and expensive in combined systems.

Necessary Evaluation: It's not enough to estimate resources for each model separately. It's essential to perform load testing and profiling of the entire combined system under realistic conditions to identify actual hardware requirements, detect potential bottlenecks (e.g., a slow pipeline stage, data I/O), and adequately size the infrastructure.

Complex Training and Validation:

Necessity: It's not enough to validate each model individually. It is crucial to validate the performance of the entire combination as an integrated system. Training strategies can also become more complex. For example, in a detector-classifier pipeline: is the classifier trained using the (potentially imperfect) outputs of an already fixed detector, or is an attempt made to train the entire pipeline end-to-end (which is technically more challenging)?

Data: Validation requires comprehensive datasets that allow evaluating how models interact. For example, the validation set must include difficult cases for the

detector to see how the subsequent classifier behaves, or ambiguous examples to evaluate the robustness of an ensemble. The already considerable challenge of obtaining ethical, legal, and representative training data is amplified, as adequate data is needed to train and validate each component and the interaction between them.

Maintenance and Lifecycle:

Additional Burden: Maintaining a system composed of multiple open-source models is more complex than maintaining a single one. Updating a component (e.g., a new framework version, a re-trained model) may require re-validating the entire system, or even re-training other components if interfaces or behaviors change. Managing different model versions, their library dependencies, and monitoring possible performance regressions over time becomes an ongoing and demanding task. The risk associated with a possible reduction in maintenance activity for projects like GluonCV becomes more significant if that component is a critical part of a larger combined system.

In essence, opting to combine open-source models not only adds to the inherent challenges of each individual component but compounds them and adds new layers of complexity. The risks already identified for models like GluonCV or RTMDet (related to the need for data, the difficulty of fine-tuning, the lack of native explainability or bias mitigation tools, and the burden of ensuring legal and ethical compliance) multiply. To these are added new risks arising from the interaction between components: interoperability issues, accumulated latency, complex end-to-end validation, and management of multiple licenses. For example, ensuring GDPR compliance requires a detailed analysis of data flow and processing across the entire chain of models, not just one. Verifying license compatibility involves auditing all involved software artifacts. Therefore, the decision to implement a combination of models must carefully weigh the potential performance benefits against this significant increase in technical complexity, management effort, and the overall project risk profile.

6.4 Conclusions and Recommendations on Open Source Models

6.4.1 Synthesis of Open Source Model Evaluation

This chapter has deeply analyzed open-source options for image recognition applied to sensitive content detection. A generalist toolkit (GluonCV, section 6.1), a family of specialized detectors (RTMDet, section 6.2), and strategies for combining these or other open-source models (section 6.3) were evaluated. Key findings are summarized below:

Recap of Individual Models:

GluonCV: Presents itself as a versatile toolkit offering a wide range of pre-trained model architectures and tools for research and development. Its main strength lies in this flexibility and documented fine-tuning capabilities. However, its weaknesses are significant for this use case: it lacks native specialization in sensitive content, requires a very considerable and expert adaptation effort (fine-tuning), its understanding of visual context is limited, there is a possible reduction in its active maintenance, and it does not include integrated functionalities for regulatory compliance (GDPR), bias mitigation, or explainability (XAI). Although the source code often has a permissive license (Apache 2.0), the Total Cost of Ownership (TCO) associated with its effective implementation is high due to data, expertise, infrastructure, and compliance costs.

RTMDet: Corresponds to a family of object detector models designed to be efficient and fast (real-time). Its primary strength is speed and accuracy in object localization. Its key weaknesses are that it primarily acts as a detector (requiring additional components for sensitive content classification), it also needs intensive fine-tuning to adapt to the specific domain, its understanding of context is very limited, there is a risk of license fragmentation (especially for pre-trained weights or optimized third-party variants), and it also offers no native compliance, ethical, or XAI tools. The base code license is usually Apache 2.0, but it is crucial to verify it for each variant and weight. The TCO is also high.

Summary of Combinations (6.3):

Opportunities: Combining models (e.g., RTMDet + GluonCV-based classifier pipeline, or ensembles) offers the theoretical possibility of improving the accuracy and robustness of the overall system, leveraging the complementary strengths of different architectures. For the most complex and context-dependent sensitive content categories, this might be the only way to achieve acceptable performance with open-source technology.

Challenges: This strategy dramatically increases technical complexity (design, integration, training, validation), computational resource requirements, maintenance effort, and, critically, compounds the risks associated with each individual component (licenses, dependencies, regulatory compliance).

Emphasis on Critical Common Challenges:

Regardless of whether an adapted individual model or a combination is chosen, several cross-cutting and critical challenges emerge when considering open-source solutions for this project:

Mandatory and Expert Adaptation: None of the analyzed open-source options work "out of the box" for sensitive content detection. Extensive, complex fine-tuning, performed by highly qualified Machine Learning personnel, is essential.

Critical Dependence on Adequate Data: The success of any adapted AI model (whether open-source or proprietary) absolutely and fundamentally depends on the availability of training and validation datasets that are large, representative of the school environment, high-quality, and, crucially, obtained and used ethically and legally compliant with GDPR, especially considering that they involve minors' data. This is, by far, the biggest obstacle and risk of the adaptation approach.

Inherent Limitations of Contextual Understanding: Purely visual models, even after fine-tuning, have intrinsic difficulties in interpreting the intent, nuances, and social or cultural context necessary to correctly classify many sensitive content categories (e.g., bullying, radicalization, irony vs. hate). This implies that there will always be a ceiling on the accuracy achievable with image analysis alone.

Total Burden of Compliance and Ethics: When using open-source tools, the entire responsibility for ensuring GDPR compliance (performing DPIAs, establishing legal bases, implementing security, ensuring transparency, managing Art. 22 rights on automated decisions), mitigating algorithmic biases, and providing system explainability (XAI) falls entirely on the implementing organization. These functionalities are not built into open-source toolkits or models.

The analysis reveals a fundamental trade-off inherent in using open source in such a sensitive context. Technical flexibility and code control are gained, and direct software license costs are avoided. However, this freedom comes at the cost of assuming immense responsibility in implementation and exposure to considerable risks (technical, legal, ethical). The burden of building, validating, deploying, maintaining, and, above all, ensuring the legal, ethical, and fair operation of the system rests entirely with the organization. This includes the Herculean task of obtaining the necessary data and the external implementation of fairness and explainability mechanisms. The initial appeal of "zero" license cost is misleading; the Total Cost of Ownership (TCO) and the risks associated with the open-source

path are, in reality, very high. This transforms the challenge from vendor management to the need for deep internal technical capability and a very robust AI governance framework.

Comparative Summary Table of Open Source Approaches:

The following table summarizes the key characteristics of the different open-source approaches:

Feature	**GluonCV (Toolkit - Fine-tuned Model)**	**RTMDet (Fine-tuned Detector + Classifier)**	**Combined Models (e.g., Pipeline/Ensemble)**
Main Strength	Flexibility (architectures)	Speed/Detection Accuracy	Potential for Higher Accuracy/Robustness
Main Weakness	Generic, needs massive adaptation	Primarily detection, needs classification	High Complexity and Resource Needs
Fine-Tuning Effort	Very High (Data, Expertise)	Very High (Data, Expertise)	Extremely High (Compounded)
Context Understanding	Limited	Very Limited	Limited (inherent visual limitation)
Data Dependence	Critical (Major Obstacle)	Critical (Major Obstacle)	Critical (Amplified Obstacle)
Licensing	Apache 2.0 (code), verify weights	Apache 2.0 (code), verify weights	Complex (Multiple components)
Compliance/XAI Tools	None Native	None Native	None Native

Infrastructure Needs	High (GPU recommended)	High (GPU recommended)	Very High (Compounded)
TCO / Risk Profile	High	High	Very High

This table provides a concise comparison of the characteristics, strengths, weaknesses, and requirements associated with the different open-source strategies discussed. It helps to quickly visualize the trade-offs in terms of effort, cost, risk, and potential performance, directly supporting the strategic recommendations presented below.

6.4.2 Strategic Recommendations for the Project

Based on the exhaustive analysis of individual and combined open-source models, and considering the highly sensitive context (minors, European schools, delicate content) and the strict regulatory framework (GDPR, EU AI Act), the following strategic recommendations are proposed:

Adopt a Cautious Approach Towards Open Source:

Justification: The extreme challenges related to ethical and legal data acquisition, the mandatory need for deep and expert fine-tuning, inherent limitations in visual context understanding, and the full burden of ensuring GDPR/AI Act compliance and ethical implementation (bias mitigation, explainability), make pursuing a purely open-source route a very high-risk and very high-effort path for this specific application.

Action: It is recommended to proceed with an exclusively open-source solution only if the organization possesses, or is willing to invest significantly to acquire, deep and dedicated internal expertise in Machine Learning, considerable resources for data management and rigorous validation, and a robust governance framework for ethical AI and legal compliance.

Prioritize Data Strategy Above All Else:

Justification: The success of any adapted AI model (whether open-source or proprietary) critically and fundamentally depends on the quality, representativeness, and, above all, the legality and ethics in the acquisition and use of training data. This aspect is absolutely non-negotiable under GDPR when dealing with minors' data.

Action: Before committing to any specific open-source model or combination, it is imperative to develop a detailed and legally validated strategy for the acquisition, annotation, secure storage, and management of the necessary data. If obtaining

adequate and compliant data proves unfeasible, the viability of any custom-trained AI solution is highly questionable. Data minimization techniques should be considered from the design phase.

Require Human Oversight:

Justification: Given the limitations in contextual understanding by visual models, the inherent risk of algorithmic biases, the impossibility of achieving perfect accuracy, and the serious consequences of errors (both false positives and false negatives) in a school environment, relying solely on automated decisions is unacceptable from an ethical and risk perspective. Furthermore, Article 22 of the GDPR likely requires the possibility of human intervention in automated decisions that produce legal effects or similarly significantly affect the data subject.

Action: Any implemented system (open-source or not) must be mandatorily designed with a clear, efficient, and well-defined workflow for human review of content flagged as potentially sensitive before any definitive action is taken. This is crucial to ensure accuracy, fairness, accountability, and regulatory compliance.

If Opting for Open Source, Start Simple and Validate Rigorously:

Justification: Combining models adds significant complexity on multiple fronts. It is more prudent and manageable to first establish a performance baseline with a single well-chosen and carefully adapted model before attempting more complex architectures.

Action: If choosing the open-source path, it is recommended to start by fine-tuning a single promising model (for example, an RTMDet variant for detection coupled with a simple classifier, or a robust classifier from the GluonCV repertoire). Perform rigorous validation focused on key metrics such as recall for sensitive content (minimizing false negatives) and fairness metrics to evaluate biases across different relevant demographic groups. Only consider implementing more complex combinations if the performance of the single model, after exhaustive optimization and validation, proves to be clearly insufficient for the project's minimum requirements.

Ensure a Comprehensive Legal and Ethical Compliance Framework:

Justification: The legal (GDPR, EU AI Act - which will likely classify these systems in schools as 'high-risk') and ethical implications are extremely high. Open-source tools do not provide built-in safeguards in this area.

Action: It is mandatory to conduct a Data Protection Impact Assessment (DPIA). An explicit and documented legal basis for data processing must be ensured. Implement robust technical and organizational security measures. Guarantee transparency to users (students, parents, school staff). Develop clear policies and

active processes for monitoring and mitigating biases. Integrate external XAI tools if necessary for auditability, debugging, and trust-building. Continuous advice from legal and ethical experts specialized in data protection, children's rights, and responsible AI is indispensable throughout the project lifecycle.

These recommendations focus on mitigating the most critical risks (legal, ethical, data) and establishing fundamental prerequisites (human oversight, robust governance) before solely focusing on technical performance optimization. Given the high-risk nature of the context (European schools, minors, sensitive content) and the fact that open-source solutions shift all responsibility for managing these risks to the implementer, this prioritized approach is essential. The technical feasibility of fine-tuning, although challenging, is potentially solvable. However, a failure in legal compliance, ethical implementation, or data strategy can not only derail the project but also cause significant harm. Therefore, strategic decisions regarding the open-source path must primarily address these fundamental risks.

7. Analysis of Proprietary AI Models for Text Recognition

7.1 Analysis of the Google Cloud Natural Language API Model

7.1.1. Model Identification

- Model Name: Google Cloud Natural Language API
- Model Type: Text Processing / Natural Language Processing (NLP). It functions as a collection of pre-trained models exposed via API endpoints.
- Provider: Google

7.1.2. General Model Description

What tasks does the model currently perform?

The Google Cloud Natural Language API offers a suite of pre-trained natural language processing functionalities, accessible via REST or RPC endpoints. These capabilities allow developers to integrate language understanding into their applications. Key functions include:

- Sentiment Analysis: Determines the overall emotional tone (positive, negative, neutral) of a text, providing a score (from -1.0 to +1.0) and a magnitude (intensity of emotion).
- Entity Analysis: Identifies and classifies named entities in text, such as people, organizations, locations, events, products, etc. Assigns prominence

scores indicating the entity's importance in the overall text. Can link entities to knowledge bases like Wikipedia if available.

- Entity Sentiment Analysis: Combines the above two capabilities to determine the sentiment expressed towards specific entities within the text.
- Syntax Analysis: Breaks down text into sentences and tokens (words, punctuation marks), identifies parts of speech (e.g., noun, verb), lemmas (root form of the word), and creates syntactic dependency trees to understand the grammatical structure of each sentence.
- Content Classification: Uses the classifyText method to categorize text documents within a predefined taxonomy of over 700 general and specific categories (e.g., /Arts & Entertainment, /Health, /Sensitive Subjects). Version 2 of classification returns more specific categories compared to V1. A minimum of approximately 20 words is required for reliable classification.
- Text Moderation: Employs the moderateText method to specifically classify text into harmful and sensitive content categories. These categories include: Toxic, Derogatory, Violent, Sexual, Insult, Profane, Death/Harm/Tragedy, Firearms, Public Safety, Health, Religion, Illicit Drugs, War, Finance, and Politics & Legal. This functionality is based on Google's PaLM 2 model.

This range of functions establishes the baseline of what the API can directly offer. The moderateText function is the closest to the objective, but its predefined categories are general safety attributes and may not adequately capture the specific forms of bullying, grooming, or radicalization sought.

What tasks is it not specialized in and would require adaptation?

The standard Google Cloud Natural Language API does not provide pre-trained models explicitly designed to detect the nuanced and contextual forms of bullying, specific types of threats (veiled or direct), grooming behaviors (conversation patterns for child sexual abuse), or radicalization narratives that are relevant in school environments and across multiple languages.

While the moderateText function covers related areas such as Toxic, Insulting, Sexual, or Violent content, it is highly likely to lack the specificity and contextual understanding needed for reliable detection of these complex phenomena. For example, distinguishing between playful insults and systematic harassment, identifying grooming tactics disguised as friendly conversation, or recognizing specific keywords and ideologies associated with radicalization requires a level of specialization that goes beyond these general categories.

Similarly, the standard classifyText categories (like /Sensitive Subjects) are too broad and not tailored to these specific risks.

Therefore, to achieve the desired detection capabilities, adaptation would be indispensable. This would involve training custom models (fine-tuning) or developing sophisticated prompting strategies, using platforms like Google Cloud's Vertex AI. The need for this adaptation indicates a significant increase in project complexity and cost compared to simply integrating a pre-trained API, as it requires delving into the Vertex AI ecosystem for customization.

Even with fine-tuning, what can it not do or what are its limitations?

Even with customization through fine-tuning, there are inherent limitations of current NLP technology that must be considered:

- Limits of Contextual Understanding: NLP models, even fine-tuned, can struggle with deep context, sarcasm, rapidly evolving slang, cultural nuances specific to different European languages and regions, and distinguishing between intent and literal meaning. It remains very difficult to detect when an adult is trying to gain a minor's trust with bad intentions or when a person begins to adopt extremist ideas, especially if they use disguised language or if it happens gradually over a long period.
- Data Dependence: The effectiveness of fine-tuning critically depends on the quality, quantity, and representativeness of labeled training data for each specific sensitive category (bullying, grooming, etc.) in the multiple required languages. Acquiring such data, especially that involving minors, presents significant ethical and practical challenges.
- New Threats (Zero-Day): Fine-tuned models are trained on historical data. They may struggle to identify entirely new forms of harmful language, emerging slang, or radicalization narratives that were not present in the training set. Continuous retraining would be necessary to maintain effectiveness.
- Over-reliance on Keywords: Models might rely too heavily on specific keywords, which could lead to false positives (flagging benign content) or false negatives (missing harmful content that avoids those keywords).
- Multilingual Nuances: Achieving high and consistent performance across multiple European languages simultaneously, especially for culturally specific phenomena like certain types of harassment or radicalization, is inherently difficult even with fine-tuning. The limited support for some languages in the moderation function adds complexity. This multilingual requirement complicates both the use of pre-trained models and the fine-tuning process, as it requires high-quality labeled datasets in each language and for each sensitive category, which is resource-intensive.

- Limits of Explainability: While tools like Vertex Explainable AI exist to understand model decisions, it is still difficult to explain why a language model classified certain sensitive content in a specific way. This can reduce trust in the model and complicate decisions based on its results.

It is fundamental to set realistic expectations for the capabilities of current NLP technology, acknowledging its inherent limitations, especially for tasks as complex, evolving, and sensitive as risk detection in school environments.

Foreseen use cases (future specialization in sensitive content detection).

The main foreseen use case is the adaptation of Google Cloud's NLP capabilities (either through thresholds in moderateText or, more likely, through custom models fine-tuned in Vertex AI) to automatically scan text communications within school platforms (e.g., chats, forums, document submissions). The objective is to detect potential instances of bullying, threats, grooming, and radicalization almost in real-time [User Query Objective].

The system would flag potentially harmful content for review by human moderators or designated school personnel, to enhance safety and allow timely intervention. Multilingual support is a critical requirement for deployment in Europe's diverse school populations.

7.1.3. Technical Capabilities

Accuracy and recall (if available).

Standard API: No specific precision and recall figures are provided in the research fragments for the standard classifyText or moderateText methods applied directly to the target sensitive content (bullying, grooming, etc.). Google generally does not publish these metrics for general-purpose APIs, as performance varies significantly depending on the specific text domain and task.

moderateText Confidence Scores: The API returns confidence scores (between 0.00 and 1.00) for its safety attributes. These scores indicate the probability of belonging to a category, not the severity of the content. It is crucial for users to define their own confidence thresholds based on exhaustive testing with their own data, balancing their tolerance for false positives and false negatives. Google explicitly warns against relying solely on these scores for reliability or accuracy in critical decisions. Therefore, the lack of published precision/recall metrics for the specific task means that self-evaluation is mandatory.

Related Benchmarks (Proxy): A comparative study of the Healthcare Natural Language API (a related but distinct service) showed high precision (99%) and recall (93%) for medical entity and relationship extraction, outperforming AWS and Azure in that specific test. This suggests that Google's underlying NLP technology

can be highly accurate for specialized tasks when properly trained. However, these figures are not directly transferable to the general NL API or the specific task of sensitive content detection in schools.

Custom Models (AutoML/Vertex AI): Performance metrics (precision, recall, F1 score, confusion matrix) are generated after training a custom model using user-provided data on the Vertex AI platform. Actual performance will depend entirely on the quality and quantity of the training data provided and the inherent complexity of the detection task. Vertex AI includes tools for evaluating these custom models.

Data Limitation: The fragments indicate an inability to access specific URLs that might have contained more detailed performance data.

Processing speed (images/second, words/second, etc.).

API Latency: As a cloud-based API, processing speed includes both network latency and Google's processing time. No specific words-per-second metrics for the NL API are provided in the available documentation. Latency can be affected by various factors, such as input text length, specific features requested (e.g., an annotateText call invoking multiple analyses will likely be slower than a single function call), the geographical location of the server processing the request, and network conditions. Benchmarks from other Google Cloud APIs (such as Text-to-Speech) suggest typical latencies in the range of hundreds of milliseconds, although this is only indicative. General cloud storage API latency can be low (tens of ms) but can increase considerably (hundreds of ms) depending on factors like authentication. Google Cloud offers tools like the Performance Dashboard to monitor latency between regions.

Request Quotas: The API has default usage quotas: 600 requests per minute and 800,000 requests per day per project. These quotas suggest that the underlying infrastructure is designed to handle significant throughput. It is possible to request an increase in these quotas if necessary.

Processing Units: Pricing is based on character units (1000 characters for most functions, 100 for moderation). This implies that processing time might scale to some extent with text length, but the dominant factor for user experience is likely to be the combined network and API call latency.

In summary, while exact speed figures are not available, the API operates within the typical latency ranges of cloud services. Quotas indicate sufficient capacity for the target volume, but real-world latency testing from end-user geographical locations is recommended.

Fine-Tuning Capability:

Does it allow Fine-Tuning?

Yes, but customization is not done directly on the standard NL API's pre-trained models. It requires using Google's platform to build custom models, which is currently Vertex AI.

Gemini Tuning: Vertex AI allows supervised fine-tuning of Gemini models (like Gemini 2.0 Flash) using user-provided labeled datasets, preferably in JSONL format. Parameter-efficient tuning (or adapter tuning), which is more resource-efficient than full fine-tuning of all model parameters, is often used.

How difficult or costly is it?

Difficulty: Considerably more complex than simply using the pre-trained API. It involves several stages:

Data Preparation: Collecting, cleaning, and precisely labeling a sufficient dataset (more than 100-500 examples are recommended) in JSONL format. This is often the most challenging part, especially for sensitive and multilingual data. Requires careful splitting into training, validation, and test sets.

Using the Vertex AI Platform: Understanding and utilizing Vertex AI services, including Datasets, Training Jobs, Model Registry, and Endpoints. Although AutoML aimed to be "no-code," Gemini tuning generally involves using the SDK (Python) or the console, which requires some technical expertise.

Hyperparameter Tuning: Potentially adjusting parameters like the number of epochs, learning rate, or adapter size to obtain optimal results. Vertex AI Vizier can help in this process.

Evaluation: Correctly interpreting the evaluation metrics (precision, recall, etc.) provided by Vertex AI after training.

Cost: Involves multiple potential cost components that go beyond NL API usage:

Training on Vertex AI: Costs associated with computational resources (GPU/TPU) used during the tuning process. User reports suggest this can be significant, even for small datasets or short durations, potentially reaching hundreds of dollars per tuning job. Pricing can be based on tokens processed during training (dataset size * epochs) or on compute hours. Gemini tuning has specific pricing, for example, per million training tokens. Currently, the tuning service for Gemini 1.5 Flash is listed as free, but costs apply for tokens used during inference of the tuned model.

Deploying Endpoints on Vertex AI: If the tuned model is deployed for real-time predictions, there are costs associated with endpoint nodes (e.g., per node-hour).

Prediction Calls on Vertex AI: Costs for making predictions against the deployed custom model (e.g., per character or token, potentially with different rates than the standard API).

Cloud Storage: Costs for storing datasets and model artifacts.

General: Customization through Vertex AI represents a substantial potential increase in total cost compared to exclusive use of the pre-trained NL API. It is crucial to perform careful cost estimation using the GCP Pricing Calculator and monitor usage. The greatest cost uncertainty lies not in basic API usage, but in the potential need for customization via Vertex AI, which lacks fixed and transparent pricing for training and has reports of high variability. This implies that the actual cost of implementing the desired solution could be substantially higher and more difficult to accurately estimate than basic API costs, representing a significant budgeting risk factor.

Workload:

Can it process between 2000 and 3000 units (images/texts) daily?

Yes. The default request quota is 600 requests per minute and 800,000 requests per day. Processing 3000 texts per day is well below these limits (equivalent to approximately 2 requests per minute on average). The API is designed to be scalable.

Does it require high CPU, GPU, RAM resources?

Not directly by the user for standard API calls. The Google Cloud Natural Language API is a managed service (SaaS/API). Google manages the underlying infrastructure. Users interact via API calls and do not need to provision or manage specific CPU/GPU/RAM resources for the API itself. Resource consumption occurs on Google's infrastructure and is abstracted, reflected in the pay-per-use price.

However, if Vertex AI is used for fine-tuning and deploying custom models, resource selection (e.g., machine types for training, number of nodes for endpoints) and associated costs do become relevant and are managed (and paid for) by the user. Although the API is "serverless" from the user's perspective (no servers are managed for API calls), achieving the required functionality (custom sensitive content detection) will likely necessitate Vertex AI, introducing infrastructure considerations (training resources, deployment nodes) and associated costs.

7.1.4. Licensing and Terms of Use

Is it open source or proprietary?

Proprietary

License type (Apache 2.0, MIT, commercial license, etc.).

Commercial. Use is governed by the Google Cloud Platform Terms of Service and any associated service-specific terms, including data processing agreements.

Does it allow commercial use?

Yes, the API is designed to be integrated into commercial applications, subject to Google Cloud's terms. Resale of the API service itself is prohibited.

Approximate cost:

Pay-as-you-go model based on usage volume, measured in "units" of Unicode characters processed per month. Different API features have different unit sizes and tiered pricing levels.

Pricing Units:

Text Moderation: 100-character units.

Entity Analysis, Sentiment, Syntax, Entity Sentiment: 1,000-character units.

Content Classification: 1,000-character units.

- Free Tiers (Monthly):

Entity Analysis, Sentiment, Syntax, Entity Sentiment: First 5,000 units free.

Content Classification: First 30,000 units free.

Text Moderation: First 50,000 units free.

Paid Tiers (Example - Content Classification per 1,000 characters):

- 30,001 - 250,000 units: $0.0020 / unit.
- 250,001 - 5,000,000 units: $0.00050 / unit.
- More than 5,000,000 units: $0.0001 / unit.

Paid Tiers (Example - Text Moderation per 100 characters):

- 50,001 - 10,000,000 units: $0.0005 / unit.
- 10,000,001 - 50,000,000 units: $0.00025 / unit.
- More than 50,000,000 units: $0.000125 / unit.

Estimated Cost for 3000 texts/day:

- Assumptions: Average text length = 500 characters (needs validation with real use case data). Combined use of Content Classification and Text Moderation.
- Daily Characters: 3000 texts * 500 characters/text = 1,500,000 characters.
- Monthly Characters: 1,500,000 * 30 days = 45,000,000 characters.
- Content Classification Units (1000 characters): 45,000,000 / 1000 = 45,000 units/month.
- Cost: (45,000 - 30,000 free units) * $0.0020/unit = 15,000 * $0.0020 = $30 / month.
- Text Moderation Units (100 characters): 45,000,000 / 100 = 450,000 units/month.
- Cost: (450,000 - 50,000 free units) * $0.0005/unit = 400,000 * $0.0005 = $200 / month.
- Estimated Total API Cost (Classification + Moderation): Approximately $230 / month. This estimate heavily depends on the actual text length, exact features used, and volume processed.

Additional Potential Costs:

- Infrastructure used to call the API (e.g., Cloud Functions, App Engine, GKE).
- Vertex AI costs if custom models are trained/deployed (training compute, endpoint hosting, prediction calls), which can be significant.

Cloud Storage costs for data storage.

annotateText cost: Billed as the sum of costs for each requested feature within the single call.

The pay-per-use model offers flexibility but requires careful monitoring, especially if text volume or average length fluctuates. The greatest cost uncertainty does not lie in basic API usage, but in the potential need for customization via Vertex AI, which lacks fixed and transparent pricing for training and has reports of high variability. This implies that the actual cost of implementing the desired solution could be substantially higher and more difficult to accurately estimate than basic API costs, representing a significant budgeting risk factor.

7.1.5. Legal and Ethical Analysis

GDPR compliance and European regulations.

Google Cloud Commitment: Google Cloud affirms its commitment to GDPR compliance across all its services. They offer a Cloud Data Processing Addendum (CDPA), which incorporates Standard Contractual Clauses (SCCs) to meet the security, contractual, and data transfer requirements of EU, UK, and Swiss data

protection laws. Customers may need to formally accept the CDPA via the Cloud Console.

Data Processing: Google acts as a data processor, processing customer data according to customer instructions, as defined in the DPA. The customer (the school or educational platform) is the data controller.

Data Use: Google states that it will not use customer content submitted to the Cloud Natural Language API to train its general models, nor will it make it public or share it, except as necessary to provide the service. For the standard NL API, data is processed in memory and not permanently stored, although metadata (like request time and size) may be temporarily logged. In the case of fine-tuning in Vertex AI, Google asserts that tuning data belongs to the customer and is used to create an adaptive layer in the customer's instance, not to train Google's base LLMs. This distinction is crucial: sensitive customer data is indeed used to train a model, even if it's for the customer's exclusive use.

Data Location: Although Google Cloud has global infrastructure, customers may need to configure services, use specific features (like Assured Workloads), or select specific regions to meet data residency requirements, if applicable. Vertex AI tuning jobs could offload compute to other US or EU regions. Processing locations for Vertex AI are documented.

Specific Considerations for Minors' Data: Processing minors' data in schools under GDPR requires special attention to the legal bases for processing (e.g., consent, legitimate interest, public task), data minimization, purpose limitation, robust security measures, and, most likely, conducting Data Protection Impact Assessments (DPIAs). Although Google provides the platform with the necessary contractual and technical safeguards, the responsibility for ensuring the legality of processing school data rests with the customer (the school or platform provider). The use of AI for monitoring purposes raises specific ethical and legal questions about necessity, proportionality, and children's rights (privacy, freedom of expression).

EU Cloud Code of Conduct: Google Cloud adheres to the EU Cloud Code of Conduct, which can help demonstrate GDPR compliance for data processors. The Natural Language API and AutoML Natural Language are listed as covered services.

In essence, while Google Cloud provides a GDPR-compliant contractual and technical framework, the customer assumes ultimate responsibility for ensuring that their specific use case (processing sensitive minors' data in schools using AI) is legal, ethical, and complies with all applicable local and European regulations. Given the nature of the processing (automated monitoring of minors), a

comprehensive DPIA conducted by the customer will almost certainly be required before any deployment.

Bias handling (racial, gender, cultural...).

Google's Position: Google acknowledges the risks of bias in AI and sets principles to avoid unfair biases. They advocate for fairness, ensuring models treat all users fairly, which requires careful data selection, model evaluation, and continuous monitoring.

Bias Potential: Pre-trained models like those in the NL API are trained on large datasets (potentially web data), which can reflect existing societal biases. This could lead to the model performing differently for texts from different demographic groups or misinterpreting culturally specific language. Sensitive content detection models could disproportionately flag content from certain groups if biases are not carefully evaluated and mitigated.

Mitigation Tools/Techniques:

- Data Selection/Augmentation: Carefully curate or augment training data for custom models to ensure adequate representation across groups.
- Model Remediation: Techniques exist, often within the TensorFlow ecosystem (usable with custom training in Vertex AI), such as MinDiff (balances errors across different data subgroups) and Counterfactual Logit Pairing (ensures changing a sensitive attribute doesn't alter prediction). Vertex AI provides some tools to explore fairness issues.
- Evaluation: It is crucial to evaluate model performance across different demographic subgroups (if identifiable and their use is permissible).

Responsibility: While Google provides tools and principles, the responsibility for identifying and mitigating biases specific to the school context and target languages rests with the organization implementing the solution. This requires continuous effort and domain expertise. Bias is a significant risk, especially in sensitive applications, and requires proactive auditing, data management, and the potential implementation of specific mitigation techniques during customization.

Explainability capacity ("explainability") of decisions.

Need for Explainability: Understanding why a model flagged a piece of text as sensitive is crucial for trust, manual review, appeal processes, and improving the model itself.

Vertex Explainable AI: Google Cloud offers Vertex Explainable AI. It primarily provides:

- Feature Attributions: Shows how much each input feature (e.g., a word or token) contributed to the final prediction. Methods like Integrated Gradients, Sampled Shapley, and XRAI are available. It is compatible with various model types (AutoML, custom TensorFlow, scikit-learn, XGBoost) and modalities, including text.
- Example-Based Explanations: Finds examples from the training set that are most similar to the input instance being explained. Primarily for TensorFlow models that can provide embeddings.

Applicability to NL API / Tuned Models:

- Standard NL API: Explainable AI is generally applied to models trained or deployed in Vertex AI. It is unclear if/how it can be directly applied to the pre-built, "black-box" endpoints of the NL API. The API returns results (categories, scores) but no inherent explanations of why those results were generated.
- Custom Models (Vertex AI): Explainable AI can be configured for custom models (including those potentially fine-tuned from Gemini, though specific support needs verification) when deployed to Vertex AI Endpoints. This requires specific configuration during model training or deployment.

Limitations: Explaining complex NLP models remains a challenge. Feature attributions for text can highlight influential words but may not fully capture complex syntax or semantic relationships that led to a classification. Explainability for newer, larger models (like Gemini) might still be evolving. Confirmation is needed on specific support for fine-tuned Gemini models.

Explainability is partially achievable if custom models deployed on Vertex AI are used, primarily through feature attributions. Direct explanation for standard NL API calls is likely not available. The effectiveness and ease of implementation for specific fine-tuned models require investigation. Furthermore, effective explainability is not an "out-of-the-box" feature but requires deliberate effort, expertise, and careful interpretation, which could limit its practical implementation.

7.1.6. Infrastructure Requirements

Can it be deployed locally or does it require the cloud?

It requires the cloud, as it is a Google Cloud service.

Is it feasible to have a server per school or a global server instead?

A global cloud server managed by Google Cloud is the most viable architecture, as the API is accessed over the Internet.

Estimated infrastructure consumption (CPU, GPU, RAM).

For NL API Calls: Minimum direct consumption by the user. As a serverless API, Google manages the compute resources. The user's infrastructure only needs to handle making HTTPS requests and processing JSON responses. This typically requires minimal CPU/RAM on the client side (e.g., a web server, a Cloud Function).

For Customization with Vertex AI (Training): May require significant GPU or TPU resources, managed by Vertex AI but selected and paid for by the user. Resource needs depend on model size, dataset size, and training duration.

For Customization with Vertex AI (Deployment/Prediction): Requires deploying models to Vertex AI Endpoints, which use underlying virtual machine nodes. The user selects machine types (CPU/GPU options available) and the number of nodes, which impacts cost and performance. Resource consumption depends on prediction traffic volume and latency requirements.

The shift to Vertex AI for customization fundamentally alters the infrastructure management paradigm, moving from purely serverless API calls to managing (and paying for) dedicated training and prediction resources in the cloud. This requires the user to be involved in infrastructure selection, scaling configuration, and cost management related to these resources, which is a change from the simpler, fully abstracted model of simply calling the pre-trained API.

7.1.7. Advantages and Limitations

Aspect	**Advantages**	**Limitations**
Accuracy	Provides moderateText for related categories with confidence scores.	Low/Medium (estimated) for specific bullying/grooming/radicalization detection; requires threshold tuning/definition; no published precision/recall for exact task; unreliable confidence.
Cost	Pay-per-use, free tiers available, potentially Low/Medium for moderate volumes.	Scales with volume; can become significant with high usage. Potentially High due to Vertex AI training/deployment costs; less predictable than API costs.
Ease of Integration	Easy via standard REST/RPC APIs and client libraries.	Requires technical expertise for advanced specialization (fine-tuning).
Explainability	Vertex Explainable AI available for custom models.	Low/Medium; primarily feature attributions; limited/uncertain for standard API. Tuned Gemini;

		practical implementation requires effort.
Flexibility for Fine-Tuning	High via Vertex AI Gemini tuning; allows specialization.	Requires significant data, effort, costs; tied to Vertex AI/Gemini ecosystem.
Regulatory Compliance	Google Cloud is committed to GDPR compliance. COPPA-compliant services for minor protection.	The user is responsible for correct implementation and data handling to ensure full compliance. Adequate consent is required for processing minors' data.

7.1.8. Conclusions and Recommendations

Is this model suitable for our use case?

The standard Google Cloud Natural Language API provides relevant but insufficient capabilities for the specific use case as presented. The moderateText function offers a starting point for detecting potentially harmful content, but it is unlikely to achieve the specificity and precision needed to reliably identify nuanced forms of bullying, grooming, and radicalization in a multilingual school context.

The suitability of the service is therefore conditional and depends on successful adaptation via Vertex AI, specifically through fine-tuning of models like Gemini. If the organization possesses the necessary resources (high-quality and specific data, ML expertise, adequate budget) to carry out this customization and can effectively address the complex legal and ethical challenges associated with it, then the platform could be adapted to meet the requirements. However, if the intention is to rely solely on the standard pre-trained API, it is likely not suitable for achieving the detection objectives with the required reliability.

What conditions or adjustments would be necessary to use it?

To effectively and responsibly use Google Cloud technology for this purpose, the following conditions and adjustments are required:

- Mandatory Customization: Significant investment in developing custom models using Vertex AI Gemini tuning is indispensable. Standard NL API capabilities cannot be relied upon for this specific task.
- Data Acquisition and Preparation: The collection, cleaning, and precise labeling of high-quality training data, representative of the European school environment and covering each sensitive category (bullying, grooming, threats, radicalization) in all target languages, is a critical and highly challenging step.
- Rigorous and Continuous Evaluation: Exhaustive testing and validation of the custom model's performance (precision, recall, fairness) must be

carried out using data specific to the school context and evaluating behavior across different languages and, if possible and permissible, demographic subgroups. Given the evolving nature of language and threats, continuous monitoring and periodic retraining will be necessary.

- Threshold Definition (if applicable): If moderateText is used as a complementary or initial tool, careful definition and validation of confidence thresholds are necessary to balance false positives and negatives.
- Robust Legal and Ethical Framework: Development and implementation of clear usage policies, data governance, and privacy. Mandatory conduct of a Data Protection Impact Assessment (DPIA) before deployment. Ensuring strict compliance with GDPR and local laws, especially concerning the processing of minors' data and the justification for monitoring. Establishment of transparent processes for human review of flagged content and for user appeals. Active implementation of bias mitigation strategies.
- Technical and Domain Expertise: Having personnel or collaborators with skills in machine learning, the Vertex AI platform, sensitive data handling, and preferably with knowledge in linguistics, psychology, or social sciences relevant to understanding the nuances of sensitive content in the school context.

7.1.9. Visual Summary

Category	**Result**
Model Type	Text (NLP API) / Requires Custom Model (Vertex AI Gemini)
Accuracy	Low (standard API for specific task) / Potentially High (with extensive fine-tuning and quality data)
Cost	Medium (API) + Potentially High (Vertex AI for fine-tuning/deployment) - $/month variable based on usage and customization.
License	Proprietary (commercial use allowed)
Fine-Tuning	Yes (via Vertex AI Gemini Tuning); Difficulty: High (requires data, expertise, ML management)
GDPR Compliance	Yes (Platform and DPA/SCCs provided by Google) / Customer Responsibility (Legal/ethical implementation, especially minors' data).
Final Recommendation	Requires adaptation (significant customization via Vertex AI is indispensable); Not recommended for direct use without adaptation.

7.2 Analysis of the Amazon Comprehend Model

7.2.1. Model Identification

- Model Name: Amazon Comprehend (Service encompassing multiple APIs and capabilities, not a single versioned model. Relevant features like DetectToxicContent, DetectPiiEntities, Custom Classification, and Custom Entity Recognition will be analyzed).
- Model Type: Text (Natural Language Processing - NLP).
- Provider: Amazon Web Services (AWS)

7.2.2. General Model Description

What tasks does the model currently perform?

Amazon Comprehend is a managed Natural Language Processing (NLP) service by AWS that extracts insights and meaning from text. Its current capabilities include a variety of fundamental NLP tasks, such as Entity Recognition (identifying people, places, organizations, dates, etc.), Key Phrase Extraction, Dominant Language Detection, Sentiment Analysis (determining if the tone is positive, negative, neutral, or mixed, both at a general level and directed at specific entities), Syntax Analysis (identifying parts of speech like nouns, verbs, adjectives), and Topic Modeling (grouping documents by common themes).

In addition to these general functions, Comprehend offers specific features geared towards "Trust and Safety." These include Personally Identifiable Information (PII) Detection, which locates data like addresses, phone numbers, or bank accounts. It also provides a Toxicity Detection API (DetectToxicContent) that classifies textual content into seven specific categories of potentially harmful content: sexual harassment, hate speech, threats/violence, abuse, profanity, insults, and explicit graphic content. Additionally, it offers a Prompt Safety Classification, designed to evaluate the safety of text inputs directed at large language models (LLMs).

A key capability of Comprehend is customization. It allows users to train custom models for specific tasks not covered by standard pre-trained models. This is achieved through Custom Classification (for categorizing texts according to user-defined labels) and Custom Entity Recognition (for identifying user-defined types of entities specific to the user's domain).

The service can process text in UTF-8 format and, for certain functionalities like custom models, can also ingest text extracted from image files (JPG, PNG, TIFF), PDFs, and Word documents.

What tasks is it not specialized in and would require adaptation?

Amazon Comprehend does not offer pre-built functionalities to directly detect the complex and nuanced forms of sensitive content that are critical in school environments, such as "bullying," "grooming" (online child sexual abuse), or "radicalization." These phenomena often involve subtle, indirect, coded language and depend heavily on context, intent, and relationships between interlocutors—aspects that go beyond the scope of standard entity, sentiment, or toxicity analysis offered by the service natively.

Therefore, detecting these specific types of sensitive content would require significant adaptation of the service. The primary avenue for this adaptation would be to use Comprehend's customization capabilities:

- Custom Classification: Train a model to classify text segments according to categories specifically defined for the school environment, such as "Possible Bullying," "Grooming Indicator," "Radicalization Content," "Threat," etc.
- Custom Entity Recognition: Train a model to identify specific terms, patterns, usernames, or entities (like URLs, phone numbers) that are frequently associated with identified risks in the school environment.

Both approaches demand considerable effort from the user, primarily in collecting, preparing, and annotating large volumes of high-quality training data that are representative of real communications in the European school environment and in the relevant languages. Furthermore, although Comprehend's core functions support multiple European languages, capturing the specific cultural and linguistic nuances of sensitive content in each language poses an additional challenge. Generic pre-trained models may not be sufficient, and training effective custom models will require specific and high-quality data for each language and cultural context.

Even with fine-tuning, what can it not do or what are its limitations?

Even with customization capabilities (which AWS refers to as "custom training" rather than "fine-tuning" of an accessible base model), Amazon Comprehend presents significant limitations for the proposed use case:

- Critical Language Limitation in Toxicity: The pre-trained DetectToxicContent API, which is the closest functionality to harmful content detection, only supports English. This is a fundamental barrier to its direct application in the multilingual environment of European schools. Using translation services like Amazon Translate as a preceding step introduces additional complexity, costs, potential latency, and, more importantly, the risk of

losing crucial linguistic nuances or introducing translation errors that affect detection accuracy.

- Language Limitations in Custom Models: Language restrictions also exist in customization features. Training custom entity recognizers using annotated PDF files is limited only to English. Similarly, training custom "native document" classifiers (which directly process formats like PDF/Word) is also restricted to English. While custom models trained with plain text (CSV or one-line-per-document format) support several European languages (German, Spanish, French, Italian, Portuguese), creating effective models to detect sensitive content still depends on the availability of high-quality and sufficiently voluminous annotated training data in each of the target languages.
- Deep Contextual Understanding: NLP models, even customized, can struggle to correctly interpret sarcasm, irony, coded language, rapidly evolving slang, and complex relational dynamics that are common in bullying, grooming, or radicalization situations. Achieving high accuracy in these scenarios often requires more sophisticated approaches than standard classification or entity recognition and represents an inherent challenge for current NLP technology.
- Limited Explainability: Amazon Comprehend provides confidence scores for its predictions, but it lacks advanced explainability functionalities (like SHAP or LIME values) that would allow understanding why the model has made a certain classification. This lack of transparency is a significant limitation in a school context, where it is crucial to be able to review, understand, and potentially appeal automated decisions about sensitive content.
- Bias Management: Machine learning models can inherit and amplify biases present in their training data. Comprehend does not include specific integrated tools for active detection or mitigation of demographic biases (racial, gender, cultural). Although AWS promotes responsible AI and offers tools like SageMaker Clarify for models in SageMaker, these do not integrate directly with Comprehend. Reports of false positives in toxicity detection suggest that accuracy problems or biases may exist in practice, which is especially concerning for applications in diverse educational environments.

Foreseen use cases (future specialization in sensitive content detection).

The main foreseen use case for Amazon Comprehend in this context is its adaptation via Custom Classification to identify texts likely containing indicators of bullying, threats, grooming, or language associated with radicalization within school communication platforms (such as emails, chats, forums, or shared documents).

Another relevant use case is leveraging PII Detection to locate and, potentially, redact (hide or mask) sensitive personal data of students before further analysis or information storage, which can be a useful tool to aid GDPR compliance.

Additionally, the use of Custom Entity Recognition could be explored to flag specific patterns, keywords, usernames, or entities (like URLs, phone numbers) that are frequently associated with identified risk activities in the school environment.

It is fundamental to understand that the viability of these use cases does not lie so much in Comprehend's ready-to-use capabilities, but in the implementing organization's ability to develop and maintain a robust data strategy. Success will depend on the ability to collect, annotate, and manage high-quality, multilingual training data specific to the nuances of online risks in the European school context. The burden of specialization falls significantly on the user.

7.2.3. Technical Capabilities

Accuracy and recall (if available).

AWS does not generally publish specific precision and recall metrics for Amazon Comprehend's standard APIs in its official documentation. Instead, the service provides confidence scores (numerical values between 0 and 1) for the detections made (entities, sentiment, toxicity, PII, custom classifications). These scores indicate the model's level of certainty in its prediction. Users are expected to define appropriate thresholds on these scores based on the sensitivity of their use case and their tolerance for false positives or false negatives.

For custom models (Classification or Entity Recognition), Amazon Comprehend does calculate and display performance metrics (Precision, Recall, F1-score) after the training process is complete. However, these metrics are calculated using the test dataset that the user themselves provides during training setup. Therefore, they reflect the model's performance exclusively on that specific dataset and do not guarantee the same precision or recall on real-world data, which may differ in characteristics or distribution.

Regarding Toxicity Detection, there are no official precision/recall figures published by AWS. Nevertheless, academic research and user reports suggest variable performance and potential issues:

- One comparative study of several moderation APIs assigned Comprehend a performance score of 0.74, trailing OpenAI (0.83) and Azure (0.81) on certain hate detection benchmarks. Another study observed that Comprehend achieved very high performance (92.2% ROC AUC) on the

well-known Jigsaw dataset, which might indicate that this dataset was part of its training data, but its performance was lower on other datasets.

- Significant false positives have been reported for the "SEXUAL" label in entirely innocuous texts, including examples related to the school environment. This indicates potential accuracy issues, at least for that specific category, and underscores the need for careful validation.

For PII (Personally Identifiable Information) Detection, AWS indicates that it internally evaluates accuracy using standard metrics (precision, recall, F1), but does not publish general figures. Actual performance will depend on the type of PII entity and the text context. Given the absence of public and independent benchmarks for the specific task of detecting bullying, grooming, or radicalization in multiple languages using Comprehend, it is imperative that any organization considering its use performs its own exhaustive and rigorous evaluation before productive deployment. Relying solely on the confidence scores provided by the API or on custom model training metrics is insufficient for critical applications where precision is paramount.

Processing speed (images/second, words/second, etc.).

Processing speed in Amazon Comprehend varies depending on the mode of operation:

- Synchronous APIs (Real-Time): These are designed to offer low latency in the analysis of relatively small texts. Size limits are typically 5 KB or 10 KB per document, depending on the specific API. AWS does not specify a speed in words or characters per second for these standard APIs, as they are subject to a "throttling" system that dynamically adjusts request capacity based on service load.
- Asynchronous APIs (Batch): These are optimized for processing large volumes of stored documents (e.g., in Amazon S3) and are not designed for real-time speed. Total processing time will depend on the dataset size, the number of documents, and the complexity of the requested analysis.
- Custom Endpoints (for real-time custom models): When using custom Classification or Entity Recognition models in synchronous mode, speed is directly determined by the number of Inference Units (IU) the user provisions for the endpoint. Each IU provides a maximum throughput of 100 characters per second. Users must calculate how many IUs they need based on the expected load (average text length and number of requests per second) and provision the corresponding capacity, which directly impacts cost.

The 100 characters/second per IU limit for custom endpoints have important implications. Processing moderate-length texts (e.g., chat messages or short emails of a few hundred characters) in near real-time could require provisioning multiple IUs, especially if peak loads or multiple concurrent requests need to be handled. This directly links the desired processing speed to the operational cost of the service, as each IU incurs a cost per second while active.

Fine-Tuning Capability:

Does it allow Fine-Tuning?

Yes, Amazon Comprehend allows for the specialization of its capabilities through custom model training. The two main avenues are:

- Custom Classification: Allows training models that classify entire texts according to a set of user-defined labels (e.g., "Bullying," "Threat," "Safe Content").
- Custom Entity Recognition: Allows training models that identify and extract specific types of entities (words or phrases) relevant to the user's domain (e.g., "Offensive Slang Term," "Risky Platform Name," "Radicalization Indicator"). It is important to note that AWS uses the term "custom training." This is not "fine-tuning" in the technical sense of taking a pre-trained base model (like BERT) and directly adjusting its weights with proprietary data. In Comprehend, the user provides the data, and AWS manages the process of training a specific model for that task within its infrastructure, using AutoML (Automated Machine Learning) techniques to simplify the process.

How difficult or costly is it?

Difficulty: The service is designed to be accessible to users without deep machine learning expertise. The main difficulty lies not in operating the AWS platform, but in preparing the training data. Creating a high-quality, correctly annotated dataset (either by document-level labels for classification or by identifying specific entities within the text) and a sufficiently large one is a laborious and critical process for success. The difficulty multiplies in a multilingual scenario and for detecting concepts as subtle as harassment or grooming. AWS provides documentation on the required data formats. Recently, AWS reduced the minimum data requirements for custom entity recognition in plain text (now starting from 25 annotations per entity and 3 documents), which facilitates initial experimentation. However, achieving robust performance will likely require considerably larger datasets. Requirements for training with semi-structured documents (PDF) remain higher (100 annotations/entity, 250 documents).

Cost: The cost of customization has several components:

Training: Billed at $3 per compute hour used to train the model, with per-second billing. Training duration depends on dataset size and model complexity.

Model Storage: Each custom model incurs a cost of $0.50 per month for its storage.

Inference (Real-Time Endpoints): If the custom model needs to be used for real-time analysis, an endpoint must be provisioned. The cost is $0.0005 per Inference Unit (IU) per second (with a minimum of 60 seconds per activation). Each IU offers a throughput of 100 characters/second. This cost is incurred continuously while the endpoint is active and is typically the most significant component of operational cost for real-time applications.

Inference (Asynchronous Batch): Using custom models in asynchronous jobs follows a pricing structure based on the volume of processed data (similar to standard asynchronous APIs).

Workload:

Can it process between 2000 and 3000 units (images/texts) daily?

Yes, this daily processing volume is feasible with Amazon Comprehend, provided that the appropriate infrastructure is configured. Viability depends on the chosen processing mode:

Asynchronous Processing (Batch): This volume is easily handled. A single asynchronous job can process up to 1 million documents or 5 GB of data. Processing 3000 texts per day (approximately 1.1 million per year) is within limits and suitable for analyses that do not require immediate response.

Synchronous Processing (Real-Time with Custom Endpoints): Also, feasible but requires careful planning of endpoint provisioning. 3000 texts per day equates to an average of approximately 2.1 texts per minute. If peak loads are higher or texts are long, multiple IUs will be needed. For example, if an average text has 300 characters and a peak of 10 texts per minute (0.17 texts/second) is expected, the calculation of necessary IUs would be: (10 texts/min * 300 chars/text) / (60 sec/min * 100 chars/sec/IU) = 0.5 IUs. In this case, provisioning 1 IU might be sufficient for the average load, but peaks might require more capacity or the use of auto-scaling (available for endpoints) to dynamically adjust IUs based on demand.

Does it require high CPU, GPU, RAM resources?

Amazon Comprehend is a fully managed service by AWS, meaning that the underlying hardware (CPU, GPU, RAM) is provisioned and managed by AWS and is abstracted from the end-user. Users interact with the service through API calls or

the AWS console, without needing to manage servers or computational resources directly. Resource consumption is indirectly reflected through the service costs:

- API Usage Costs: Per-unit text processing fees for standard APIs implicitly reflect the compute used by AWS.
- Custom Endpoint Costs: The number of Inference Units (IUs) provisioned for real-time inference is directly related to the computational resources AWS allocates for that endpoint. More IUs mean greater resource allocation by AWS to maintain performance, and thus a higher cost.
- Custom Model Training Costs: The $3 per training hour fee also reflects the intensive resource usage during that model creation phase.

7.2.4. Licensing and Terms of Use

Is it open source or proprietary?

Proprietary

License type (Apache 2.0, MIT, commercial license, etc.).

Use of Amazon Comprehend is governed by the AWS Service Terms, and specifically by the conditions applicable to AWS AI Services. Use of the service implies acceptance of these contractual terms.

A key aspect of these terms is AWS's use of customer data. The terms state that AWS may store and use text inputs processed by Comprehend to provide and maintain the service, as well as to develop and improve the quality of Comprehend and other Amazon machine learning and artificial intelligence technologies. However, customers have the option to opt out of this use of their content for service improvement by configuring an opt-out policy at the AWS Organizations level.

Does it allow commercial use?

Yes. The AWS Service Terms permit the use of its services, including Comprehend, for commercial purposes, provided these terms are met. The service is widely used in commercial applications for customer analysis, document processing, content moderation, and other business purposes.

Approximate cost:

Annual or monthly license: There is no licensing fee as such. The pricing model is pay-per-use. However, there is a recurring monthly cost for storing custom models, set at $0.50 per custom model per month.

Cost per use (if applicable, e.g., pay-per-API, etc.): Yes, based primarily on the amount of text processed and resources used:

API Calls (Standard Functions): Billed per "unit," where 1 unit equals 100 characters. There is a minimum charge per synchronous request (e.g., 3 units or 300 characters for Toxicity Detection). Prices are tiered, decreasing the cost per unit as monthly volume increases. Example prices for the first tier (up to 10 million units/month):

Entity Recognition, Sentiment Analysis, Key Phrase Extraction, Language Detection, PII Detection: $0.0001 per unit.

Syntax Analysis: $0.00005 per unit.

Toxicity Detection: Follows the per-unit cost pattern for synchronous requests, probably like Entities/Sentiment.

Asynchronous Jobs (Batch): Cost is usually based on the total size of processed documents (measured in MB, for example, for Topic Modeling) or, depending on the job type, might follow a per-unit pricing similar to synchronous calls.

Custom Model Training: $3 per compute hour (billed per second).

Custom Model Endpoints (Real-Time Inference): $0.0005 per Inference Unit (IU) per second (minimum 60 seconds). Each IU provides 100 characters/second of throughput. This is a continuous cost as long as the endpoint is active.

Free Tier: AWS offers a free tier for the first 12 months from the first request to Comprehend. It includes 50,000 text units (5 million characters) per month for various APIs (Key Phrase Extraction, Sentiment, Entities, Language Detection, PII, Toxicity, Syntax, Prompt Safety).

The cost model implies that for large data volumes where latency is not critical, asynchronous batch processing tends to be more economical. Conversely, if real-time detection using custom models is required, the continuous cost of endpoints can quickly become significant. A single IU, running 24/7, would cost approximately $1300 per month (calculated from $0.0005/second). Therefore, a careful evaluation of latency requirements versus budget is essential when choosing between synchronous and asynchronous processing for custom models.

7.2.5. Legal and Ethical Analysis

GDPR compliance and European regulations.

AWS provides a general framework designed to comply with GDPR. This includes offering a Data Processing Addendum (DPA) that incorporates the EU-approved Standard Contractual Clauses (SCCs) as a mechanism for legal data transfer if

necessary. When customers use services like Amazon Comprehend to process personal data (such as texts from school communications), AWS acts as a data processor under GDPR terminology.

However, GDPR compliance operates under a shared responsibility model. AWS is responsible for the security of the cloud (the physical infrastructure and base services), but the customer is responsible for security in the cloud. This means that the organization using Comprehend (the school or educational service provider) is responsible for:

- Region Choice: Selecting the AWS region(s) where data will be processed and stored, ensuring compliance with GDPR data residency requirements and applicable local laws. It is crucial to choose regions within the EU if data is required not to leave the European Economic Area.
- Implementation of Technical and Organizational Controls: Properly configuring access controls (using AWS IAM), applying encryption to data both in transit (SSL/TLS) and at rest (using AWS KMS), and enabling detailed activity logging (using AWS CloudTrail).
- Data Lifecycle Management: Defining clear policies for the minimization, retention, and secure deletion of processed data.
- Data Subject Rights Management: Establishing procedures to respond to requests from data subjects (students, parents, staff) regarding their rights of access, rectification, erasure, etc., as required by GDPR.
- Consideration of AWS Data Use: Evaluating whether AWS's potential use of processed content for service improvement is compatible with their obligations under GDPR for the specific data being processed. If not, the opt-out policy must be actively implemented through AWS Organizations.

The protection of minors' data is an especially critical aspect in the school context. GDPR (Article 8) imposes strict requirements for processing children's data, generally requiring verifiable parental consent below a certain age (which may vary slightly among EU Member States). Although AWS service terms mention the customer's responsibility to comply with COPPA (the US child privacy law), the responsibility to comply with specific GDPR requirements for minors rests entirely with the customer. Comprehend's PII detection functionality can be an auxiliary tool to identify certain sensitive data, but it does not in itself guarantee compliance. The organization must ensure that necessary consents are obtained, clearly inform about processing, and apply reinforced protection measures. In summary, achieving GDPR compliance when using Comprehend in European schools demands significant diligence on the part of the implementing organization. It is not enough to rely on the compliance of AWS's infrastructure. Robust data governance policies, appropriate consent management, and rigorous

implementation of technical and organizational controls adapted to the sensitivity of minors' data are required.

Bias handling (racial, gender, cultural...).

Amazon Comprehend, in its current form, does not offer specific integrated tools to detect, measure, or mitigate demographic biases in its pre-trained or custom models. The main indicator it provides are confidence scores, which measure model certainty but do not inform about the fairness of its predictions across different groups.

Like all machine learning models, Comprehend's models are susceptible to reflecting and potentially amplifying biases present in the data they were trained with. This represents a significant ethical risk, especially in sensitive applications like content moderation in diverse educational environments.

Reported problems with the Toxicity Detection API, such as false positives in the "SEXUAL" category for innocuous texts, and findings from academic research suggesting uneven performance (over- or under-moderation) for content related to certain identity groups (specifically mentioning LGBTQIA+, Black, Jewish, and Muslim individuals in one study that evaluated various APIs, including Comprehend), point to the real existence of fairness or bias issues in practice.

AWS does offer a tool called SageMaker Clarify designed for bias detection and model explainability. However, Clarify is integrated with the Amazon SageMaker ecosystem and is intended for analyzing models trained or deployed within SageMaker. It cannot be directly applied to the endpoints of Comprehend's pre-trained or custom models. Using Clarify would require a more complex workflow, such as exporting Comprehend's predictions for offline analysis or, alternatively, building and training a completely custom model within SageMaker instead of using Comprehend.

The absence of integrated tools for bias management in Comprehend constitutes a considerable ethical risk for its deployment in European schools, which are inherently diverse environments. It would be indispensable to implement a proactive and separate strategy for bias evaluation, using diverse test datasets and, possibly, external methodologies or tools. This additional effort adds complexity and cost to the project. The documented problems with the Toxicity API reinforce this concern, indicating that the service's fairness cannot be assumed without rigorous verification.

Explainability capacity ("explainability") of decisions.

Amazon Comprehend's ability to explain its decisions is limited. The service provides confidence scores associated with each prediction (for example, the probability that an entity is of type "PERSON," that the sentiment is "NEGATIVE," that a text belongs to the custom category "BULLYING," or that it contains "INSULT" toxicity). These scores indicate the model's degree of certainty but do not explain the factors or parts of the input text that led to that conclusion.

Comprehend lacks built-in deep explainability methods.

This low explainability represents a significant disadvantage for the school context. In an environment where transparency, accountability, and the possibility of reviewing and appealing automated decisions (such as flagging a message as inappropriate) are fundamental, the inability to understand the "why" of a decision is problematic. It hinders model debugging, identifying failures (including biases), improving the system, and building trust among users (students, teachers, parents, and administrators).

7.2.6. Infrastructure Requirements

Can it be deployed locally or does it require the cloud?

It requires the cloud, as it is an AWS service.

Is it feasible to have a server per school or a global server instead?

A global cloud server managed by AWS is the most viable architecture, as the API is accessed over the Internet.

Estimated infrastructure consumption (CPU, GPU, RAM).

As a managed service, the direct consumption of computational resources (CPU, GPU, RAM) by Comprehend models is managed by AWS and abstracted from the user. The user does not need to provision or administer these resources directly for Comprehend.

Resource consumption is reflected indirectly through service costs:

- API Usage Costs: Per-unit text processing fees for standard APIs implicitly reflect the compute used by AWS.
- Custom Endpoint Costs: The number of Inference Units (IUs) provisioned for real-time inference is directly related to the computational resources AWS allocates for that endpoint. More IUs mean more resources and higher cost ($0.0005 per IU per second).
- Training Costs: The $3 per training hour fee covers the resources consumed during the custom model creation phase.

It is important to differentiate this from the infrastructure consumption of the client's application components that interact with Comprehend. The orchestrating application (running on Lambda, EC2, etc.), databases, queuing systems, etc., will have their own CPU, RAM, and storage requirements, and their own associated costs, which are independent of Comprehend's direct costs.

7.2.7. Advantages and Limitations

Aspect	**Advantages**	**Limitations**
Accuracy	Robust basic functions; specific APIs for PII/Toxicity (English); potential for high accuracy with well-trained custom models.	Pre-trained accuracy is medium; Toxicity API is English-only and has false positives; does not directly detect bullying/grooming/radicalization; custom accuracy critically depends on high-quality data.
Cost	Flexible pay-per-use model with initial free tier; cost-effective asynchronous processing for large volumes.	Pay-per-use model, but real-time endpoints and training can be expensive.
Ease of Integration	Well-documented APIs, SDKs, and easy integration with other AWS services (S3, Lambda).	Requires technical expertise for advanced specialization (fine-tuning).
Explainability	Provides confidence scores as an indicator of model certainty.	Only provides confidence scores, lacks deep explanations (like SHAP or LIME) on which parts of the text influenced the decision.
Flexibility for Fine-Tuning	Allows creating custom classifiers/recognizers with a simplified training process (AutoML).	Allows customization, but requires significant effort in preparing high-quality, multilingual data and has language limitations for training with certain formats like annotated PDFs (English only).
Regulatory Compliance	AWS offers a basic framework (DPA/SCCs) and security tools; PII detection can aid compliance.	AWS offers a basic framework, but full compliance requires considerable effort on the part of the customer (shared responsibility), especially for minors' data.

7.2.8. Conclusions and Recommendations

Is this model suitable for our use case?

Amazon Comprehend presents a set of NLP tools that could be adapted for sensitive content detection in European school environments, but it is not a ready-to-use solution and presents significant challenges. It offers relevant building blocks such as Custom Classification, Custom Entity Recognition, and PII Detection, which are necessary to address the problem.

However, significant obstacles must be carefully considered:

- The critical limitation of the Toxicity API to the English language makes it unfeasible for direct multilingual deployment in Europe.
- The need to create and maintain high-quality, high-volume, multilingual training datasets for custom models (especially for Custom Classification) represents a very considerable technical and resource effort. Detecting bullying, grooming, or radicalization requires very specific and contextualized data.
- The low explainability (only confidence scores) and the absence of integrated tools for bias management are serious disadvantages for an environment as sensitive as the school, where fairness, transparency, and accountability are paramount.
- GDPR compliance, especially concerning minors' data, largely rests with the customer and requires meticulous implementation of technical and organizational controls.
- The cost of custom endpoints for real-time inference can be high if low latency and high performance are required.

Therefore, Comprehend's suitability is conditional and heavily depends on the organization's ability and willingness to invest in overcoming these limitations.

What conditions or adjustments would be necessary to use it?

To use Amazon Comprehend effectively and responsibly in this context, the following conditions and adjustments would be required:

- Robust Data Strategy: Develop a comprehensive plan for the collection, annotation (ensuring high quality and consistency), and continuous management of training data that is representative, diverse, and specific to each language and target sensitive content type. Prioritize Custom Classification.

- Address Multilingualism: Make a strategic decision: (a) Limit toxicity/sensitive content detection to English communications, (b) Implement a preceding translation workflow (carefully evaluating the impact on cost, latency, and accuracy), or (c) Undertake the effort to train and maintain separate custom classification models for each relevant European language.
- Rigorous GDPR Implementation: Ensure the choice of AWS regions within the EU, implement robust encryption (KMS), strict access controls (IAM), auditing (CloudTrail), and establish clear data minimization, retention, and deletion policies. Properly manage parental consent and data subject rights. Configure opt-out from AWS data usage if deemed necessary for compliance reasons.
- Independent and Continuous Validation: Conduct exhaustive performance testing (precision, recall, F1, error analysis) using real-world data not seen during training, for each language and content category. Establish confidence thresholds based on this empirical validation, not just default values. Continuously monitor performance and retrain models periodically.
- Bias Evaluation and Mitigation: Since Comprehend lacks integrated tools, implement an external process to evaluate potential demographic biases in model predictions. This could involve using test datasets specifically designed to measure fairness, or adopting responsible AI evaluation methodologies. Careful curation of diverse training data is fundamental.
- Human Review-Centric Design: Due to low explainability, design workflows where system detections (especially high-sensitivity or low-confidence ones) are reviewed by trained personnel before definitive actions are taken. The system should be seen as a support tool, not as an autonomous final decision-maker.
- Detailed Budget Planning: Carefully estimate all cost components: API usage, model training, model storage, endpoint cost (if real-time is required), potential translation costs, costs of the surrounding application infrastructure, and the cost (in time and resources) of data management and validation. Compare the cost of 24/7 endpoints with asynchronous processing if latency allows.
- Potential Hybrid Approach: Consider using Comprehend for simpler tasks or as a first filter (e.g., PII detection, basic explicit language filtering if in English), complementing it with human review or more specialized tools for the most complex and nuanced cases.

7.2.10. Visual Summary

Category	Result
Model Type	Text (Cloud NLP Service with pre-trained and customizable APIs)
Accuracy	Medium/Low (direct) for specific sensitive content (bullying, etc.). Potentially Medium/High (adapted) with high-quality data, but requires exhaustive validation. Known issues (Toxicity).
Cost	Variable (Low to High). Pay-as-you-go API. Custom training and real-time endpoints can be expensive.
License	Proprietary (commercial use allowed)
Fine-Tuning	Yes (Customization) via Custom Classification/Entities. Difficulty: Medium (requires good data). Language limitations.
GDPR Compliance	Partial (Requires client effort). AWS provides framework, but client is responsible for implementation (minors' data, consent, controls).
Final Recommendation	Requires significant adaptation and rigorous validation. Not recommended for direct multi-language use due to limitations (Toxicity, Explainability, Bias). Consider alternatives if language, explainability, or bias requirements are strict.

7.3 Analysis of IBM Watson Natural Language Understanding

7.3.1. Model Identification

- Model Name: IBM Watson® Natural Language Understanding (NLU). This is IBM's primary cloud-based service for advanced text analysis. It is distinct from, but related to, other Watson services like Watson Discovery, watsonx Assistant, and Watson NLP Library for Embed.
- Model Type: Text (Natural Language Processing - NLP).
- Provider: IBM

7.3.2. General Model Description

What tasks does the model currently perform?

IBM Watson NLU performs a wide range of general Natural Language Processing (NLP) tasks by default. These capabilities include: entity recognition (people, places, organizations), keyword extraction, sentiment analysis (at document and specific phrase level), emotion analysis (joy, anger, sadness, fear), category classification (using a hierarchical taxonomy), concept identification, relationship

extraction (between entities), syntax analysis, and metadata extraction (author, title).

These are pre-trained capabilities designed for broad applicability in various business domains, such as customer feedback analysis, content management, and market research. The model is designed to extract meaning and metadata from large volumes of unstructured text.

What tasks is it not specialized in and would require adaptation?

The model is not specifically pre-trained to identify nuanced categories such as bullying, grooming, threats, or radicalization in the specific context of European schools. Standard features do not match these sensitive categories.

It lacks inherent understanding of school-specific slang, jargon, evolving keywords, or linguistic patterns used by children and adolescents in multiple European languages. These elements are critical for detecting subtle threats or grooming attempts.

Multilingual Nuances: Although it supports multiple languages, standard models may not capture the specific cultural and linguistic subtleties of sensitive content in diverse European contexts. It's important to note that language support varies significantly by model feature or function.

Need for Adaptation: To effectively address sensitive content detection in school environments, custom model training is required, specifically, custom classification models. These models must be trained with labeled data that reflects the sensitive categories and relevant linguistic context.

Even with fine-tuning, what can it not do or what are its limitations?

Implicit Meaning/Sarcasm: Deeply contextual or sarcastic harmful content might go undetected. NLU models, even customized, struggle with hidden intent behind seemingly innocuous language. Performance will heavily depend on the quality and representativeness of the training data used for customization.

Evolving Language: The model cannot automatically adapt to rapidly changing slang, emojis, or coded language among young people without continuous retraining and data updates. Constant effort is required to keep the model current.

Visual/Audio Content: Watson NLU is strictly a text analysis tool. It cannot analyze images, videos, or audio components that are often present in online interactions where bullying or grooming can occur. This would require separate and complementary tools.

Real-Time Intervention: NLU provides analysis after text is generated. It does not prevent harmful content from being initially sent; it is a detection tool for subsequent review or action.

Explainability Limits: Even with customization, explaining why a specific text was flagged (especially by complex deep learning models) may have limitations directly within the NLU API.

Foreseen use cases (future specialization in sensitive content detection).

Analyzing student communications (e.g., messages on school platforms, forum posts, subject to strict GDPR/privacy restrictions) to flag potential instances of bullying, threats, self-harm indicators, grooming attempts, or exposure to radicalizing content.

Providing alerts to designated school staff or safeguarding officers for review and possible intervention.

Aggregating anonymized data (if feasible under GDPR) to identify trends or hotspots within the school environment.

The core value proposition here shifts from IBM's general business use cases (sentiment analysis, keywords for marketing or content management) to a highly specialized, sensitive, and ethically complex public safety application. This implies that standard performance metrics and general ethical considerations may not suffice. Harm detection requires high precision and recall for specific, rare, and evolving categories, demanding custom models trained with sensitive, domain-specific data, going beyond standard NLU capabilities and requiring rigorous ethical oversight.

7.3.3. Technical Capabilities

Accuracy and recall (if available).

Standard Features: Benchmarks exist that compare Watson NLU/Assistant (which uses NLU) with other platforms for general tasks like intent classification or sentiment analysis. Some studies show that Watson performs well, sometimes best, in specific contexts (e.g., F1 > 84% in intent classification in software engineering tasks, high accuracy for negative sentiment, outperforming others in F1 scores in the retail sector). However, performance varies by dataset and task.

Custom Classification (Sensitive Content): No specific public benchmarks were found for this exact use case (bullying, grooming, etc., in European schools). Performance (Precision, Recall, F1-score) will largely depend on:

- The quality, quantity, and representativeness of the custom training data. A minimum of 5 examples per label, a maximum of 3000 labels, and a maximum of 20000 examples are required.
- The specific languages being processed.
- The complexity and subtlety of the content being detected.

Metrics like Precision (accuracy of positive predictions) and Recall (ability to find all true positives) are critical. For sensitive content, high recall (minimizing missed cases/false negatives) is often paramount, even if it means lower precision (more false positives requiring human review). The F1 score balances both metrics. Evaluation needs tools like confusion matrices. The absence of ready-to-use specific benchmarks for this sensitive task means that the implementing organization must budget for and carry out its own rigorous testing and validation using representative data before deployment. Relying on general benchmarks would be inadequate and potentially dangerous. General benchmarks test common NLP tasks. Sensitive content detection is a niche and high-risk task. The definition of "bullying" or "grooming" can be subjective and context-dependent. Therefore, pre-trained models or general benchmarks will not suffice. The user must create a custom dataset that reflects their specific definitions and context, train a custom model, and evaluate its Precision/Recall/F1 on a test set representative of real school communications.

Processing speed (images/second, words/second, etc.).

API latency is measured in milliseconds (ms). Benchmarks for general Watson services suggest that average response times can be low (e.g., ~100-200ms mentioned as desirable targets for NLP APIs in general; Watson Assistant + NLU reports around 700ms in a specific 2017 test case, although this is outdated).

Performance metrics (requests/second or records/second) are also tracked by IBM monitoring tools like OpenScale.

Performance can be influenced by factors such as input text size (NLU items are based on 10,000-character chunks), the number of features requested, the use of custom models, and the overall load on IBM Cloud.

Performance gains have been observed when using Intel-optimized libraries/hardware (up to a 35% performance increase reported for some NLP/NLU tasks).

Speed is measured per API call. Processing 3000 texts/day translates to approximately 1 text every 29 seconds on average, which is within typical API latency capabilities. The bottleneck is more likely to be the overall processing architecture and cost, not the speed of the individual API call.

Fine-Tuning Capability:

Does it allow Fine-Tuning?

Yes, primarily through Custom Classification Models. Custom Entities and Relationships are also possible via Watson Knowledge Studio (WKS), but classification is the key feature for this use case. "Fine-tuning" here means training a model with user-provided labeled data. Prompting (instruction-based tuning) is less relevant for the NLU API compared to generative models.

How difficult or costly is it?

Difficulty: Requires expertise in data preparation (labeling text according to defined sensitive categories), data formatting (JSON or CSV), and using the NLU API or potentially Watson Studio/Knowledge Studio. It is not "no-code" for training classification models via the API (requires scripting). It needs careful definition of sensitive categories and consistent labeling. It is considered of intermediate difficulty.

Cost: Specific cost per custom classification model: USD $25 per model per month on the Standard plan. The Lite plan allows 1 free custom model. There's also the cost of data preparation/labeling (human effort) and potentially using Watson Knowledge Studio if entities/relationships are needed (USD $800/month per model). API usage costs apply for training/analysis.

Workload:

Can it process between 2000 and 3000 units (images/texts) daily?

Yes. This volume is relatively low for a cloud-based API service designed to scale. Standard plan tiers support millions of "NLU items" per month. 3000 texts/day is ~90,000 texts/month. Assuming 1 text unit and 1 custom classification feature per text, this equates to ~90,000 NLU items/month, which falls within the first tier of the Standard plan.

Does it require high CPU, GPU, RAM resources?

Cloud API (Managed): For cloud API users, IBM manages the underlying infrastructure. Resource requirements are abstracted; users pay based on usage (NLU items), not direct resource consumption. Scalability is managed by IBM Cloud.

On-Premise (Cloud Pak for Data): If deployed on-premise via Cloud Pak for Data, then yes, significant local CPU, GPU (especially for training/certain models), and RAM resources would be required locally. Sizing depends heavily on the specific CP4D configuration and workload. General NLP tasks can be resource-intensive.

Training: Training custom models (especially complex ones or with large datasets) can be computationally intensive, potentially requiring GPU resources. This is managed by IBM's infrastructure when using the cloud service API for training.

The low daily volume suggests that the primary challenge is not technical performance, but cost-effectiveness (especially custom model fees and API calls if many features per text are used) and achieving high precision/recall for sensitive, low-frequency events within that volume. 3000 texts/day is trivial for a cloud API. The cost model is per NLU item (text chunk * features). If only custom classification is used on short texts, the cost is relatively low (e.g., 90k items/month * $0.003/item = $270 + $25/model = $295/month). But adding other potentially useful features (like entity recognition for names, sentiment/emotion analysis) significantly multiplies the cost per text. More importantly, finding rare events (like grooming) in low volume requires a very precise (high recall) custom model, which is the main technical/data science challenge, not infrastructure scaling.

7.3.4. Licensing and Terms of Use

Is it open source or proprietary?

Proprietary

License type (Apache 2.0, MIT, commercial license, etc.).

Commercial via the IBM Cloud services agreement.

Does it allow commercial use?

Yes, it is designed for commercial use.

Approximate cost:

Annual or monthly license: Cost is primarily based on usage and custom models. Plans are typically billed monthly.

Cost per use (if applicable, e.g., pay-per-API, etc.): Yes, based on "NLU Items."

Definition of NLU Item:

1 text unit (<= 10,000 characters) analyzed for 1 feature (e.g., custom classification, sentiment, entities). Example: a 15,000-character text analyzed for 3 features = 2 text units * 3 features = 6 NLU items.

Lite Plan: Free tier allows up to 30,000 NLU items/month and 1 free custom model. Suspends if limit is exceeded until the next month. Suitable for Proof of Concept (PoC).

Standard Plan (Pay-as-you-go, Monthly Billing): Tiered pricing per NLU item:

- Tier 1: $0.003 / NLU item (1 - 250,000 items/month)
- Tier 2: $0.001 / NLU item (250,001 - 5,000,000 items/month)
- Tier 3: $0.0002 / NLU item (> 5,000,000 items/month)
- Custom Model Costs (Standard Plan):
- Custom Classification Model: USD $25 / model / month.
- Custom Entities/Relationships Model (via WKS): USD $800 / model / month.

Deployment Costs: Cloud usage costs are included in the API price. On-premise deployment via Cloud Pak for Data involves separate licensing and infrastructure costs for CP4D.

Cost Estimator: IBM provides a cloud cost estimator tool.

The pricing model strongly incentivizes using fewer features per text analysis and processing shorter texts to minimize "NLU items." For the sensitive content use case, if only the custom classification feature is needed, the cost per text is relatively low ($0.003 at Tier 1). However, adding other potentially useful features (like entity recognition for names, sentiment/emotion analysis) significantly multiplies the cost per text. The flat fee of $25/month for a custom classification model is lower compared to potential usage costs or the $800/month for custom entity models. Decision-makers need to have costs clearly defined.

7.3.5. Legal and Ethical Analysis

GDPR compliance and European regulations.

The IBM Cloud platform offers features and agreements to support GDPR compliance.

"EU Supported" Configuration: Accounts can be configured as "EU Supported," restricting data access and support to EU-based personnel for certain services/regions.

Frankfurt Data Center: The NLU service is hosted in Frankfurt (among others). The Frankfurt region has specific EU-managed controls and can be used for data residency requirements. Using the Frankfurt endpoint is critical if EU data residency is mandatory. Custom model training must also be done with data residency in mind; if the Frankfurt endpoint is used for training API calls, it is assumed that processing associated with training occurs within that designated region.

Customer Responsibility: IBM emphasizes that the customer (data controller) is ultimately responsible for ensuring their use of the service complies with GDPR and other relevant laws. This includes the legal basis for processing, data

minimization, purpose limitation, consent management (if applicable), and handling data subject rights.

Sensitive Data: Processing sensitive student communications requires a clear legal basis under GDPR (e.g., legitimate interest related to safety, possibly explicit consent depending on context and local laws) and robust safeguards. Article 9 conditions for processing special categories of data may apply.

Although IBM provides compliant infrastructure (Frankfurt region, DPA), the use case itself (analyzing student communications for sensitive content) carries significant GDPR risk. The customer must have a robust Data Protection Impact Assessment (DPIA), clear policies, a defined legal basis, strict access controls, data minimization practices, and processes for transparency and data subject rights (especially erasure/rectification) to use NLU compliantly in this context. Relying solely on IBM's platform compliance is insufficient. GDPR compliance has two parts: the processor's infrastructure/agreements (IBM) and the controller's use case (school/organization). IBM offers tools (Frankfurt DC, DPA, EU Support). However, the act of processing student messages for sensitive content is high-risk. The controller must justify this processing under GDPR (Art 6, Art 9), ensure necessity and proportionality, manage consent/rights (Art 7, 15-22), and implement technical/organizational measures (Art 32). The NLU tool is just one part of a much larger compliance puzzle that the client must solve.

This use case directly involves the processing of data relating to minors, which requires scrutiny under GDPR and specific national child protection laws across Europe.

Requires robust age verification (if applicable), potential parental consent mechanisms (depending on age and local law), and strict data access controls limited to authorized safeguarding personnel.

Data minimization is crucial – only analyze data necessary for the specific purpose of detecting harm.

Retention policies must be clearly defined and enforced, deleting data when no longer needed for the safeguarding purpose (GDPR storage limitation principle).

Bias handling (racial, gender, cultural...).

Bias Potential: AI models, including NLU, can inherit biases present in their training data. Custom models trained for sensitive content are particularly vulnerable if training data underrepresents certain demographics or overrepresents specific types of harmful language associated with particular groups.

IBM's Stance/Tools: IBM emphasizes AI ethics and offers tools like AI Fairness 360 (open-source toolkit) and features within Watson OpenScale (part of the broader Watson platform, not directly in the NLU API) for bias detection and mitigation.

Mitigation: Requires careful curation and balancing of training data for custom models. Continuous monitoring and auditing are essential to detect biased performance across different student groups (defined by relevant protected characteristics). This likely requires effort beyond the standard NLU service capabilities.

Detecting biases in sensitive content classification is complex. Slang, cultural references, and context differ across groups. A model might incorrectly flag common language in one group as harmful or overlook genuinely harmful content specific to another group if not carefully trained and evaluated with diverse data. The very definition of "fairness" needs careful consideration in this context. Bias arises from data. If training data for "bullying" primarily contains examples from one demographic group, the model might perform poorly (either false positives or negatives) on bullying expressed differently by another group. Mitigating this requires: 1) Collecting diverse and representative training data (difficult for sensitive topics). 2) Using fairness metrics during evaluation (e.g., comparing recall rates across groups). 3) Potentially using advanced techniques (available in toolkits like AIF360 or OpenScale but requiring integration effort).

Explainability capacity ("explainability") of decisions.

NLU API: Primarily provides confidence scores for its predictions (e.g., for categories, sentiment, custom classifications). Standard NLU features like Categories have an optional explanation parameter (English-only). No explicit mention of built-in feature attribution methods like LIME/SHAP for custom classification model predictions directly in the NLU API documentation was found. Custom model predictions return labels and their confidence scores.

Broader Watson Platform: IBM offers explainability tools like LIME and SHAP integration, but typically through Watson OpenScale or related platforms designed for model monitoring and governance. These require separate setup and integration. General XAI techniques like LIME/SHAP are described in contexts outside the NLU API.

IBM Principles: IBM advocates for transparency and explainability as key pillars of trust.

Context: For sensitive decisions (flagging potential harm), understanding why a model made a prediction is crucial for human reviewers and for building trust. Confidence scores alone may be insufficient.

There appears to be a gap between IBM's high-level principles on explainability and the specific features readily available within the NLU API for custom models. Achieving deep explainability (feature importance, LIME/SHAP) for custom classification models likely requires significant additional effort: either integrating with Watson OpenScale or implementing external explainability libraries to work with NLU's predictions. This adds complexity and potentially cost. The user needs to know why a message was flagged. The NLU API gives a classification and a confidence score. This does not explain what words or patterns triggered the flag. Techniques like LIME/SHAP provide this word-level importance. NLU API documents do not show this for custom models. Other Watson tools (OpenScale) do. Therefore, using only the NLU API limits explainability for custom models; achieving deeper explanations requires using additional tools/libraries.

7.3.6. Infrastructure Requirements

Can it be deployed locally or does it require the cloud?

Cloud (Primary): IBM Watson NLU is fundamentally a cloud-based API service hosted on IBM Cloud.

On-Premise Option: Can be deployed "behind your firewall" via IBM Cloud Pak for Data (CP4D). CP4D itself can run on-premise on Red Hat OpenShift or on private/public clouds (AWS, Azure, Google Cloud).

The cloud API is the simplest and standard way to use NLU. The on-premise option via CP4D offers data control but adds significant infrastructure management complexity and cost (CP4D license, OpenShift management, hardware resources).

Is it feasible to have a server per school or a global server instead?

Cloud API: A "server per school" model is not applicable. Clients access IBM Cloud's central NLU service via the API. A central application/backend would handle API calls for all involved schools.

On-Premise (CP4D): Deploying a full CP4D instance per school is probably technically possible but economically and operationally unfeasible due to cost and management overhead. A centralized CP4D deployment serving multiple schools (e.g., for a district or region) would be the most realistic on-premise approach.

The choice is not server per school vs. global server, but Cloud API (IBM centralized service) vs. On-Premise CP4D (self-managed centralized service). The decision depends on data residency control needs, budget, and available technical expertise to manage an on-premise platform like CP4D. For most, the cloud API is the most viable.

Estimated infrastructure consumption (CPU, GPU, RAM).

Cloud API: Abstracted. Managed by IBM. Not directly relevant to the user beyond API usage costs.

On-Premise (CP4D): Significant. Requires a robust Kubernetes/OpenShift cluster. Specific needs depend on the scale of use (number of concurrent requests, data volume), which CP4D services are deployed alongside NLU, and if model training is performed locally. It will require substantial CPU, RAM, and potentially GPU resources, especially if 3000 requests/day are handled concurrently or training is performed. Detailed sizing requires consulting IBM CP4D documentation and potentially IBM architects.

7.3.7. Advantages and Limitations

Aspect	**Advantages**	**Limitations**
Accuracy	Generally good for common tasks and potentially High (with good data/training).	Critically dependent on training data quality/quantity; requires specific validation.
Cost	Scalable with tiers; Free Lite Plan for PoC.	Can be expensive if many features or long texts are used; high tiers difficult to reach with low volume.
Ease of Integration	Standard REST APIs, SDKs available (Python, Java, etc.).	Requires development to integrate into the application.
Explainability	Provides confidence scores.	Deep explanations (LIME/SHAP) for custom models not directly included; requires additional tools.
Flexibility for Fine-Tuning	Allows custom classification models for specific tasks.	Requires labeled data and training effort; limited to NLU platform capabilities.
Regulatory Compliance	GDPR-supported infrastructure (Frankfurt DC, DPA, EU Support).	Use case compliance is entirely customer's responsibility; high compliance burden for minors' data.

7.3.8. Conclusions and Recommendations

Is this model suitable for our use case?

IBM Watson NLU is potentially suitable, but its suitability is subject to significant conditions. The platform provides the necessary core technical capability, which is custom classification, to address sensitive content detection. Its cloud infrastructure, particularly the Frankfurt data center, offers support for GDPR requirements, including EU data residency. Furthermore, the cost of the basic custom classification model is reasonable ($25/month plus usage).

However, the success of implementation critically depends on the organization's ability to create and maintain high-quality, representative, and multilingual training data for the specific sensitive content categories. Standard pre-trained models are insufficient for this task. Limitations in the direct explainability of custom model predictions via the NLU API must be addressed. Fundamentally, legal and ethical responsibility (GDPR compliance, child protection) rests entirely with the customer and requires thorough and ongoing due diligence.

What conditions or adjustments would be necessary to use it?

To use IBM Watson NLU effectively and responsibly in this context, the following conditions and adjustments would be required:

Data: It is imperative to develop a robust, diverse, accurately labeled, and multilingual dataset for custom classification model training. This must cover all target sensitive categories and relevant European languages/cultures. Continuous data collection and re-labeling are needed to adapt to evolving youth language.

Training and Validation: Train custom classification models using the NLU API. Rigorously evaluate performance (Precision, Recall, F1) on unseen, representative test data before deployment. Establish appropriate confidence thresholds based on the tolerance for false positives versus false negatives, likely prioritizing high recall to minimize missed cases.

Compliance: Conduct a comprehensive Data Protection Impact Assessment (DPIA). Establish a clear legal basis for personal data processing, especially for minors. Implement strict access controls, data minimization, and clear retention policies. Ensure use of the Frankfurt endpoint if EU data residency is required. Comply with all relevant child protection laws.

Explainability: Implement a strategy to explain model decisions to human reviewers. This may involve integrating with Watson OpenScale or using external libraries if confidence scores alone are insufficient to understand why a text was flagged.

Workflow: Design a clear workflow for human review of flagged content, including escalation procedures and actions to be taken. Clearly define roles and responsibilities.

Monitoring: Continuously monitor model performance, drift, and potential bias in production. Plan for regular retraining with new data.

7.3.9. Visual Summary

Category	**Result**
Model Type	Text (Cloud NLP Service with pre-trained and customizable APIs)
Accuracy	Medium (Standard) / Potentially High (Custom, critically dependent on data and rigorous validation)
Cost	Medium (Standard Plan: $0.003/NLU item Tier 1 + $25/month per custom model)
License	Proprietary (commercial use allowed)
Fine-Tuning	Yes (Custom Classification Model; intermediate difficulty, requires data and validation)
GDPR Compliance	Yes (Platform with GDPR support, DPA, Frankfurt Region). Use case compliance is customer's responsibility.
Final Recommendation	Requires adaptation (Custom model is essential; high burden of performance validation and legal/ethical compliance)

8. Analysis of Open Source AI Models for Text Recognition

8.1 Analysis of the Multilingual Toxic-BERT Model (Fine-tuned variants from XLM-RoBERTa Base/Large)

8.1.1. Model Identification

- Model Name: Multilingual Toxic-BERT (fine-tuned variants from XLM-RoBERTa Base/Large)
- Model Type: Text (Transformer-based language model).
- Provider: Base model XLM-RoBERTa by Meta AI. Fine-tuned variants by various groups/researchers (e.g., Unitary AI, TextDetox Project).

8.1.2. General Model Description

What tasks does the model currently perform?

The base XLM-RoBERTa model is pre-trained for general language understanding across 100 languages using Masked Language Modeling (MLM). It provides a robust foundation for downstream tasks.

Existing "Toxic-BERT" variants are fine-tuned to detect general toxicity in text. Some models predict multiple labels (e.g., toxic, severely toxic, obscene, threat, insult, identity-based hate, identity attack, sexually explicit content), while others focus on binary classification (toxic vs. neutral). The languages covered vary depending on the fine-tuned model (e.g., 7 languages for Unitary, 15 for TextDetox-v24).

What tasks is it not specialized in and would require adaptation?

Neither the base XLM-RoBERTa nor existing Toxic-BERT variants are specialized in detecting the specific and often subtle forms of sensitive content relevant to school environments, such as grooming or radicalization. The definition of "toxicity" in existing training datasets (e.g., Jigsaw 2, general compilations) focuses more on explicit aggression, insults, or hate speech, and is unlikely to capture the manipulative or persuasive tactics of these phenomena.

Although some labels like threats or insults are relevant for bullying, specific and reliable detection of bullying in the school context (considering youth slang, social context, etc.) would also require adaptation.

Even with fine-tuning, what can it not do or what are its limitations?

Deep Contextual Understanding: Even when fine-tuned, the model may struggle with sarcasm, inside jokes, rapidly evolving slang, cultural nuances specific to different European languages and regions, and distinguishing between actual intent and literal meaning. It remains very difficult to detect when an adult is attempting to groom a minor or when an individual begins to adopt extremist ideas, especially if disguised language is used or if the process is gradual over time.

Dynamic Adaptation: It cannot adapt in real-time to new slang, memes, or evasion tactics without periodic retraining.

Causal or Ethical Reasoning: It does not possess true ethical reasoning or understand the deep social implications of the content it detects. Its classification is based on patterns learned from data.

Complete Explainability: The "black box" nature persists. While XAI techniques (LIME, SHAP) can offer local explanations, they do not provide full transparency into the internal decision process.

Inherent Biases: Despite mitigation efforts, there will always be a residual risk of biases inherited from pre-training data or introduced during fine-tuning.

Foreseen use cases (future specialization in sensitive content detection).

The objective is to adapt (through fine-tuning) the model to reliably detect in a multilingual manner:

- Bullying in its various forms.
- Direct or veiled threats.
- Online grooming tactics.
- Content related to radicalization.

The tool would be used to support human moderators in European school environments to identify and flag potentially harmful communications on digital platforms used by students.

8.1.3. Technical Capabilities

Accuracy and recall (if available).

XLM-RoBERTa Base: Not directly applicable (it's a pre-trained language model). Its performance is measured in benchmarks like XNLI (80.9% average zero-shot accuracy, 83.6% with translate-train), MLQA (70.7% F1), NER (89.4% F1 multi-train).

Existing Toxic-BERT Variants:

- unitary/multilingual-toxic-xlm-roberta: Reports high AUCs (>0.91) on Jigsaw datasets. Specific Precision/Recall/F1 per label or language are not provided in the snippets.
- textdetox/xlmr-large-toxicity-classifier-v2: Reports F1 scores for binary classification (toxic/neutral) that vary significantly by language: from 0.97 down to 0.56. For key European languages: EN (0.92), RU (0.95), UK (0.96), DE (0.73), ES (0.71), IT (0.59), FR (0.92).

Projected Performance (After Specific Fine-tuning):

- Bullying/Threats: Reasonable to expect F1 > 0.90 in well-represented languages, similar to cyberbullying benchmarks.
- Grooming/Radicalization: Likely lower, perhaps F1 0.70-0.85, due to subtlety and context dependence.
- Precision/Recall Balance: It will be crucial to adjust the decision threshold to balance minimizing false positives (high precision) and false negatives (high recall), depending on the severity of the content.

Processing speed (images/second, words/second, etc.).

Speed (latency and throughput) critically depends on hardware (GPU vs. CPU), model size (Base vs. Large), batch size, sequence length, and applied optimizations (FP16, INT8, TensorRT, ONNX).

CPU Inference: Very slow, likely unsuitable for real-time. An AWS Serverless (CPU) benchmark for XLM-R Large (token classification) showed ~1.5 seconds average latency.

GPU Inference: Much faster, on the order of tens or hundreds of milliseconds per instance, but varies greatly. More powerful GPUs (A100 vs. T4) and optimizations (TensorRT) significantly reduce latency. Throughput can be improved with larger batches and techniques like sequence packing.

Fine-Tuning Capability:

Does it allow Fine-Tuning?

Yes, XLM-RoBERTa is explicitly designed to be fine-tuned for downstream tasks. It is a standard practice for Transformer models.

How difficult or costly is it?

Technical Difficulty: Moderate to high. Requires expertise in NLP, deep learning frameworks (PyTorch/TensorFlow), and hyperparameter tuning. Implementing optimizations (quantization, LoRA) adds complexity.

Cost: Potentially high. Requires access to powerful GPUs for significant periods for training/experimentation. The greatest cost lies in the creation/acquisition and annotation of high-quality, domain-specific, and multilingual training data, which is extremely challenging for ethical, privacy, and scarcity reasons.

Workload:

Can it process between 2000 and 3000 units (images/texts) daily?

Yes, with adequate infrastructure, processing 3000 texts daily is very feasible. This equates to less than 1 text every 28 seconds. Even with conservative GPU latencies (e.g., 100-500 ms per text), a single GPU could easily handle this volume. Scaling to much larger volumes is possible with more resources (multiple GPUs, scalable cloud deployment).

Does it require high CPU, GPU, RAM resources?

GPU: Highly recommended, almost indispensable for acceptable inference performance. The amount of VRAM needed depends on the model (Large > Base)

and precision (FP32 > FP16 > INT8). XLM-R Large (FP16) may need >14GB VRAM for reasonable batch sizes. XLM-R Base requires less.

CPU: Necessary, but generally less critical than the GPU for inference. Powerful CPUs may be needed if the system becomes CPU-bound at low load or if intensive pre/post-processing tasks are performed.

RAM (System): Tens of GB, depending on model size and whether it fits entirely in VRAM or if CPU is used.

Storage: Space for the model (several GB for Large), training/evaluation data, logs.

Network: Sufficient bandwidth if deployment is centralized and receives data from multiple schools.

8.1.4. Licensing and Terms of Use

Is it open source or proprietary?

XLM-RoBERTa (Base/Large): The base model from Meta AI is distributed as open source.

Fine-tuned Toxic-BERT Variants: Depends on the specific model. Many are publicly available on platforms like Hugging Face, but under different licenses.

- unitary/multilingual-toxic-xlm-roberta: Open source.
- textdetox/xlmr-large-toxicity-classifier-v2: Open source (under the OpenRAIL++ license).

License type (Apache 2.0, MIT, commercial license, etc.).

XLM-RoBERTa (Base/Large): Distributed under the MIT License.

unitary/multilingual-toxic-xlm-roberta: Distributed under the Apache 2.0 License.

textdetox/xlmr-large-toxicity-classifier-v2: Distributed under the OpenRAIL++ License. It's important to note that OpenRAIL is not an OSI-approved open source license due to its use restrictions.

Does it allow commercial use?

MIT License (XLM-R Base/Large): Yes, it permits commercial use.

Apache 2.0 License (unitary/...): Yes, it permits commercial use.

OpenRAIL++ License (textdetox/...): Generally, yes, but with important conditions. The license explicitly prohibits using the model for certain harmful purposes listed in an appendix (behavioral use restrictions). These restrictions must be passed on to downstream users. There are OpenRAIL variants (e.g., OpenRAIL++-M-NC) that prohibit commercial use. It is crucial to review the specific license version.

Approximate cost:

Annual or monthly license: For models under MIT, Apache 2.0, or OpenRAIL++ licenses (commercial variants): €0 (the licenses themselves are free).

Cost per use (if applicable, e.g., pay-per-API, etc.): If the model is used directly (deployed locally or in your own cloud): €0 for the model itself. Costs are for infrastructure and operation.

If accessed via a third-party API that uses these models: Will depend on the API provider's pricing policy.

8.1.5. Legal and Ethical Analysis

GDPR compliance and European regulations.

Complex and High-Risk: Monitoring minors' communications with AI is a high-risk processing activity under the GDPR.

Legal Basis: Consent is problematic; legitimate interest requires a very rigorous Legitimate Interest Assessment (LIA), balancing safety with minors' privacy and freedom of expression.

Sensitive Data: Content may reveal special categories of data (Art. 9), requiring an additional legal basis (e.g., essential public interest based on national law).

Key Principles: Demands strict purpose limitation, data minimization (is scanning everything necessary?), accuracy (error management), storage limitation, and robust security.

Mandatory DPIA: A Data Protection Impact Assessment (Art. 35) is required before deployment.

Transparency: Obligation to clearly inform students/parents about monitoring, AI logic (simplified), and their rights (Art. 13/14).

Rights: Procedures must be enabled for exercising rights of access, rectification, erasure, and, crucially, objection (Art. 15-22).

Automated Decisions: Avoid decisions with significant effects based solely on AI; ensure human intervention (Art. 22).

Transfers: If providers outside the EEA are used, comply with international transfer rules. The Data Protection Officer must be actively involved.

Bias handling (racial, gender, cultural...).

Inherent Risk: XLM-R, pre-trained on CommonCrawl, inherits societal biases present on the web. Fine-tuning can amplify or introduce new biases.

Potential Impact: Risk of uneven performance and discrimination against certain student groups (based on language, dialect, ethnicity, etc.).

Necessary Mitigation: Requires proactive bias audits (using benchmarks like BBQ, CrowS-Pairs) and mitigation strategies (data diversification, model tuning), although complete elimination is difficult. Multilingual training can help mitigate some biases compared to monolingual models.

Explainability capacity ("explainability") of decisions.

Challenge ("Black Box"): Transformer models like XLM-R are inherently opaque. Understanding why a specific text is classified is difficult.

XAI Techniques: Post-hoc methods like LIME and SHAP exist that attempt to explain individual predictions by attributing importance to input words.

XAI Limitations: These explanations are local, approximate, and may not faithfully reflect the model's complex internal reasoning. They can be difficult for non-experts to interpret.

Legal/Ethical Requirement: There is a tension between the need for meaningful explanations (potentially required by GDPR Art. 22) and the current technical ability to provide them completely and reliably. Transparency focuses more on the general process, data, and safeguards.

8.1.6. Infrastructure Requirements

Can it be deployed locally or does it require the cloud?

Both options are possible.

- Local (On-Premise): Requires investment in servers with powerful GPUs and technical staff for management. Offers maximum data control. Less flexible for scaling.
- Cloud: Platforms like AWS, Azure, GCP offer managed infrastructure (GPU instances, serverless inference), on-demand scalability, and MLOps tools. Involves considerations about data transfer and recurring costs.

The choice will depend on budget, internal technical expertise, scalability needs, and, crucially, GDPR data control requirements.

Is it feasible to have a server per school or a global server instead?

Server per School: Likely economically unfeasible and technically complex to manage and maintain consistently across multiple locations. Would require hardware and specialized personnel at each school.

Global Server (or Centralized Regional): Much more viable. Allows centralizing infrastructure (local or cloud), management, maintenance, updates, and technical expertise. Facilitates consistent policy application and regulatory compliance. Monitoring would be done remotely, sending data (or results) to the central location.

Estimated infrastructure consumption (CPU, GPU, RAM).

GPU: This is the key resource. At least one modern GPU with sufficient VRAM is needed. For XLM-R Large (FP16), >14GB VRAM is a reasonable estimate for inference with batches. XLM-R Base requires less. Multiple GPUs may be necessary for high load or low latency.

CPU: Necessary, but generally less critical than the GPU for inference. Powerful CPUs may be needed if the system becomes CPU-bound at low load or if intensive pre/post-processing tasks are performed.

RAM (System): Tens of GB, depending on model size and whether it fits entirely in VRAM.

Storage: Space for the model (several GB for Large), training/evaluation data, logs.

Network: Sufficient bandwidth if deployment is centralized and receives data from multiple schools.

8.1.7. Advantages and Limitations

Aspect	Advantages	Limitations
Accuracy	Potentially High (fine-tuning): The base XLM-R model has achieved outstanding results in international tests evaluating its multilingual understanding capabilities. Fine-tuning can achieve high precision/F1 (>0.90) for explicit bullying/threats.	Variable: Inconsistent performance across languages in existing models. Lower for subtle tasks: Grooming/radicalization will likely have lower precision/F1 (estimated 0.70-0.85). Errors: Prone to errors with sarcasm, slang, context.
Cost	Free Base Model: MIT/Apache/OpenRAIL++ licenses are free.	High Operational Cost: Requires powerful GPUs (hardware/cloud cost). Data Cost: Acquisition/annotation of fine-tuning data is very expensive. Personnel Cost: Needs ML/NLP experts, IT, human

		moderators. Compliance Cost: GDPR (DPIAs, audits) is expensive.
Ease of Integration	Standard Frameworks: Integrates with Hugging Face Transformers, PyTorch/TensorFlow. Available APIs: Pre-fine-tuned models exist on Hugging Face.	Technical Complexity: Requires ML/NLP expertise for fine-tuning, deployment, and optimization (quantization, etc.). Data Dependence: Effective integration critically depends on having appropriate fine-tuning data.
Explainability	Available XAI Techniques: Post-hoc methods like LIME/SHAP can be applied for local explanations.	"Black Box": Internal workings remain opaque. Limited Explanations: LIME/SHAP are local approximations, do not reveal full reasoning, and can be difficult to interpret. Tension with GDPR requirements.
Flexibility for Fine-Tuning	High: Designed for fine-tuning. XLM-R is a robust multilingual base. Allows adaptation to new tasks/domains/languages.	Data-Dependent: Effectiveness of fine-tuning is limited by the quality and quantity of available specific data. Costly: Requires computational resources (GPU) and time.
Regulatory Compliance	Possible (with effort): The model itself does not violate GDPR, but its application requires a robust compliance framework.	High Risk and Complexity: Complying with GDPR (legal basis, minimization, DPIA, transparency, rights) is a major and mandatory challenge. Licenses: OpenRAIL++ imposes use restrictions that must be managed.

8.1.8. Conclusions and Recommendations

Is this model suitable for our use case?

Potentially suitable, but with VERY significant reservations. XLM-RoBERTa provides the necessary multilingual foundation. However, existing Toxic-BERT variants are not directly suitable for detecting grooming or radicalization, and their performance in general toxicity varies by language. Final suitability critically depends on the ability to overcome adaptation (data), compliance (GDPR), and ethical challenges. It is not a plug-and-play solution.

What conditions or adjustments would be necessary to use it?

Specific and Rigorous Fine-tuning: Additional fine-tuning is essential, using high-quality datasets specific to the school context, and explicitly covering bullying, threats, grooming, and radicalization in the relevant European languages. This requires a massive investment in ethical data creation/acquisition and annotation.

Robust GDPR Compliance Framework: Conduct comprehensive DPIAs, establish a solid legal basis (likely legitimate interest with a detailed LIA), ensure full transparency, implement data minimization and security, and effectively manage data subject rights.

Active Bias Mitigation: Perform continuous bias audits and apply mitigation techniques to ensure fairness across different student groups.

Mandatory Human Oversight: The system should only be used as a supporting tool; final decisions must always rest with qualified human moderators.

Adequate Infrastructure: Investment in hardware (GPU) or cloud services to ensure adequate performance and scalability.

Legal License Review: Carefully review the terms of the specific model license chosen, especially if it is OpenRAIL++.

8.1.9. Visual Summary

Category	Result
Model Type	Text (Multilingual Transformer-based)
Accuracy	Potential: High (Bullying/Threats) / Medium (Grooming/Radicalization) - Requires extensive fine-tuning. Current performance varies by language.
Cost	Model: €0 (license). Operation: High (GPU Infrastructure, Data, Personnel, Compliance).
License	Base (XLM-R): MIT (Permissive, commercial use OK). Fine-tuned: Variable (Apache 2.0 - OK; OpenRAIL++ - commercial OK with behavioral use restrictions).
Fine-Tuning	Difficulty: Moderate-High (Technical) / Very High (Data and Ethics).
GDPR Compliance	Not by default. Requires significant effort, mandatory DPIA, robust compliance framework. High risk.
Final Recommendation	Requires VERY significant adaptation and extreme caution. Proceed only with massive investment in data, rigorous GDPR compliance, bias mitigation, human oversight, and after controlled pilots.

9. Analysis of Proprietary Multimodal AI Models

9.1 Analysis of the Microsoft Azure Content Safety Model

9.1.1. Model Identification

- Model Name: Microsoft Azure AI Content Safety (includes APIs for Text, Image, Multimodal, Prompt Protection, Coherence, Protected Material (Copyrighted or legally protected content), Custom Categories)
- Model Type: Multimodal (Capable of analyzing Text, Images, and Image+Text combinations).
- Provider: Microsoft

9.1.2. General Model Description

What tasks does the model currently perform?

Harmful Content Detection: The core of the service detects objectionable content in text and images, classifying it into four main categories: Hate, Sexual, Violence, and Self-Harm. It assigns severity levels (0, 2, 4, 6) for each detected category, allowing for granular filtering.

Multimodal Analysis: Jointly analyzes images and associated text (including text extracted from the image via OCR) for a more comprehensive contextual understanding of potentially harmful content.

Protection against LLM Attacks (Prompt Shields): Detects attempts to manipulate large language models (LLMs), such as "jailbreak" (attempts to bypass system rules) and indirect attacks (malicious instructions hidden in provided documents or data).

Groundedness Detection: Verifies if AI-generated responses are based on user-provided information (sources), helping to mitigate "hallucinations" or non-factual information. Includes a "reasoning" mode to explain detections.

Protected Material Detection: Identifies whether AI-generated text or code matches known copyrighted material (song lyrics, articles, code from public repositories).

Multilingual Support: The main models (Hate, Sexual, Violence, Self-Harm) are specifically trained and tested in 8 languages (Chinese, English, French, German, Spanish, Italian, Japanese, Portuguese) but can work in more than 100 languages with variable quality. It automatically detects the input language for these categories.

What tasks is it not specialized in and would require adaptation?

Detection of Specific Categories in School Environments: The model does not come pre-trained to detect very specific and contextual categories relevant to European schools, such as:

- Bullying - Although it can partially overlap with "Hate" or "Violence," it requires specific nuances.
- Specific Drug Consumption/Promotion.
- Weapon Use/Display in a school context.
- Radicalization (political, religious).
- Eating disorders.
- Gambling.
- Specific adult content not necessarily "sexual" as per the model's definition.
- Cybersecurity (Phishing, Malware - although Prompt Shields covers some manipulation aspects).
- Specific forms of discrimination (Homophobia, Racism - partially covered by "Hate," but may require cultural/linguistic specificity).

Deep Understanding of European Cultural Context: Although it supports many European languages, interpreting cultural nuances, local slang, or specific references may not be as accurate as a system specifically trained in that context.

Detection of Subtle Biases: While Microsoft works on mitigating biases in its models, the detection of implicit or very subtle biases in generated or analyzed content is not an explicitly detailed function and would require specific validation.

PII (Personally Identifiable Information) Identification and Redaction: The service does not natively offer PII detection or redaction, which is crucial in school environments.

Even with fine-tuning, what can it not do or what are its limitations?

Does not allow Direct Fine-Tuning of the Base Model: It is not possible to retrain the core Azure Content Safety models (Hate, Sexual, Violence, Self-Harm) with proprietary data to modify their fundamental behavior. Adaptation is done through configuration and additional features.

Limitations of "Custom Categories":

- Standard: Only works with text and exclusively in English. Requires a minimum of 50 positive examples (max 5k) and a total of 10k samples. Training can take hours and is limited to 3 categories per user. This is a significant barrier for multilingual use cases requiring custom-trained models.
- Rapid: Works with text and images and supports multiple languages (the same as base text moderation). It is faster (no explicit training, uses LLM) but limited to 100 incidents/categories per resource and 1000 samples per incident (text: 500 chars, image: 4MB). It may be less accurate for complex nuances than a trained ML model (Standard).

Dependence on Adaptation Data Quality: The effectiveness of Custom Categories (Standard and Rapid) and Blocklists depends entirely on the quality, representativeness, and quantity of the data/terms provided by the user.

Persistent Linguistic Limitations: Even with Custom Categories (Rapid), detection quality in languages not specifically trained (outside the 8 main ones) can vary and requires exhaustive validation. Key features, such as fact-grounded content verification and protection against sensitive materials, are only available in English.

Limited Contextual Understanding: Although multimodal analysis helps, AI can struggle with sarcasm, irony, very specific cultural contexts, or situations where the text-image combination is ambiguously harmful. It can generate false positives (blocking benign content) or false negatives (failing to detect harmful content).

Foreseen use cases (future specialization in sensitive content detection).

- Filter images and texts uploaded by students to school platforms (forums, chats, assignments).
- Moderate AI-generated content within educational tools used in school.
- Analyze communications in virtual learning environments to detect harassment, hate, or other risks.
- Identify inappropriate content in avatars or usernames on educational platforms.
- Use Custom Categories to detect specific patterns of cyberbullying, radicalization, eating disorder promotion, etc., adapted to European school language and context.
- Implement blocklists with slang or problematic terms specific to the local student population.
- Configure differentiated severity thresholds based on platform type or user age.

9.1.3. Technical Capabilities

Accuracy and recall (if available).

Microsoft does not publish specific precision and recall metrics for the base categories (Hate, Sexual, Violence, Self-Harm) or for the multimodal API.

Documentation mentions that the models are "state-of-the-art" and use advanced technology. It indicates that they aim for accuracy similar to human review.

It is acknowledged that errors (false positives/negatives) can occur, and vulnerabilities allowing detection evasion with certain techniques have been reported.

Azure AI Studio and other Azure tools allow evaluating models (including metrics like precision, recall, F1-score for custom classification tasks in other services like Document Intelligence or CLU), and Content Safety Studio allows monitoring KPIs like blocking rate and category distribution, but it does not provide predefined benchmarks for the base model.

A demonstration video shows an example of threshold adjustment in Content Safety Studio to improve the F1-score (reaching 0.80 in that specific example), suggesting that evaluation and adjustment are possible, but must be performed by the user.

Implication: It is absolutely necessary to perform exhaustive validation with data representative of the multilingual European school environment to determine the

real precision and recall for the categories of interest (both base and custom) before large-scale deployment.

Processing speed (images/second, words/second, etc.).

Documentation does not directly specify speed in units like images/second or words/second.

However, it indicates generous default rate limits for connected APIs: 1000 requests per minute (approximately 16.67 requests per second) per resource, depending on the model. This applies to both text and image/multimodal APIs.

Average latencies reported by third parties for injection attacks are low: 0.070 seconds for Text and 0.29 seconds for Image (although this is for a specific task and may not reflect standard moderation). Azure Monitor allows tracking latency.

Performance can vary depending on load and content complexity. For very high loads, resources can be scaled by deploying multiple instances of the Azure AI Content Safety service.

Disconnected containers have performance metrics that depend on the underlying hardware (CPU, GPU). Examples of RPS and latency for T4 and A100 GPUs are provided, showing significantly higher performance with A100.

Implication: Cloud API speed appears sufficient for the projected daily volume. Container performance will depend on hardware investment.

Fine-Tuning Capability:

Does it allow Fine-Tuning?

No, not in the traditional sense of retraining base models. Adaptation is achieved by:

- Severity Threshold Configuration: Adjusting the level (0, 2, 4, 6) at which content is considered to be flagged or blocked for each base category.
- Blocklists: Creating custom lists of terms or phrases that should be blocked regardless of AI model detection.
- Custom Categories:
    - Standard: Train an ML model (text only, English only) to detect user-defined categories with example data.
    - Rapid: Define "incidents" with descriptions and examples (text and image, multilingual) for an LLM to quickly detect similar content.

How difficult or costly is it?

Threshold/Blocklist Configuration: Easy. Done via API or Azure AI Content Safety Studio. Cost is normal API usage.

Custom Categories (Rapid): Low/medium difficulty. Does not require formal training, just defining the incident and uploading examples. Cost is normal API usage when including the category in the analysis.

Custom Categories (Standard): Medium/high difficulty. Requires collecting and labeling a dataset (minimum 50 positives), configuring training (which can take hours), and evaluating the model. Limited to English and text. Cost includes data storage in Azure Blob Storage and potentially compute time for training (although pricing documentation does not explicitly detail this, the process consumes resources).

Workload:

Can it process between 2000 and 3000 units (images/texts) daily?

Yes, widely. The default API rate limit (1000 requests/minute) allows processing up to 1,440,000 daily requests per resource. The volume of 2000-3000 daily units is well below this limit.

Does it require high CPU, GPU, RAM resources?

Cloud API: Does not require significant local resources. Computational load is managed by Microsoft Azure. Only capacity to make API calls and process responses is needed.

On-premise: Yes, requires high resources.

CPU: Minimum 8 cores recommended.

RAM: Minimum 10-16 GB, recommended 24 GB depending on the specific container (Read, Layout, Invoice, etc. - although Content Safety is not specifically detailed here, a similar or higher need is inferred). 16 GB is recommended.

GPU: Required. Needs NVIDIA CUDA. Minimum T4, A100 recommended for optimal performance. NVIDIA Container Toolkit must be installed on the host.

Implication: Local deployment involves significant investment in specialized hardware (powerful GPUs) and their maintenance.

9.1.4. Licensing and Terms of Use

Is it open source or proprietary?

Proprietary.

License type (Apache 2.0, MIT, commercial license, etc.).

Proprietary commercial license. Use is governed by the terms of the customer's Azure subscription agreement (e.g., Microsoft Customer Agreement - MCA, or

Microsoft Online Subscription Agreement - MOSA), Product Terms, Data Protection Addendum (DPA), and specific terms of the Azure AI Content Safety offering.

Does it allow commercial use?

Yes. Azure AI services, including Content Safety, are designed for commercial use, provided that service terms, responsible AI use requirements, and Microsoft content policies are met. A paid subscription (Pay-as-you-go or higher) must be used for commercial use; free or trial accounts may have restrictions. Preview versions may not be suitable for use in production systems because they may have limitations or not be fully stable.

Approximate cost:

Annual or monthly license:

Cloud API: No explicit licensing cost. It's pay-per-use. There is a limited free tier (5,000 text records and 5,000 images/multimodal per month).

On-premise: Requires the purchase of an annual commitment plan. Prices are high.

Cost per use:

Cloud API (Standard Tier):

- Text: $0.38 per 1,000 text records. A "text record" is up to 1,000 Unicode characters. Longer texts are counted as multiple records (e.g., 7,500 characters = 8 records). This includes base text analysis, Prompt Shields, Protected Material, and Groundedness.
- Image/Multimodal: $0.75 per 1,000 images (multimodal analysis is billed per image, associated text up to 1k characters is included in that price).

API Cost Example: Processing 3,000 image+text combinations daily (~90,000 per month): Image/Multimodal Cost: (90,000 / 1,000) * $0.75 = $67.50 per month.

If additional text analysis were performed separately (e.g., for chats), the text cost would be added.

9.1.5. Legal and Ethical Analysis

GDPR compliance and European regulations.

Azure Platform: Microsoft Azure as a platform complies with GDPR and offers tools and agreements to help customers meet their own obligations. Microsoft acts as Data Processor and the customer (the school organization) as Data Controller.

Data Processing Agreement (DPA): Microsoft provides a DPA (Microsoft Products and Services Data Protection Addendum) covering data processing in Azure. These

terms must be reviewed and accepted. A specific agreement with Microsoft Ireland Operations Ltd. for processing must be established.

Data Transfers: Azure offers options to keep processed data within specific geographies (e.g., Europe). However, some operations or deployment types ("Global") may involve processing outside the designated geography, even if data at rest remains there. Transfers to the USA are based on adequacy decisions like the Data Privacy Framework (DPF).

Controller Responsibility (School/Organization): The school/organization is responsible for:

- Having a legal basis for processing (likely consent or legitimate interest, with careful evaluation).
- Conducting a Data Protection Impact Assessment (DPIA) if processing is considered "high-risk" (very likely when processing minors' data and sensitive content at scale).
- Informing users (students, staff) about processing (transparency).
- Managing Data Subject Rights Requests (DSRs).
- Implementing appropriate technical and organizational measures to protect data.

Protection of Minors: GDPR has specific provisions for processing minors' data. Any implementation must strictly consider these rules, especially regarding the legal basis and clarity of information provided.

Bias handling (racial, gender, cultural...).

Inherent Risk: Like all AI models trained with large datasets, Azure Content Safety can inherit and potentially perpetuate existing biases in the data. This could lead to uneven moderation (e.g., being stricter with certain groups or failing to detect culturally specific forms of hate).

Mitigation by Microsoft: Microsoft states it works on bias mitigation through techniques like "debiasing" and the use of diverse training data. They promote Responsible AI principles, including Fairness. They offer tools like Fairlearn (open source) to evaluate and mitigate biases in ML models, although its direct applicability to tuning Content Safety is limited.

Need for Audit: It is crucial for the user organization to actively audit model results in their specific context to detect potential biases. This includes analyzing whether certain types of content from specific groups are unfairly blocked or allowed.

Custom Categories and Local Data: Using Custom Categories trained with diverse and representative local data could help mitigate some context-specific biases but could also introduce new biases if training data is not carefully selected.

Explainability capacity ("explainability") of decisions.

Model Outputs: The service provides the detected harm category (Hate, Sexual, etc.) and a severity level (0-6). If a blocklist or Custom Category is used, the API will indicate which specific list or category caused the detection.

Groundedness: The groundedness detection feature allows identifying phrases that may not be based on real or verifiable facts and explains why they are considered to lack a solid basis.

Limitations: The service does not explain why the model assigned a specific severity to content within a base category (e.g., why something is "Hate" level 4 vs. level 2). The internal logic of the deep learning model is largely a "black box."

Transparency: Microsoft publishes Transparency Notes for some AI services to explain how the technology works, but this does not reach the level of explaining individual decisions in real-time.

Implications: The lack of detailed explainability can hinder the appeal of moderation decisions and the building of trust. For complex or sensitive cases, a human review process may be necessary to complement the AI's decision.

9.1.6. Infrastructure Requirements

Can it be deployed locally or does it require the cloud?

Both options are possible:

- Cloud (API): This is the primary mode of use. Requires Internet connectivity to send requests to Azure endpoints and receive responses. It is managed by Microsoft.
- Locally: Allows running text and image analysis models on the client's own infrastructure, without an Internet connection for normal operation (although it requires initial connection for configuration and license download, and potentially periodically to report usage if a fully disconnected plan is not purchased). This option requires prior approval from Microsoft and the purchase of an annual commitment plan.

Is it feasible to have a server per school or a global server instead?

Global Server (Using Cloud API): This is the most viable and scalable option for most cases. A single Azure Content Safety endpoint (or several for

redundancy/scalability) can serve multiple schools. Management is centralized, and cost is per-use.

Server per School: Technically possible, but unfeasible due to the cost of both the commitment plan and the infrastructure.

Global Server (on-premise): If local deployment is chosen for privacy or disconnection reasons, a powerful centralized server (or cluster) serving all schools would be more logical from a cost and management perspective, although it would still be a considerable investment.

Estimated infrastructure consumption (CPU, GPU, RAM).

Cloud API: Minimal. Only the infrastructure needed to make HTTP REST calls or use Azure SDKs is required.

On-premise: High.

CPU: 8+ cores recommended.

RAM: 16 GB - 24 GB recommended.

GPU: Indispensable. NVIDIA GPU with CUDA support (T4 minimum, A100 recommended for optimal performance). Multiple GPUs may be needed for high performance or concurrency.

9.1.7. Advantages and Limitations

Aspect	**Advantages**	**Limitations**
Accuracy	"State-of-the-art" base models for 4 categories. Contextual multimodal analysis. Adaptation improves accuracy for specific cases.	No public benchmarks. Requires exhaustive user validation. Can generate false positives/negatives. Limited in untrailed cultural/linguistic nuances.
Cost	Cloud API: Low initial cost, pay-per-use ($0.38/1k texts, $0.75/1k images). Free tier available.	On-premise: Very high annual fixed cost ($100k-$200k+). Local infrastructure cost (GPU). Possible hidden cost of personnel for Custom Categories.
Ease of Integration	SDKs available (Python, C#, Java, JS). Well-documented REST API. Integration with Azure ecosystem (Studio, Monitor).	Requires development effort for integration logic and moderation workflow.

Explainability	Provides category/severity. Blocklist/Custom Cat hits are clear. "Reasoning" mode for Groundedness.	Little information on the reason for severity assigned by base models. May require additional XAI tools for deep analysis.
Flexibility for Fine-Tuning	Custom Categories (Standard/Rapid). Severity threshold configuration. Custom blocklists.	No direct fine-tuning of the base model. Custom Cats Standard only English/text. Limits on number of categories/samples. Quality depends on user data.
Regulatory Compliance	Azure platform complies with GDPR. DPA available. Data residency options (regions, containers).	User (Controller) responsibility for implementation. Careful configuration required.

9.1.8. Conclusions and Recommendations

Is this model suitable for our use case?

Azure AI Content Safety is potentially suitable but requires significant adaptation and validation for the specific use case of sensitive content detection in multilingual school environments.

Strengths: Offers native multimodal (image+text) capability, robust detection in core categories (hate, violence, sexual, self-harm), customization options (thresholds, blocklists, Custom Categories), inherent scalability of the cloud API, and integrates into the Azure ecosystem.

Weaknesses and Critical Challenges:

- The specific risk categories for the school environment (harassment, drugs, radicalization, etc.) are not predefined and mandatorily require the use of Custom Categories.
- The Custom Categories Standard functionality, which allows training more precise ML models, is limited to text in English, which is a significant barrier for a multilingual environment like Europe. Reliance would be on Custom Categories Rapid (multilingual but potentially less precise for nuances) or blocklists for other languages.
- Absence of public precision/recall benchmarks, making rigorous internal validation indispensable before deployment.

- Risk of biases and errors (false positives/negatives), especially in diverse cultural and linguistic contexts.
- Local deployment via containers is technically complex and economically very costly.

What conditions or adjustments would be necessary to use it?

Deployment: Cloud API is strongly recommended due to its lower cost, scalability, and ease of management, unless strict disconnection or data residency requirements justify the complexity and cost of containers.

Configuration: Carefully adjusting severity thresholds for base categories, based on the school's specific policies and pilot test results, is essential. Implement blocklists with relevant local terms and slang.

Customization (Critical):

- Define a clear set of specific risk categories for the school environment (e.g., School Bullying, Drug Advocacy, Radicalization, etc.).
- Collect representative datasets (images and texts, in multiple languages) for each category, ensuring ethical and diverse origins.
- Use Custom Categories Standard for English content (if applicable), accepting the training and evaluation effort.
- Use Custom Categories Rapid for content in other languages and for images, being aware of its limitations and need for validation.

Rigorous Validation: Conduct mandatory pilot tests using real and diverse data from the target European schools. Measure precision, recall, F1-score, and false positive/negative rates for each relevant category (base and custom) in each main language.

Legal and Ethical Compliance: Formalize the DPA with Microsoft. Conduct a comprehensive DPIA. Establish clear processes for DSR management and breach notification. Ensure transparency to users about tool usage. Implement measures to audit and mitigate biases.

Workflow: Design and implement a clear workflow for moderation, which will likely include human review for flagged content, especially in ambiguous cases, appeals, or sensitive categories like self-harm or radicalization.

9.1.9. Visual Summary

Category	Result
Model Type	Multimodal (Image + Text)
Accuracy	Medium/High (Potential, requires extensive validation and customization for specific school risks)
Cost	Cloud API: Low ($0.75/1k images+text). Disconnected Container: Very High ($100k-$200k+/year + hardware).
License	Proprietary Commercial (Azure Terms).
Fine-Tuning	No direct. Adaptation via Custom Categories (Standard: Text/English, medium difficulty; Rapid: Text-Image/Multi, low difficulty), Threshold Configuration, Blocklists.
GDPR Compliance	Yes (Azure Platform and DPA available). Requires implementation and responsibility from the Data Controller (user).
Final Recommendation	Requires adaptation and exhaustive validation. Suitable as a base if investment is made in customization (Custom Categories, blocklists) and rigorous validation.

10. Case Studies and Pilot Projects of AI Implementation in Educational Institutions

10.1 Introduction

This chapter, while offering a global perspective on various AI implementations in education, lays the groundwork for understanding how these technologies can be (and in some cases already are being) adapted or specifically designed for sensitive content detection. We will explore cases and pilot projects that, although not all focused exclusively on risk detection, provide relevant technological and implementation context for the broader report's main focus on AI models for sensitive content detection in European school environments. AI's capabilities to analyze large volumes of textual and visual data, identify patterns, and flag anomalies are fundamental for these specific applications, although their implementation poses even greater ethical and practical challenges.

The Role of International Organizations (UNESCO, OECD, World Bank) in Shaping Educational AI

International organizations such as UNESCO, the OECD, and the World Bank play a crucial role in shaping the global discourse and establishing guidelines for the implementation of AI in education.

UNESCO leads the promotion of a human-centered approach to AI, emphasizing ethics, inclusion, and equity. Its goal is to ensure that AI does not widen technological gaps and that its benefits, such as access to knowledge and innovation, are accessible to all. It has published key documents such as the "Recommendation on the Ethics of Artificial Intelligence" (2021), the first global standard on the subject adopted by 193 Member States, the guide "Artificial intelligence and education: A guide for policy-makers," and AI competency frameworks for students and teachers. UNESCO organizes international forums on AI and education and supports initiatives in various regions, including Africa.

The Organisation for Economic Co-operation and Development (OECD) contributes with analyses of national AI strategies, the use of AI in the public sector (including education), and the promotion of responsible innovation. Its Observatory of Public Sector Innovation (OPSI) and the AI Policy Observatory provide valuable resources for understanding how governments are using AI. The OECD has also established AI principles that aim to foster trust and guide policy development, aligning with the objectives of the Learning Compass 2030, which emphasizes general competencies and student agency.

The World Bank recognizes the transformative potential of AI to personalize learning, support teachers, and optimize educational management. It emphasizes the importance of developing these tools in close collaboration with education experts and cognitive scientists to ensure that they are based on the science of human learning and effectively address educational needs.

The concerted activity of these international organizations is fundamental to guiding the adoption of AI in education worldwide. Their guidelines, frameworks, and recommendations are prompting governments and institutions to explore AI, although responsiveness and infrastructure vary enormously across countries and regions. This external pressure, combined with the growing availability of AI tools, fosters an environment of experimentation and policy development, although the lack of formal institutional guidelines suggests that adoption is often more reactive or fragmented than strategic.

10.2. Government Strategies and National AI Initiatives in Education: A Global Mosaic

10.2.1 Analysis of governmental approaches and policy frameworks in different regions

Globally, there is a growing trend for governments to recognize the strategic importance of AI and, consequently, to formulate specific national plans for its integration into various sectors, including education. According to the OECD, of

the 50 countries with general AI strategies, 36 have developed or plan to develop specific strategies for the public sector, which often encompass education. However, approaches, policy maturity, and implementation levels vary considerably across regions and countries.

10.2.2 Asia-Pacific

This region shows remarkable dynamism, with several countries leading the implementation of AI policies in education.

China has adopted a strong government-driven approach, establishing a robust regulatory framework, such as the 2023 generative AI law. Its national strategy emphasizes AI integration from primary education, actively promotes public-private collaboration with tech giants like Tencent and Alibaba, and EdTech companies like Squirrel AI, and assigns a central role to universities in research, development, and talent training.

South Korea launched its National AI Strategy in 2019 and AI Ethics Guidelines in 2020. Its plan for public education is particularly ambitious, highlighting the national development and deployment of AI-powered digital textbooks to personalize learning, planned between 2025 and 2028. This initiative is accompanied by strong investment in teacher training and the improvement of school technological infrastructure.

Singapore is another regional leader, with clear policies and guidelines for AI in education. Its Ministry of Education (MOE) has actively implemented AI solutions for learning personalization and automated grading, seeking to optimize processes and support teachers.

India has focused its National AI Strategy on leveraging this technology for social good, with priority applications in education, health, and agriculture.

Australia published the "Australian Framework for Generative AI in Schools" in 2023, establishing principles on teaching, well-being, transparency, fairness, accountability, and safety. There is active debate on the need for a more coherent national policy, student data protection, and adequate teacher training.

Japan has AI Governance Guidelines since 2022 and is working on integrating AI into curricula and teacher training.

10.2.3 North America

In the **United States**, the approach is more decentralized. At the federal level, programs like the FCC's "Schools and Libraries Cybersecurity Pilot Program" seek to fund cybersecurity equipment and services, including AI-based ones, for schools. Initiatives like CISA's explore the use of AI to detect vulnerabilities. At the

state level, specific pilot programs emerge, such as the school safety with AI program in North Carolina. However, the adoption of student monitoring technologies generates considerable debate on privacy and ethics. Frameworks like "SAFE" (Safety, Accountability, Fairness, Transparency, and Efficacy) attempt to guide responsible integration.

Canada was a pioneer in launching the first National AI Strategy in 2017.

10.2.4 Europe

The European Union and its member states have adopted a markedly regulatory and ethical approach.

The **European Union** includes AI and data usage as priorities in its Digital Education Action Plan (2021-2027). It has published "Ethical guidelines on the use of AI and data in teaching and learning." The EU AI Act, a horizontal regulatory framework, classifies certain educational applications (assessment, admission) as high-risk and prohibits uses considered unacceptable, such as emotion detection in educational or work contexts. Compliance with the General Data Protection Regulation (GDPR) is a fundamental requirement for any implementation.

The **United Kingdom**, although outside the EU, follows a similar line with its National AI Strategy (2021) and government projects to foster AI tools supporting teachers.

France has had a national strategy (SNIA) since 2018, investing in research and talent, and seeking to position its universities as global leaders in AI through the 'AI Cluster' initiative. Its data protection authority (CNIL) actively promotes privacy-respecting AI.

Germany updated its National AI Strategy in 2020, seeking to balance competitiveness with responsible development and ethical integration. Educational policies exist at the state level for the implementation of digital technologies.

Spain participates in regional and European initiatives and projects, applying the EU regulatory framework.

10.2.5 Latin America and the Caribbean

The region shows growing interest, often driven by multilateral organizations.

The **Inter-American Development Bank (IDB)** and the **Organization of Ibero-American States (OEI)** collaborate to promote digital transformation and the use of AI in regional education. The IDB initiative supports pilot projects with AI components in several countries.

Uruguay, through its Plan Ceibal, is a regional benchmark in educational technology and has implemented AI-based virtual assistants.

Argentina is developing initiatives for AI literacy and the creation of resources for teachers.

Countries **like Brazil, Chile, Ecuador, Mexico, and Peru** participate in IDB pilot projects focused on challenges such as student assignment, school dropout prediction, and teacher allocation.

10.2.6 Africa

AI adoption in education is in earlier stages, with strong support from international organizations.

UNESCO and other partners actively support AI integration, focusing on capacity building, policy development, and adapting solutions to the African context.

By 2024, six sub-Saharan countries **(Kenya, Nigeria, Uganda, Mauritius, Rwanda, South Africa**) had adopted National AI Strategies.

Concrete projects exist such as **RobotsMali** (content creation in local languages) and the **STEPS** project (science textbooks in Benin, Cameroon, DRC) which demonstrate practical and contextualized AI applications. The "**Technology-enabled Open School Systems**" project (UNESCO-Huawei) in Egypt, Ethiopia, and Ghana seeks to build resilient school systems.

10.2.3 Fostering pilot projects and public-private collaboration

A common feature in many national strategies is the use of pilot projects as a step prior to large-scale implementations. These pilots allow governments and educational institutions to test technologies, refine AI-based teaching methodologies, evaluate their real impact, and adapt solutions to specific contexts before committing significant resources. Notable examples include the testing of digital educational materials and digital textbooks in South Korea, pilot programs in selected cities in China, and school safety pilots in the USA.

Public-private collaboration is another fundamental pillar in numerous national strategies, particularly visible in Asia, but also present in global initiatives such as the UNESCO-Huawei one in Africa. These partnerships seek to combine strategic vision and public resources with the agility and technological expertise of the private sector to accelerate innovation, the development of tools and platforms, and the implementation of advanced educational solutions.

The global landscape of AI strategies in education reveals considerable diversity, but also a shared recognition of the importance of this technology. Countries like China and South Korea demonstrate centralized and ambitious planning, investing heavily in infrastructure, content development, and teacher training. This strong investment and explicit political support will likely accelerate adoption and generate a large amount of data on the effectiveness and challenges of AI in real educational contexts. Lessons learned (and potential mistakes) in these pioneering countries will undoubtedly be closely observed and could offer valuable information for other nations in earlier stages of developing and implementing their own strategies.

However, this governmental drive, often motivated by global economic competitiveness objectives and the pursuit of efficiency in educational systems, coexists with a growing current of concern expressed by educators, unions, civil society organizations, and regulatory bodies. These concerns focus on ethical risks (data privacy, algorithmic biases, equity in access), the need to maintain human and pedagogical control, the lack of solid evidence on learning benefits, and the potential additional burden on teachers. This fundamental tension between the promised potential of technology and the inherent risks of its implementation defines the complex path that AI is charting in education worldwide.

10.3. Case Studies and Pilot Projects Highlighted by Region

AI implementation in educational settings manifests in diverse ways around the world. Below are representative case studies and pilot projects from different regions, detailing their objectives, technologies employed, observed results, and key lessons learned from their experience.

10.3.1 North America (USA, Canada)

Georgia Institute of Technology (USA): AI Assistant "Jill Watson"

- Objectives: Alleviate the workload of teaching assistants (TAs) in massive online courses, efficiently answering frequently asked student questions.
- Technologies: IBM Watson platform, Natural Language Processing (NLP).
- Results/Impact: Drastic reduction in student query response time. Jill Watson handled most routine questions with a reported 97% accuracy, freeing up humans to address more complex and personalized issues.
- Learnings/Challenges: Demonstrated the potential of AI to scale educational support services efficiently. Underlined the critical importance of a well-curated training dataset and the need for continuous monitoring and adjustment to maintain AI system performance and relevance.

Ivy Tech Community College (USA): Student Risk Prediction

- Objectives: Early identification (within the first two weeks of the semester) of students at high risk of failing courses, allowing for proactive and personalized interventions.
- Technologies: AI-based predictive analysis system using historical data on student performance, attendance, and other academic indicators, employing machine learning algorithms.
- Results/Impact: The system achieved 80% accuracy in predicting at-risk students. In a pilot study across 10,000 course sections, 16,000 at-risk students were identified. Thanks to targeted interventions (addressing non-academic obstacles), it was reported that 3,000 students were "saved" from failing, and 98% of contacted students achieved a grade of C or higher. The "Project Student Success" program has assisted over 34,700 students.
- Learnings/Challenges: AI can be a powerful tool for student retention by enabling early and targeted interventions. Data quality and integration are fundamental for model accuracy. Ethical considerations regarding privacy and bias require continuous auditing.

Brainly (Global online platform): Homework Assistance

- Objectives: Facilitate obtaining help with homework, especially via mobile devices, by providing instant and relevant answers through image recognition.
- Technologies: Google Cloud Vision AI to process photos of questions, multilingual capabilities for global reach.
- Results/Impact: Achieved 70% student satisfaction and increased interaction through photo queries sixfold. Contributed to an increase in subscription revenue and democratized access to educational help.
- Learnings/Challenges: Image recognition using AI is effective for improving accessibility and participation on educational platforms. Cloud infrastructure scalability was crucial to handle increased demand, especially during remote learning.

Other relevant initiatives in the USA: Include AI school safety pilot programs in North Carolina (using Eviden technology to detect physical threats) and Palm Beach County (using Lightspeed Alert to monitor digital activity for risks). Institutions like MIT and Stanford University also explore AI for student retention and adaptive learning, with Stanford developing a "wheel-spinning" predictor for stuck students.

10.3.2 Europe (EU and individual countries)

University of Alicante (Spain): "Help Me See" Application

- Objectives: Improve campus and educational material accessibility for visually impaired students, fostering their independence.
- Technologies: Computer vision and machine learning to recognize and narrate objects, texts, and environmental elements.
- Results/Impact: Significant improvement in students' ability to navigate the campus independently and safely, increasing their confidence and participation in university life.
- Learnings/Challenges: Demonstrated AI's transformative impact on creating inclusive educational environments. Highlighted the critical importance of user-centered design, ensuring technology responds to real user needs to be truly effective.

Harris Federation (United Kingdom): Workload Management and Linguistic Accessibility

- Objectives: Reduce the time teachers spend on administrative tasks and adapting materials for students from diverse linguistic backgrounds.
- Technologies: ChatGPT for text adaptation and summarization; Microsoft Live for simultaneous translation and real-time subtitles.
- Results/Impact: Considerable reduction in time spent by teachers on these tasks, allowing them to focus more on direct teaching and student interaction. Improved curriculum accessibility for non-English speaking students.
- Learnings/Challenges: AI can be a valuable tool to support teaching staff and overcome linguistic barriers in diverse classrooms.

Oak National Academy (United Kingdom): Improving Digital Curriculum Resources

- Objectives: Use AI to continuously improve digital curriculum resources, reduce teacher workload associated with planning and material creation.
- Technologies: AI tools to assist with lesson planning and creating quizzes and formative assessments.
- Results/Impact: The explicit goal of the government initiative was to reduce teacher workload by up to five hours per week, streamlining resource creation.
- Learnings/Challenges: Illustrates AI's potential to optimize educational administration and resource development. Emphasizes the need for

continuous collaboration between educators and AI developers to ensure tools are practical and genuinely alleviate workload.

Berlitz (Global company with strong presence in Europe): Language Learning

Objectives: Adapt to the demand for flexible online learning, maintaining a focus on oral practice and overcoming the limitations of traditional pronunciation assessment tools.

Technologies: Microsoft Azure AI Speech, specifically its pronunciation assessment and text-to-speech (TTS) capabilities to generate voices with various accents.

Results/Impact: Significant improvement in the learning experience for 500,000 users. Greater accessibility for diverse learning profiles. Successful launch of new products and reduction in development costs and time.

Learnings/Challenges: Voice AI can drastically improve online language learning, especially oral practice. Accuracy in assessment (low number of false negatives) is crucial. AI can increase accessibility and efficiency in educational product development.

European University of Madrid (Spain): Multiple Applications

- Objectives: Implement AI to personalize training, increase student motivation, offer realistic simulations, facilitate language learning, and promote universal access.
- Technologies: AI for personalized learning, advanced simulation platforms (e.g., Simulated Hospital with mannequins that react in real time), real-time translation and subtitle generation tools.
- Results/Impact: Reports of increased student interest and motivation. Practical and safe learning experiences in simulated environments. Improved comprehension and language learning for an international student population (35%). Greater accessibility for students with hearing disabilities or different languages.
- Learnings/Challenges: AI offers a versatile toolkit to address multiple challenges and opportunities in higher education, from pedagogy to inclusion.

University of London Worldwide (United Kingdom): AI Tutor "Walter"

- Objectives: Investigate the impact of an AI tutor ("study buddy") on engagement, performance, and tutor support in distance learning law courses (undergraduate and postgraduate).
- Technologies: AI tutor "Walter" developed by Noodle Factory.

- Results/Impact: (Qualitative evaluation ongoing/completed). Preliminary findings emphasize the irreplaceable value of human interaction in the educational process and the critical need for careful pedagogical integration of AI tools.
- Learnings/Challenges: The development of comprehensive ethical guidelines and the promotion of human-AI collaboration models are recommended to ensure technology complements, rather than replaces, traditional teaching methods. Continuous evaluation and feedback from all stakeholders are crucial.

10.3.3 Asia-Pacific (China, South Korea, Singapore, Australia, India, Pakistan, Japan, Indonesia)

Ministry of Education of Singapore (Singapore): Integrated AI Solutions

- Objectives: Provide personalized learning experiences to a diverse student population and optimize grading and feedback processes for teachers.
- Technologies: Automated English exam grading systems (using NLP) for primary and secondary school; adaptive learning systems enhanced with machine learning.
- Results/Impact: Significant reduction in teacher workload associated with grading, allowing more time for direct interaction. Effective personalization of education to meet individual paces and needs.
- Learnings/Challenges: Successful integration requires AI to complement existing pedagogical methods. Continuous teacher training and careful management of ethics and data privacy are essential.

China (Multiple government and private initiatives): Systemic Integration

- Objectives: Integrate AI into the national curriculum from basic levels to university; personalize the educational experience at scale; position China as a global leader in AI.
- Technologies: Wide range of AI-based educational platforms developed in public-private collaboration (Tencent, Alibaba, Squirrel AI, etc.); extensive use of machine learning for personalization.
- Results/Impact: Rapid development and expansion of the AI-based EdTech market. Specific studies, such as one in legal education using generative AI, showed significant improvements in learning outcomes and student "flow" experience.
- Learnings/Challenges: Strong government direction and public-private collaboration act as accelerators of innovation. AI is a national strategic priority with deep educational implications.
-

South Korea (National educational system): AI Digital Textbooks

- Objectives: Transform public education through deep AI integration to improve quality, personalize learning, and support teachers in the face of declining student populations.
- Technologies: Development of AI digital textbooks that collect real-time performance data, adapt content (pace, difficulty), and provide detailed information to teachers. Improved network infrastructure and device distribution (1 per student).
- Results/Impact: National deployment planned from March 2025, initially covering key subjects in various grades, with expansion to all schools by 2028. Korea positions itself as a global pioneer in this national-scale implementation.
- Learnings/Challenges: Focus on continuous improvement through pilot testing and rigorous impact evaluation. Strong regulatory framework for data privacy (aligned with Korea's Personal Information Protection Act) and guidelines for responsible AI development.

Pakistan (Pilot with Knowledge Platform): Personalized Learning

- Objectives: Improve student academic performance through AI-powered personalized learning paths.
- Technologies: "Knowledge Platform" platform offering adaptive learning paths and AI-based content creation tools.
- Results/Impact: A pilot program involving 75 schools and 26,000 students reported an average 60% improvement in grades for students who used the personalized approach.
- Learnings/Challenges: Personalized AI implementation can have a significant quantitative impact on student performance, demonstrating its potential in resource-limited contexts.

Singapore (AICET): "Codavari" and "Cikgo" Tools

- Objectives: "Codavari" acts as an AI-based programming "coach" to offer quality learning support. "Cikgo" personalizes the learning experience by adapting to individual needs, helping teachers manage larger classes.
- Technologies: Specific AI platforms for programming tutoring and general adaptive learning.
- Results/Impact: Tools designed to increase teaching capacity and provide individualized support where access to traditional instruction might be limited.

- Learnings/Challenges: AI can play a role in mitigating the shortage of qualified teachers and improving classroom management in large classes, acting as an intelligent assistant.

Japan and Latin America (Edwin AI): English Language Learning

- Objectives: Offer an affordable and personalized solution for English as a foreign language learning, with individualized practice.
- Technologies: Combination of adaptive learning, natural language understanding (NLU), and AI-based speech recognition and evaluation technology.
- Results/Impact: Over 800,000 students in the region have used the platform to improve their English skills.
- Learnings/Challenges: AI can effectively scale personalized language learning, making it accessible to a large number of students in diverse geographies.

Australia (New Town High School): STEM Improvement

- Objectives: Increase engagement and learning outcomes in STEM (Science, Technology, Engineering, and Mathematics) subjects, providing individualized support in large classes.
- Technologies: Machine learning to analyze student performance and adapt support.
- Results/Impact: A notable improvement in mathematics performance was observed, with increased student engagement and better test scores. Teachers were able to identify and address individual learning gaps more effectively.
- Learnings/Challenges: AI can act as a "force multiplier," enabling teachers to implement personalized learning strategies that would be difficult to manage manually in large classes.

Indonesia (Prakerja Platform): Training and Employment

- Objectives: Connect training with labor market needs, offering personalized training and relevant employment recommendations.
- Technologies: AI for personalizing training modules, analyzing acquired skills, recommending jobs, and performing "liveness checks" to validate participation.
- Results/Impact: Successful implementation highlighted as a regional example of AI use for employability.

- Learnings/Challenges: AI can create more effective bridges between education/training and the world of work, adapting training offerings to market demand.

Los Angeles Pacific University (USA, relevant study): AI Course Assistants

- Objectives: Evaluate the impact of AI-based course assistants (developed by Nectar) on the learning experience (grades, motivation, self-efficacy, engagement) in online university courses.
- Technologies: AI course assistants.
- Results/Impact: Statistically significant improvement in student grades and intrinsic motivation in the group with access to AI assistants. Positive impact also on self-efficacy. No significant differences were found in feelings of engagement, encouragement, and support, possibly due to the high-interaction model already existing at the university.
- Learnings/Challenges: AI assistants can improve key aspects of online academic performance and motivation. More research is needed on long-term effects and their interaction with other support mechanisms.

10.3.4 Latin America and the Caribbean (Regional and national initiatives)

fAIr LAC Projects - IDB (Various countries: Uruguay, Ecuador, Brazil, Chile, Peru, Mexico)

- Objectives: Address key systemic challenges such as school dropout, inefficient student and teacher allocation, academic risk prediction, and centralization of admission processes.
- Technologies: Machine learning for predictive risk models (dropout, failure), optimization algorithms for centralized allocation, virtual assistants for support.
- Results/Impact: Projects in various phases, from pilot design to implementation. The overall objective is to systematize learnings on AI applications with social impact and foster scalability and replication in the region.
- Learnings/Challenges: AI is being explored as a tool to improve efficiency and equity in the management of the region's educational systems, addressing structural problems.

Plan Ceibal (Uruguay): Virtual Assistant

- Objectives: Support national educational transformation, improve learning quality, and social equity through technology.

- Technologies: Implementation of a virtual assistant to provide information and support to students and educators within the Ceibal ecosystem.
- Results/Impact: Integral part of a long-standing national policy for technology integration into education.
- Learnings/Challenges: Uruguay serves as a case study on the sustained integration of technology (and now AI) into a national educational system.

Capybara AI (Chile): Bullying Prevention

- Objectives: Go beyond direct bullying detection, measuring peer cooperation and analyzing classroom social dynamics to identify risks and positive leadership early on.
- Technologies: AI to analyze social interactions (possibly through digital surveys or network analysis) and calculate cooperation, popularity, aggression indices, etc.
- Results/Impact: The pilot experience revealed students with positive leadership qualities who had not been identified by traditional methods, offering a more nuanced view of classroom dynamics.
- Learnings/Challenges: AI can offer complementary tools to teacher observation for understanding and fostering positive social relationships and proactively preventing bullying.

Mendoza (Argentina): Teacher Training in AI

- Objectives: Digitally literate teachers in the pedagogical use of AI, providing courses, manuals, and assistants for creating didactic sequences.
- Technologies: Focused on the practical application of AI tools in the classroom.
- Results/Impact: Subnational governmental initiative to train teachers and management teams in the effective and responsible use of AI.
- Learnings/Challenges: Recognition of the critical need for teacher professional development as a prerequisite for successful AI integration in schools.

10.3.5 Africa (Regional and national initiatives)

RobotsMali (Mali): Creation of Children's Books in Bambara

- Objectives: Address the scarcity of culturally relevant reading materials in Bambara, an important local language, to promote early literacy.
- Technologies: Combined use of generative AI (ChatGPT) for initial drafting, machine translation, and human oversight/editing.
- Results/Impact: Accelerated production (over 180 books in less than a year) and at a significantly reduced cost compared to traditional methods. Great potential to impact literacy in Bambara.
- Learnings/Challenges: Generative AI, guided by humans, can be a powerful tool for rapid and economical creation of localized educational content, especially in resource-limited contexts and minority languages.

STEPS Project (Benin, Cameroon, DRC): Science Textbooks

- Objectives: Develop high-quality primary science textbooks aligned with national curricula and culturally relevant for each country.
- Technologies: Combination of Open Educational Resources (OER) with AI to assist in initial drafting, suggest contextualized examples, and facilitate translation and localization of materials.
- Results/Impact: Cited as a concrete example of responsible and inclusive use of AI in African education, led by Global South organizations and contributing to the Sustainable Development Goals.
- Learnings/Challenges: AI can facilitate curriculum adaptation and the creation of educational materials that better respond to the specific needs and cultural contexts of African students.

Maseno University (Kenya): English-Kenyan Sign Language Translator

- Objectives: Improve the inclusion of deaf students by facilitating communication with peers and teachers.
- Technologies: AI-based translation tool, developed in collaboration with the deaf community.
- Results/Impact: Development of a functional tool designed to address a specific communication barrier in the Kenyan educational context.
- Learnings/Challenges: AI can create vital accessibility tools. Collaboration with end-user communities is essential to ensure the relevance and effectiveness of these tools.

"**Technology-enabled Open School Systems**" Project (UNESCO-Huawei in Egypt, Ethiopia, Ghana)

- Objectives: Build crisis-resilient school systems, enabled by technology, that integrate in-person and distance learning to ensure educational continuity and quality.
- Technologies: Focus on digital infrastructure, learning platforms, and open educational resources, with AI as a potential component within this broader technological framework.
- Results/Impact: Three-year project (2020-2023) focused on the design, piloting, and scaling of these systems in the three participating countries. Included teacher training in distance learning.
- Learnings/Challenges: Highlights the importance of educational system resilience and the need for flexible hybrid models, a key lesson reinforced by the COVID-19 pandemic.

UNESCO/KIX Africa Seminar (Dakar, Senegal)

- Objectives: Foster conversation and collaboration among French-speaking and Portuguese-speaking African countries, aiming to improve teachers' and students' digital and artificial intelligence (AI) skills, referencing UNESCO's established guides or frameworks.
- Technologies: Discussion focused on conceptual frameworks and strategies for developing AI competencies, rather than specific technologies.
- Results/Impact: Renewed commitment from participating countries and organizers (UNESCO, KIX) to support AI integration into educational systems, with an emphasis on contextual adaptation and the use of African languages.
- Learnings/Challenges: South-South collaboration and adapting global frameworks to African realities and needs are crucial for successful AI competency development on the continent.

A cross-sectional analysis of these cases reveals that most documented pilot projects to date have focused predominantly on "general" AI applications, such as learning personalization, administrative efficiency, teacher support, and improving access to resources. While sensitive content detection is an area of growing interest, concrete implementations in this field appear to be less numerous or in earlier stages, especially outside the USA. However, the technological infrastructure, data analysis capabilities, and experience accumulated in these general projects are fundamental. They establish a foundation upon which future applications for sensitive content detection could be built or integrated.

The lessons learned from these pioneering projects are directly relevant and, in fact, even more critical when considering the use of AI to monitor and detect sensitive content. The imperative need for careful pedagogical integration, ensuring data quality and representativeness, ethical and technical training for teachers, user-centered design, and multidisciplinary collaboration are magnified in the context of risk detection, where the implications of an error or bias can be particularly severe. If the implementation of an AI tutor already raises ethical dilemmas, a system designed to actively monitor student communication exponentially intensifies concerns about privacy, bias, and the potential for harm.

Finally, a pattern of significant regional disparity emerges. While nations like South Korea, China, and the United States are implementing or testing large-scale systems, often with strong governmental backing or significant private investment, many other regions, particularly in Africa and Latin America, focus on more foundational projects, capacity building, or smaller-scale pilots, frequently with the crucial support of international organizations like UNESCO or the IDB. This gap has direct implications for the ability to implement complex and costly sensitive content detection systems, which require not only advanced technology but also robust infrastructure, trained personnel, and well-established regulatory and ethical frameworks.

10.4. Specific AI Applications for Sensitive Content Detection in School Environments

10.4.1 Contextualization: The growing need for safe learning environments

Student safety and well-being are paramount concerns for educational communities worldwide. The digital environment, while offering countless learning opportunities, also exposes young people to a variety of risks, including cyberbullying, violence, sexual exploitation, radicalization, disinformation, and content that can exacerbate mental health problems such as eating disorders or suicidal ideation. The increasing prevalence of these risks, along with the difficulty of manually monitoring the vast amount of digital interactions, has prompted schools and districts to explore the use of Artificial Intelligence as a potential tool for early detection and intervention. These technologies promise to analyze online communications and activities for indicators of danger, alerting responsible staff so they can act.

10.4.2 Detection of School Bullying and Cyberbullying

Peer bullying, both in its traditional form and its digital manifestation (cyberbullying), is a persistent problem with serious consequences for student well-being. Several AI tools have been developed to address this challenge:

Bark for Schools (Primarily USA): This platform uses deep learning algorithms to analyze content in school Google Workspace and Microsoft 365 accounts (emails, chats, documents), as well as web Browse on school devices. Its goal is to identify not only keywords, but also context and sentiment to detect potential cases of bullying, violence, suicidal ideation, among other risks. Bark reports having alerted on hundreds of thousands of cases of severe bullying and hate speech. The company argues that its algorithmic approach can reduce individual bias and detect problems that would otherwise go unnoticed. However, it faces significant criticism regarding the generation of false positives, the lack of nuanced understanding of adolescent language and slang, the burden on school administrators to review alerts, and concerns about privacy and the potential chilling effect on student communication. Despite this, it is used by thousands of school districts in the USA.

Lightspeed Alert (USA): Similar to Bark, Lightspeed Alert monitors students' digital activity (Browse, searches, Google Suite documents, social media accessed from school accounts) for risk indicators, including bullying, violence, and self-harm. It uses AI filters and a team of human reviewers ("Safety Specialists") available 24/7 to assess the severity of alerts and escalate the most critical ones to school-designated contacts or even law enforcement if there is an imminent threat. It is reported to have enabled early interventions. However, like Bark, it has raised concerns among students about privacy, freedom of expression, and the possibility of false positives, especially when researching sensitive topics for schoolwork. The platform claims to comply with regulations such as FERPA, CCPA, and GDPR.

NetSupport DNA: Offers a classroom management solution that includes features to monitor students' technology use and detect behaviors associated with bullying. It seeks to support affected students and foster a safer school environment.

STOPit: Focuses on facilitating anonymous reporting of bullying through an application, including a live chat function for students to communicate confidentially with school counselors, thus promoting an environment of trust.

SameBullying (Spain): Developed by ENCAMINA, this tool integrates with school platforms like Microsoft Teams or ClickEDU to monitor conversations (text, image, video) using AI and Azure Cognitive Services. It detects bullying patterns, alerts

parents and guardians, offers a control panel, and sends educational content to victims and bullies. It ensures anonymity in access to conversations, which is exclusive to specialized educators who train the model.

WatsomApp (Spain): Uses a playful approach, employing online games and interactive robots (Snow and QBO from IBM) to collect information about classroom climate and detect possible cases of bullying. It has been tested with reported good results.

Capybara AI (Chile): As mentioned earlier, it adopts a preventive approach by analyzing classroom social and cooperation dynamics to identify bullying risks and foster positive relationships.

10.4.3 Violence Prevention and School Safety Improvement

Concern for physical safety in schools has led to exploring AI for threat detection in the physical environment:

North Carolina Pilot (USA): New Hanover and Davidson counties are implementing a system (from provider Eviden) that integrates AI with existing security cameras. The goal is to automatically detect threats such as firearms, intruders, fallen individuals, or open doors, and then track the threat while alerting authorities. The system is designed for continuous monitoring, overcoming human limitations. Although funded by public funds and a report is expected in 2026, serious privacy concerns persist, especially regarding the (albeit optional) use of facial recognition.

Other tools: Both **Lightspeed Alert** and **Bark for Schools** also include the detection of violence threats and potential school shootings among their digital content monitoring functionalities.

10.4.4 Identification of Risks to Student Well-being (Self-Harm, Suicide, Eating Disorders, Addictions)

Student mental health and well-being are areas where AI is also being applied, with the aim of identifying early warning signs:

Bark for Schools and **Lightspeed Alert (USA):** Both platforms actively monitor digital communications and activities for indicators of suicidal ideation, self-harm, severe depression, and other mental health issues. Bark reports generating a significant number of daily alerts for imminent risk of self-harm/suicide. Lightspeed uses human reviewers to evaluate and escalate these critical alerts.

Research and Development: Studies like the RAND Corporation report analyze the use of these technologies. While recognizing that they can help identify at-risk students, they point to a lack of solid research on their actual accuracy, their

effective impact on improving mental health, and the risks associated with privacy and equity. The tendency to generate false positives and the difficulty of interpreting context are important challenges. Applications like **Vira**, though developed in a clinical context, explore how mobile technologies can be used to monitor risk in real-time and offer "just-in-time" interventions for high-risk youth, highlighting the importance of user-centered design for adoption and effectiveness.

10.4.5 Detection of Content Related to Radicalization and Extremism

The use of online platforms for the dissemination of extremist propaganda and radicalization, especially among young people, is a growing concern. AI is being explored as a tool for both detection and prevention:

INSIKT Project (EU): This EU-funded initiative developed a data mining platform specifically for law enforcement to detect online violent radicalization in real-time. It uses NLP, text mining, social network analysis, and machine learning/deep learning to identify radical content, vulnerable individuals, and covert radicalization processes. The goal is to enable early intervention and limit the spread of extremist content.

Bark for Schools (USA): Reports having detected instances of conversations related to radicalization by hate groups.

General Context: It is recognized that AI can analyze large volumes of data to identify patterns associated with radicalization. However, extremist groups are also learning to use AI (especially generative AI) to create more sophisticated and personalized propaganda. Therefore, a purely content-detection and removal approach is considered insufficient and inefficient. A preventive approach is needed that combines technology with the promotion of media and information literacy (MIL) to build resilience in young people. Human oversight remains essential due to AI's limitations in understanding cultural and linguistic context, and the need for human rights-based regulatory frameworks is urgent.

10.4.6 Cybersecurity and Protection against Digital Threats

Educational institutions are attractive targets for cyberattacks due to the large amount of sensitive data they handle. AI is being used to strengthen defenses:

Schools and Libraries Cybersecurity Pilot Program (USA): This FCC program provides funds for schools and libraries to acquire advanced cybersecurity services and equipment, which may include next-generation firewalls, endpoint protection, and network detection and response systems, many of which incorporate AI capabilities to identify and mitigate threats in real-time.

CISA Pilot (USA): The US Cybersecurity and Infrastructure Security Agency conducted a pilot to evaluate the effectiveness of AI-based vulnerability detection tools (including LLMs). The conclusion was that, currently, AI is more useful for complementing and improving existing tools than for replacing them, although the field is constantly evolving.

10.4.7 Plagiarism Detection and Promotion of Academic Integrity

The ease with which generative AI can produce text has intensified concerns about academic integrity. AI tools are being used and developed to detect both traditional plagiarism and AI-generated content:

Grading and Detection Tools: Platforms like **Gradescope** use AI to streamline grading and potentially identify similarities. **Copyscape** is an AI-powered plagiarism checker. Specific protocols and tools exist that use algorithms to compare texts and detect similarities or characteristic patterns of AI generation.

Challenges of AI Detection: There is considerable controversy regarding the reliability of AI-generated text detectors. Studies have shown that these tools can fail, have low accuracy rates, and, worryingly, exhibit bias against non-native English writers, incorrectly classifying their work as AI-generated. This raises serious ethical and equity concerns. Therefore, pedagogical approaches that prioritize process over product, transparency in the use of detection tools, and the development of alternative assessment methods that are less susceptible to AI generation are advocated.

11. Psychological and Pedagogical Impact of Sensitive Content Detection on Students and Educators

11.1 Introduction

The implementation of artificial intelligence models for sensitive content detection in school environments represents a complex initiative, driven by the primary goal of creating safer digital spaces for students. However, this technological endeavor brings with it a series of significant psychological and pedagogical impacts for both students and educators, which must be meticulously analyzed.

It is fundamental to contextualize this discussion within the legal frameworks that shape the ethical deployment of AI in education in Europe, primarily the EU Artificial Intelligence Act and the General Data Protection Regulation (GDPR). The EU AI Act is particularly relevant, as it classifies certain AI applications in the educational sphere as "high-risk." This categorization imposes strict requirements

in terms of transparency, human oversight, and risk management, especially when it comes to monitoring student activities. GDPR principles, such as data minimization and purpose limitation, are also crucial and must guide any processing of personal data in the school context. The interaction of these regulations highlights an inherent tension: while innovation and AI adoption are encouraged to improve education, this must occur within robust ethical and legal boundaries that prioritize fundamental rights. The successful integration of AI in European schools will largely depend on institutions' ability to navigate this complex regulatory landscape, and not solely on the technological effectiveness of the tools implemented.

Recent reports, such as the Vodafone Foundation survey, indicate that more than two-thirds of adolescent students in Europe have access, at least to some extent, to AI-equipped devices. This increasing ubiquity of technology makes understanding the impacts of specific applications, such as sensitive content filtering, even more critical. The very definition of "sensitive content" is not static and can be influenced by algorithmic design and the diverse cultural contexts present in Europe. AI systems can be trained to detect different categories such as "prohibited behavior," "bullying, self-harm, or suicide," or "inappropriate content," among others. The EU AI Act considers the "detection of prohibited student behavior" as high-risk. This variability suggests that what AI identifies as "sensitive" can differ considerably, posing a challenge for standardization and fairness across different schools. Without clear and shared definitions, and without transparent algorithms, there is a risk of inconsistent application and potential biases in what is flagged as problematic, affecting students unevenly.

While sensitive content detection using AI aims to protect students, its implementation without careful consideration of the psychological and pedagogical impacts, and without strict adherence to ethical and legal guidelines, can inadvertently harm student well-being, undermine trust in the educational institution, and negatively affect the fundamental educational mission. It is imperative, therefore, to analyze these consequences to ensure that technological solutions are applied in a way that is not only effective but also ethical, and that fosters a positive and trusting school climate, exploring the possible repercussions on students' emotional well-being, their development, the teaching-learning dynamic, and the role of educators in the face of these new tools.

11.2 Psychological Impact on Students

The introduction of content monitoring and detection systems, even with the best intentions, can generate various reactions and psychological effects in students. These effects are not limited to simple discomfort but can have profound ramifications on their emotional well-being, behavior, and holistic development.

11.2.1 Anxiety and Stress from the Perception of Continuous Surveillance

The awareness that their online communications and activities are being constantly supervised by AI systems can generate feelings of anxiety, stress, and a sense of being under continuous surveillance in students. This "panopticon effect," where the individual internalizes the gaze of the observer, could lead students to feel inhibited and worried about being misinterpreted by an algorithm, even when their behavior is innocuous. The constant algorithmic measurement of behavior can create a silent but persistent pressure, contributing to an increase in performance-related anxiety and self-expression. General studies on the impact of AI on student well-being, although not exclusively focused on content filtering, indicate that excessive reliance on technology and prolonged screen time can contribute to digital fatigue, isolation, and anxiety.

The concern about triggering a false alert and the possible consequences – being questioned, parents being contacted, etc. – can be a significant source of stress. This psychological burden can be amplified by the inherent opacity of many AI systems. If students do not understand how they are being monitored, what criteria the algorithm uses to flag content, or how decisions are made based on those alerts, anxiety can intensify due to a perceived lack of control and fairness. The feeling of being observed by an inscrutable and potentially fallible entity is inherently stressful. Furthermore, the "always-on" nature of digital learning environments, coupled with AI monitoring, can blur the lines between school life and personal life, especially if school-provided and monitored devices are used at home. This extension of surveillance into personal spaces could prevent students from relaxing and disconnecting, exacerbating stress and the feeling of not having a private space free from scrutiny. The potential insecurity of these surveillance systems, as evidenced in cases of sensitive student data breaches, adds another layer of anxiety related to privacy and the exposure of intimate information.

11.2.2 "Chilling Effect" on Expression and Exploration

One of the most cited and concerning risks is the "chilling effect" on freedom of expression and the natural exploration of ideas. This phenomenon, rooted in theories like Schauer (1978) and Noelle-Neumann's Spiral of Silence (1974), describes how individuals suppress their self-expression due to fear of legal sanctions or social rejection, even in the absence of direct threats. In the digital environment, this effect is exacerbated by the perception of surveillance, over-regulation, algorithmic biases, and opaque moderation practices.

Students might self-censor, avoiding discussing topics they fear could be flagged as sensitive by AI, even if doing so in an appropriate academic context or as part of a personal search for information or support. This could limit their curiosity, their

ability to ask difficult questions, or seek help on delicate topics, negatively affecting their intellectual and emotional development. For example, a student researching eating disorders for a school project, or topics of mental health, gender identity, or controversial political issues, might fear that their searches or discussions will be misinterpreted by the AI system and generate an alert. The ambiguity about what exactly constitutes "sensitive" or "harmful" content for an algorithm can intensify this effect, leading students to err on the side of caution and, therefore, significantly restrict their intellectual and emotional exploration.

This inhibitory effect is not uniform and can disproportionately affect students who are already cautious, those belonging to minority groups, or those exploring unconventional ideas. Research suggests the effect is particularly pronounced for marginalized groups when discussing contentious topics. If AI surveillance leads these students to self-censor more, there is a risk of fostering less diverse and critical discourse within the school environment, which in turn could undermine fundamental educational goals that seek to promote critical thinking, inquiry, and the ability to engage with complex and sometimes uncomfortable topics. In the long term, a significant chilling effect could result in a student population less prepared to face the complexities of the real world and less willing to participate in robust civic debates. Granular surveillance in schools, therefore, not only affects individual expression but can also have broader consequences for the intellectual and democratic climate of the institution.

11.2.3 Impact on Trust, Relationship with the Institution, and Student Perception

The perception of constant and automated surveillance can erode students' trust in the educational institution. If AI systems are perceived as intrusive, unfair, or error-prone, students might feel less safe and respected, damaging the relationship and sense of community within the school. Lack of transparency about how these systems work, what data is collected, how it is used, and how algorithmic decisions can be appealed, can exacerbate this distrust.

European surveys on student perception of AI in education, such as the Vodafone Foundation report, reveal a trust gap: although students recognize AI's importance, less than half feel adequately prepared by their schools or perceive their teachers as competent in using AI. While these data do not specifically refer to content monitoring, they indicate a breeding ground where the implementation of intrusive surveillance systems could further erode trust. A study by Slimi (2025), although focused on higher education and with a limited sample, found that students valued AI tools like ChatGPT and Grammarly for their benefits in critical thinking and feedback, but this does not necessarily imply acceptance of sensitive content monitoring, which is a qualitatively different and more invasive AI application.

A breakdown in trust between students and the institution, due to the perception of intrusive surveillance, can have cascading negative effects on the school climate, student participation, and willingness to seek help from school staff for genuine problems. If students perceive the school primarily as a policing entity rather than a supportive environment, they may be less likely to report bullying, seek mental health support, or interact openly with educators. Ironically, this could undermine some of the safety goals that AI monitoring aims to achieve. The way these monitoring systems are introduced is as crucial as the technology itself; a lack of consultation with students and parents can foster distrust from the outset.

11.2.4 False Positives and Their Emotional and Academic Consequences

AI systems are not infallible and can generate false positives, flagging innocuous content as sensitive or problematic. Being unfairly accused due to an algorithmic error can be a deeply negative experience for a student, generating feelings of frustration, injustice, shame, and even stigmatization. The process of clarifying a false positive can itself be stressful and burdensome for the student.

The literature on AI detection tools, such as those used to identify plagiarism or AI-generated text, abounds with examples of their unreliability and high false positive rates. For example, Turnitin's AI detector has been reported to incorrectly flag student work, leading to academic dishonesty investigations based solely on AI scores. Although these tools are not identical to sensitive content detection tools, the underlying AI problems in understanding nuance, context, and human intent are comparable.

The emotional consequences of a false positive can be severe. Students may experience anxiety, damage to their academic reputation, and a loss of trust in the institution's evaluation processes. The burden of proof often unfairly falls on the student, who must prove their innocence against an opaque algorithm, a process that can be intimidating and perceived as unfair. Some institutions, recognizing these problems, have deactivated certain AI detectors, but others continue to use them, creating a worrying inconsistency. Even a seemingly small error rate can translate into a large number of false accusations when applied to an entire student population.

The use of AI detection tools with known high false positive rates, without robust due process and critical human oversight, fundamentally undermines the principles of justice and fairness within the educational system. The psychological harm from a false accusation can go beyond immediate stress, potentially affecting the student's academic trajectory, their self-perception as a learner, and

their future engagement with education. Moreover, there is a dangerous feedback loop: the fear of false positives can generate even more anxiety and self-censorship (chilling effect), exacerbating psychological harms.

11.2.5 Risk of Stigmatization, Labeling, and Algorithmic Bias

Even when a sensitive content detection is deemed "correct" by the algorithm, the handling of this information is crucial. If not addressed with due sensitivity, confidentiality, and understanding of context, the student involved could feel labeled or stigmatized, affecting their self-esteem and their relationships with peers and educators. This risk is magnified if AI systems are trained with biased data, which can lead to discriminatory outcomes.

Sensitive content detection is particularly vulnerable to algorithmic biases, where cultural nuances or expressions from marginalized groups may be misinterpreted. There is a significant risk for vulnerable student groups. For example, LGBTQ+ students might be inadvertently identified or flagged for discussing topics related to their identity. Students with disabilities or who are neurodivergent may have communication styles that AI misinterprets as problematic or indicative of sensitive content. Similarly, the writing styles of non-native English speakers have sometimes been erroneously flagged by AI-based content or plagiarism detectors. The EU AI Act recognizes the need to protect vulnerable groups, including children, from manipulative AI or systems that exploit their vulnerabilities. Indeed, emotion recognition systems, which are prone to bias and could be used to infer emotional states from sensitive content, are prohibited in education and workplaces under this Act.

Algorithmic bias in sensitive content detection is not merely a technical flaw, but a matter of social justice within education. It could lead to systemic discrimination against already marginalized student populations, undermining efforts towards equity and inclusion that are fundamental in European educational systems. The "black box" nature of many AI systems makes it extremely difficult to identify, challenge, and rectify these biases, which can perpetuate a cycle of harm. AI explainability (XAI) therefore becomes crucial not only for trust but also for fairness. Even if an alert is later clarified, the very process of being flagged by an AI for "sensitive content" can be stigmatizing, especially if not handled with extreme sensitivity and confidentiality, potentially having social repercussions or generating internal shame.

11.2.6 Impact on Identity Development and Student Digital Autonomy

Widespread AI surveillance in school environments can profoundly influence how students construct their digital identities and perceive their online autonomy. Adolescence is a critical period for identity formation, and online spaces have

become key arenas for this exploration. The "Visions of artificial intelligence, biometrics and digital surveillance in schools" project, although its final results are pending, specifically investigates changes in the conceptualization of personal identity due to technology, underscoring the relevance of this concern.

Constant monitoring could lead students to adopt a more performative and less authentic online self-presentation, as they meticulously curate their digital footprint to avoid algorithmic scrutiny. This could hinder the development of a secure and integrated personal identity. The way students perceive themselves in the digital environment, their "digital self-perception," can be affected if their online interactions are constantly judged by AI, potentially eroding their confidence to express themselves or their sense of agency in digital spaces.

The European Union's emphasis on human agency within ethical AI guidelines is pertinent in this context. Excessive surveillance could diminish this agency, making students feel like passive subjects of technology rather than active and responsible digital citizens. This could also affect their understanding and affirmation of their digital rights, such as privacy and freedom of expression. If students learn, from their own experience or from the fear of AI detection, that certain explorations or expressions carry negative consequences, they may internalize a caution that restricts their identity development. This relates to the "chilling effect," but focuses on its consequences for development rather than purely expressive ones. The broader implication is a generation potentially less willing to take intellectual risks or express dissenting opinions. Lack of control over how their data is interpreted by AI systems can undermine students' sense of digital ownership and autonomy, crucial components of digital citizenship that schools aim to foster.

11.3 Pedagogical Impact and Impact on Educators

The introduction of AI technologies for sensitive content detection not only affects students but also has profound implications for pedagogical practice and the role of educators. These impacts range from workload management to the need for new competencies and confronting complex ethical dilemmas.

11.3.1 Additional Workload and "Digital Overload" for Educators

While AI can help identify risks, reviewing generated alerts, investigating context, and determining appropriate responses largely fall on school staff. This can lead to a significant additional workload, diverting time and resources from other essential pedagogical tasks. Educators may feel overwhelmed if the volume of alerts, especially false positives, is high. Although some general studies on AI in education suggest that 44% of teachers have used AI, but their workload remains virtually unchanged (only 3% report a large reduction), this data refers to the

general use of AI and not specifically to managing alerts from content detection systems. The burden of reviewing AI-generated alerts, especially if false positive rates are high, can be considerable and contradict the promise that AI will reduce teacher workload.

Many teachers already report increasing "digital overload" that increases stress and burnout, especially when AI implementation is rapid and lacks adequate support. Time spent investigating AI alerts, which often turn out to be false positives, is time taken away from lesson planning, direct student interaction, and other fundamental pedagogical tasks. Furthermore, the phenomenon of "alert fatigue" can arise, where a high volume of notifications (especially if they are false positives or non-critical) can lead to desensitization or overwhelm, with the risk of educators overlooking genuinely critical incidents. Some AI solutions, such as Lightspeed Alert, incorporate a human review team to evaluate context, implicitly recognizing the magnitude and complexity of this task. This burden of alert management could disproportionately fall on specific personnel (such as counsellors, IT staff, or safeguarding officers), creating bottlenecks and specialized stress within the school or organization.

11.3.2 Need for Continuous Training and Professional Development in AI and Ethics Competencies

Educators need specific training not only on how these tools work but also on how to interpret their results, how to approach difficult conversations with students about sensitive content, and how to integrate these technologies within an ethical and supportive framework. The EU AI Act requires personnel using AI systems to possess a sufficient level of AI literacy, a provision that came into effect in February 2025. This underscores the critical need for professional development.

The European Commission's Ethical Guidelines on the use of AI and data in teaching and learning for educators and UNESCO's AI competency frameworks for teachers emphasize understanding AI, its ethical use, and its pedagogical integration. Training must cover AI principles, data privacy (GDPR), bias detection and mitigation, critical interpretation of AI results, managing false positives, ethical decision-making, and how to discuss AI surveillance and online safety with students. There is notable concern about the lack of support and professional development opportunities for teachers regarding AI, indicating a significant gap between AI tool deployment and educator preparedness. Without bridging this gap, AI tools risk being misused or underutilized, and ethical breaches are more likely. Effective AI training for educators must transcend technical skills to encompass deep ethical reasoning, critical digital pedagogy, and strategies to foster student well-being in AI-mediated environments.

11.3.3 Displacement of the Educator's Role and Pedagogical Agency

There is a risk that excessive reliance on technology could lead to a displacement of the educator's role, shifting from a guide and mentor to an alert supervisor. It is crucial that technology serves as a support for human work and does not replace it, especially in building trusting relationships and in the nuanced understanding of individual student needs. AI must complement, not supplant, the essential role of teachers in developing critical thinking and empathy.

Excessive reliance on AI to monitor and flag student behavior could erode educators' pedagogical agency. If decisions about teaching and student well-being are delegated to automated platforms without critical human judgment, there is a risk of de-professionalization. The apparent "objectivity" of AI-generated alerts could subtly undermine educators' professional judgment and their confidence in their own observations and intuitions about students. If an AI flags something, an educator might feel compelled to act even if their own judgment suggests otherwise, or they might begin to doubt their ability to identify problems without AI. This could lead to a gradual erosion of pedagogical autonomy. Furthermore, a shift towards AI-driven "problem detection" could foster a more deficit-centered approach to students, rather than one based on their strengths or development.

11.3.4 Ethical Dilemmas and Decision-Making in Teaching Practice

Educators may face complex ethical dilemmas when interpreting AI alerts. Deciding when to intervene, how to do so, and how to balance safety with student privacy and autonomy requires sensitive and well-informed professional judgment. The lack of explainability of some AI models can hinder this decision-making. These dilemmas can be analyzed through the prism of the European Commission's Ethical Guidelines: human agency versus safety, fairness in intervention, humanity in approach, and justified choice when acting on an AI alert.

The difficulty of interpreting AI-flagged content without full context, which AI often lacks, is a central challenge. A student's search query or an online comment may be flagged as "sensitive" without the AI understanding the academic context, satirical intent, or personal exploration behind it. Although the EU AI Act requires human oversight for high-risk systems, this oversight is fraught with ethical challenges if AI logic is opaque. Educators find themselves in the difficult position of mediating between an algorithmic judgment and a human student, often without sufficient information or training to do so ethically and effectively. The pressure to act on AI alerts to ensure student safety could lead to interventions that, though well-intentioned, might be disproportionate or premature if the AI's assessment is inaccurate or lacks nuance.

11.3.5 Impact on Classroom Dynamics and Teacher-Student Relationship

If students perceive that educators primarily rely on AI to monitor their behavior, this could affect classroom dynamics and the trust relationship between teachers and students. It is fundamental that students continue to see educators as accessible and understanding supportive figures. A perception of adversarial surveillance rather than supportive guidance can damage the trust essential for open communication and a positive learning environment. Students might become more reserved in their interactions with teachers, both online and offline, for fear that any disclosure or expression might be captured and algorithmically judged.

The ideal is for students to see educators as trusted adults they can turn to for help. If educators are primarily seen as enforcers of AI-detected violations, this crucial supportive role is undermined. The introduction of AI-based content detection can inadvertently transform the teacher-student relationship, moving from one based on pedagogical connection and pastoral care to one mediated by surveillance and suspicion. Students may be less likely to confide in teachers about sensitive personal issues (precisely those that AI might be trying to detect signals for) if they fear this information will be formally recorded or trigger an algorithmic response, thus frustrating a key purpose of safeguarding.

11.4 Fostering a Balanced, Ethical, and Supportive Approach: Strategies and Recommendations

To mitigate negative psychological and pedagogical impacts and responsibly leverage any potential benefits, it is essential to adopt a multifaceted approach that integrates legal, ethical, pedagogical, and community engagement considerations. A purely technological approach to student safety is insufficient and potentially harmful. Effective strategies must take into account both people and technology, considering all social and technical aspects holistically.

11.4.1 Strict Adherence to the European Regulatory Framework: The EU AI Act and GDPR

Compliance with the EU AI Act and GDPR is not an option but a legal and ethical obligation. The AI Act classifies AI systems used to monitor students or determine access to education as "high-risk." This classification entails strict requirements, including comprehensive risk assessments, the use of high-quality datasets to minimize bias, activity logging, detailed documentation, clear user information, adequate human oversight, and high levels of accuracy and cybersecurity.

Prohibited AI practices in the educational sphere, such as emotion recognition systems for student monitoring or systems that manipulate student behavior, must be completely avoided. Likewise, GDPR obligations are paramount: there must be a legal basis for data processing, data minimization must be applied (collecting only necessary data), purpose limitation (using data only for specified security purposes), ensuring data subject rights (access, rectification, erasure), and applying data protection by design and by default, especially when dealing with sensitive student data.

Key Provisions of the EU AI Act Relevant to Student Monitoring

Category of Provision	**Specific Requirement (Examples)**	**Implication for Schools**
Risk Classification	AI systems for monitoring students or evaluating learning outcomes classified as "high-risk."	Obligation to comply with all requirements for high-risk systems before implementation and during use.
Data Governance and Quality	Use of high-quality, relevant, representative, and error- and bias-free training datasets.	Need to investigate the origin and composition of training data for acquired AI systems to prevent the perpetuation of biases.
Transparency and Information Provision	Clear and adequate information to users (educators, students) about system operation and risks.	Development of understandable informational materials for the entire educational community on AI systems in use.
Human Oversight	High-risk systems designed to allow effective human oversight; final decisions must not be fully automated.	Establishment of clear protocols for human intervention in alert review and decision-making; staff training for this task.

Accuracy, Robustness, and Cybersecurity	Appropriate levels of accuracy, robustness against errors or inconsistencies, and cybersecurity.	Rigorous evaluation of the technical reliability of systems before adoption; implementation of security measures to protect the system and data.
Prohibited AI Practices	Prohibition of AI systems that exploit vulnerabilities of specific groups (e.g., children) or use subliminal techniques; prohibition of emotion recognition in education.	Refrain from using any AI system that falls into these prohibited categories, regardless of perceived benefits.
Logging and Documentation	Maintenance of technical documentation and operational logs.	Implementation of systems to maintain required documentation and ensure traceability of AI system operations.

11.4.2 Radical Transparency and Proactive Communication with the Educational Community

It is fundamental to adopt a policy of total transparency with students, parents, and staff about which AI tools are used, their purpose, how they work (explained in language appropriate for each age and audience), what data they collect, how that data is used and stored, and for how long. Explainability (XAI) is a key component of this transparency; whenever possible, systems should provide reasons for their results or alerts, allowing students and educators to understand why certain content was flagged. Students' rights regarding their data and AI systems should be clearly communicated, including available appeal mechanisms in case of disagreement with an algorithmic decision.

11.4.3 Student Participation and Co-design of Digital Safety Policies

Involving students in discussions about online safety, AI ethics, and the development of acceptable use policies is crucial. This participation not only fosters acceptance and trust but can also lead to more effective and student-

centered solutions. Creating student councils or focus groups to gather perspectives on AI monitoring tools before and during their implementation could be considered, ensuring their voices are heard and taken into account in the decision-making process.

11.4.4 Qualified and Indispensable Human Oversight at All Stages

It must be constantly reiterated that AI is a support tool and not a substitute for human judgment. Final decisions regarding student discipline, welfare interventions, or other significant actions must be made by trained human personnel, who consider AI results as one piece of information among others. Staff responsible for reviewing alerts and intervening need specialized training not only in interpreting AI data but also in understanding context, detecting potential biases, and conducting sensitive conversations with students and their families.

11.4.5 Prioritization of Preventive Education: Digital Citizenship and Critical Media Literacy

Technological detection must be complemented by robust educational programs that empower students to navigate the digital world safely and responsibly. This includes developing digital citizenship skills, such as responsible online behavior, understanding digital rights and responsibilities, online safety practices (privacy, security), and ethical conduct. Critical media literacy is equally essential, teaching students to critically evaluate online information (including AI-generated content), identify disinformation, understand algorithmic influence, and recognize manipulative content. AI literacy, which involves educating students on how AI works, its capabilities, limitations, and ethical implications, is promoted by OECD and European Commission initiatives. Finally, fostering social-emotional learning is vital to building resilience and helping students positively manage online risks and interactions.

11.4.6 Establishment of Clear, Fair, and Robust Appeal Processes

It is imperative to develop clear, transparent, and consistently applied protocols for managing AI-generated alerts. These protocols must include steps for human verification of alerts, contextual investigation, and sensitive communication with students and parents. Crucially, fair and accessible appeal processes must be established for students who believe they have been incorrectly flagged or sanctioned due to an AI error. This is especially important given the high false positive rates of some tools and is a requirement for high-risk AI systems under the EU AI Act. Due process must be guaranteed, and AI detection should not be treated as infallible proof.

11.4.7 Privacy Protection and Data Governance by Design and by Default

Strict adherence to GDPR is required, implying data minimization (collecting only necessary data), purpose limitation (using data only for specified security purposes), secure storage, defining retention periods, and clear consent protocols where applicable. The principles of privacy by design and by default must be implemented in the selection and configuration of AI tools. Before deploying AI monitoring systems, it is advisable to conduct Data Protection Impact Assessments. Transparency about data handling practices with students and parents is fundamental. Following recommendations like those from the UK DfE, personal data should be avoided in generative AI tools, and student work should not be used to train AI models, a widely applicable data protection principle.

11.4.8 Continuous Evaluation, Research, and Adaptation of Strategies

Educational institutions must regularly monitor and evaluate the impact of AI detection systems on student well-being, school climate, teacher workload, and the effectiveness of the systems themselves. Gathering feedback from students, educators, and parents is important. Staying updated on evolving AI technologies, ethical best practices, and regulatory changes is essential. Schools must be prepared to adapt or even discontinue the use of AI tools that prove to be harmful, ineffective, or ethically problematic. Supporting and participating in AI research in education can inform and improve practices.

11.4.9 Fostering Explainability (XAI) and Auditability of AI Systems

Priority should be given to AI systems that offer transparency in their decision-making processes (Explainable Artificial Intelligence - XAI). This helps build trust, allows for better human oversight, and facilitates the identification and mitigation of biases. It is crucial to ensure that systems are auditable, allowing for the review of how decisions were made, especially in cases of error or dispute. XAI is fundamental for accountability and fairness when AI is used for evaluation or monitoring.

The European approach, characterized by strong regulation and an emphasis on ethical guidelines, provides a unique framework for navigating AI in education. This framework could set a global standard for responsible implementation, going beyond purely technological or market-driven approaches. The successful mitigation of the negative impacts of AI content detection depends on proactive measures, including initial investment in teacher training, robust policy development, and fostering a culture of digital citizenship before widespread AI deployment.

Summary of Psychological Impacts on Students and Recommended Mitigation Strategies

Psychological Impact	**Key Contributing Factors (AI System/Implementation)**	**Recommended Mitigation Strategies (Reference to Subsection)**
Anxiety and Stress from Surveillance	Perception of constant monitoring, system opacity, fear of errors and consequences.	Radical transparency (11.4.2), Explainability (XAI) (11.4.9), AI education (11.4.5), Clear protocols (11.4.6).
Chilling Effect (Self-Censorship)	Algorithmic rule ambiguity, fear of misinterpretation, perception of surveillance.	Transparency (11.4.2), Digital rights education (11.4.5), Student participation (11.4.3), Human oversight (11.4.4).
Reduced Trust in Institution	Perception of intrusiveness, unfairness, lack of transparency, negative experiences (e.g., false positives).	Transparency (11.4.2), Student participation (11.4.3), Fair protocols and appeals (11.4.6), Proactive communication.
Emotional/Academic Harm from False Positives	AI fallibility, lack of due process, burden of proof on student.	Indispensable human oversight (11.4.4), Clear protocols with robust appeal processes (11.4.6), Continuous tool evaluation (11.4.8), Prioritize XAI (11.4.9).
Stigmatization, Labeling, and Algorithmic Bias	Biased training data, AI's lack of contextual understanding, insensitive alert handling.	Adherence to AI Act (bias prevention) (11.4.1), Data governance (11.4.7), Teacher training on biases (11.3.2), XAI (11.4.9),

		Sensitive alert handling protocols (11.4.6).
Negative Impact on Identity/Autonomy Development	Perceived surveillance as limiting exploration, lack of control over digital narrative, performativity.	Foster student agency (11.4.5), Digital citizenship and rights education (11.4.5), Limit monitoring to strictly necessary (11.4.7), Student participation in policies (11.4.3).

Pedagogical Challenges for Educators and Recommended Support Mechanisms

Pedagogical Challenge	**Key Contributing Factors (AI System/Implementation)**	**Recommended Support/Training (Reference to Subsection)**
Additional Workload / Digital Overload	High volume of alerts (especially false positives), need for investigation and follow-up.	Efficient and well-calibrated AI tools, Optimized human oversight (11.4.4), Training in alert management (11.3.2), Continuous workload evaluation (11.4.8).
Skills Gap / AI and Ethics Literacy	Rapid AI evolution, lack of adequate and continuous training.	Robust and continuous professional development in AI and ethics (11.3.2), Adherence to ethical guidelines (11.4.1), Resources and communities of practice.

Displacement of Educator's Role / Agency	Excessive reliance on AI for monitoring and decision-making, educator as "alert supervisor."	Emphasis on AI as support, not replacement (11.4.4), Training in critical AI pedagogy (11.3.2), Foster professional teacher judgment.
Ethical Dilemmas and Complex Decision-Making	AI opacity, difficulty balancing safety/privacy/autonomy, pressure to act on alerts.	Comprehensive ethical training (11.3.2), Prioritize XAI (11.4.9), Clear decision-making protocols (11.4.6), Collegial support for difficult cases.
Affected Classroom Dynamics and Teacher-Student Relationship	Student perception of surveillance and distrust, reduced open communication.	Transparency with students (11.4.2), Foster AI as a supportive, not punitive, tool (11.4.4), Strengthen digital citizenship education and trust relationships (11.4.5).

11.5 Towards a Responsible and Human-Centered Implementation

The implementation of artificial intelligence systems for sensitive content detection in school environments is a complex task that transcends mere technological selection. The psychological impacts on students – from anxiety and stress derived from continuous surveillance, through the inhibitory effect on their expression and exploration, to emotional harm caused by false positives and the risk of stigmatization – are primary considerations. Similarly, the pedagogical implications for educators – including potential workload overload, the imperative need for continuous training in digital and ethical competencies, the ethical dilemmas inherent in interpreting algorithmic alerts, and the possible displacement of their fundamental role – must be proactively considered and managed.

The deployment of these technologies in the European context is framed by a robust set of values and a regulatory framework that prioritizes human dignity,

fundamental rights, equity, and transparency, as enshrined in regulations such as the EU AI Act and GDPR. These principles must constitute the unwavering compass guiding any implementation. The solution to online safety challenges cannot create greater problems in terms of psychological well-being or pedagogical integrity. Therefore, AI's success in this context should not be measured solely by its detection accuracy, but by its overall impact on the school ecosystem, including student well-being, trust, learning quality, and the burden on educators.

An approach that prioritizes radical transparency, qualified and indispensable human oversight at all stages of the process, robust preventive education in digital citizenship and media literacy, and the holistic well-being of the entire educational community is fundamental. This means investing significantly in educator training, ensuring the explainability of AI systems as much as possible, promoting active student participation in shaping digital safety policies, and establishing clear, fair, and effective appeal processes.

The European Union, with its regulatory framework and emphasis on ethical guidelines, is uniquely positioned to lead a global conversation on the ethical implementation of AI in education, transcending purely technological or market-driven approaches. The ultimate goal is not to achieve perfect algorithmic detection, but to create genuinely safe, supportive, and stimulating learning environments where technology serves human ends, empowers students and educators, and reflects the highest ethical standards of European society. The path forward demands continuous critical reflection, rigorous research, and an unavoidable commitment to adapt strategies as technology and our understanding of its impacts evolve. How European educational institutions navigate the inherent tension between safety and freedom will largely define the future of coexistence and learning in the digital age.

12. Practical Implementation Strategies and Change Management

12.1 Introduction

The decision to implement Artificial Intelligence (AI) models for detecting sensitive content in educational institutions goes beyond merely selecting a technological tool. It requires careful strategic planning, a deep understanding of each institution's specific needs and context, and a proactive approach to change management that involves the entire educational community.

The current landscape shows a growing penetration of AI in the education sector. More than 47% of educational leaders already use AI daily, and it's estimated that 60% of teachers actively employ AI-based tools in their daily work. The global AI in education market, valued at USD 7.57 billion in 2025, is projected to reach USD 30.28 billion by 2029, with a compound annual growth rate (CAGR) for generative AI of 41.4%. This rapid adoption occurs at a time of increasing concern for students' online safety and their exposure to various sensitive content. These include cyberbullying, which, according to a 2024 WHO/Europe study, affects one in six school-aged children, child sexual abuse material (CSAM), whose production and dissemination are even facilitated by AI itself, online grooming, misinformation, and radicalization. The accelerated digitalization of young people's lives, especially intensified after the COVID-19 pandemic, has increased the window of exposure to these risks.

Students are increasingly frequent users of AI tools, both for academic purposes and entertainment, often without adequate supervision. The "AI in European Schools" report from January 2025, based on surveys of 7,000 students aged 12-17 in seven European countries, reveals that 48% use ChatGPT on their own initiative and 47% do so under teacher instruction. This ubiquity of AI in student life underscores the urgency of proactive approaches. In response, the European Union is developing robust regulatory frameworks, notably the EU Artificial Intelligence Act (AI Act) and the continuous strengthening of the General Data Protection Regulation (GDPR). It's crucial to note that the AI Act classifies certain AI systems used in education, such as those for evaluating learning outcomes or detecting prohibited behavior, as "high-risk," imposing strict requirements.

AI presents a dual landscape of challenges and opportunities for safety in the school environment. On one hand, it offers the opportunity to significantly improve student safety and well-being, allow for deeper learning personalization, and free up teacher time for higher-value pedagogical tasks. On the other hand, its implementation carries inherent risks related to privacy and security of minors' data, the perpetuation of algorithmic biases that can lead to discrimination, a lack of transparency in AI model decision-making, potential over-reliance that might diminish students' critical thinking, and the need for greater digital and AI literacy for both staff and students. Furthermore, the costs associated with the full lifecycle of these technologies and the sustainability of their funding are unavoidable considerations. The existing digital divide, in terms of both infrastructure access and competencies, could be exacerbated if AI implementation is not approached with an equity perspective.

The "International AI Safety Report 2025" highlights the rapid evolution of general-purpose AI capabilities and the critical need for a shared, updated scientific

understanding of its inherent risks. This context demands that the implementation of AI for sensitive content detection is not merely an isolated technological decision, but a profound transformation requiring committed leadership, meticulous planning, and a people-centric approach involving the entire educational community. The ethical principles of fairness, transparency, explainability, accountability, and, fundamentally, human oversight, must be the pillars guiding the entire process. Organizations like UNESCO and the Council of Europe offer valuable frameworks and guidelines that can guide educational institutions on this complex path.

The growing adoption of AI in education, coupled with persistent digital divides in infrastructure and competencies, poses a significant risk of exacerbating existing inequalities. Educational institutions with greater resources and better connectivity could leverage AI more effectively, potentially widening the achievement gap for students from less resourced institutions or regions. This is a critical point, as the promise of AI includes improving educational equity. To avoid this adverse outcome, it is essential that implementation strategies actively consider reducing the digital divide as a central objective.

Moreover, AI presents an inherent duality: it is a tool to enhance safety through content detection, but it can also be a vector for new threats, such as AI-generated CSAM (Child Sexual Abuse Material) or deepfakes. As students are increasingly frequent users of AI tools, they are in a vulnerable position, being both potential beneficiaries of AI safety tools and potential targets or even unwitting creators/distributors of AI-generated harm. This creates a complex scenario that demands not only reactive measures, but also proactive strategies focused on AI literacy and the development of critical thinking in both students and teaching staff. A purely technological solution is, therefore, insufficient.

Consequently, the successful and ethical implementation of AI for sensitive content detection is not merely a technical challenge, but a complex sociotechnical challenge, intrinsically linked to educational policy, digital equity, and the evolving nature of online risks. A fragmented approach is unlikely to succeed; a holistic, strategic, and adaptable framework is essential.

12.2 Phase 1: Strategic Planning and Specific Needs Assessment

Before embarking on the implementation of any AI technology, it is imperative for educational institutions to undertake a thorough preparation phase. This initial stage lays the groundwork for an adoption that is not only technically viable, but also aligned with pedagogical values, ethical and legal requirements, and the particularities of each educational community. It involves deep introspection and a rigorous analysis of the institutional context and capabilities.

12.2.1 Clear Definition of Objectives, Scope, and Types of Sensitive Content to Detect

The first fundamental step is for educational institutions to define with the utmost precision the specific problems they seek to solve through AI. This goes beyond a general aspiration to "improve security"; it requires a granular identification of the concrete types of sensitive content whose detection will be prioritized. Recent reports from organizations such as Europol and the World Health Organization (WHO) offer an updated overview of the most prevalent online threats to youth. Europol's "Intelligence Notification: Violent online communities threaten children" and the Internet Organised Crime Threat Assessment (IOCTA) 2024 are crucial sources for understanding the nature of these risks, which range from cyberbullying and hate speech to child sexual abuse material (CSAM) – including AI-generated –, promotion of self-harm, dissemination of extremist ideologies, and disinformation. The European Schoolnet survey on sensitive topics in the classroom can also guide this definition, noting that prejudice based on sexual orientation and gender identity, religious issues, and anti-democratic attitudes are perceived by teachers as particularly sensitive to address.

In parallel, the desired objectives must be Specific, Measurable, Achievable, Relevant, and Time-bound (SMART). Is the goal a quantifiable reduction in cyberbullying incidents? To improve response times to student well-being alerts? To strengthen compliance with online child protection regulations? These objectives will guide the tool selection and impact evaluation.

Finally, the scope of implementation must be clearly defined. This involves answering questions such as:

- Which student groups will be monitored (considering age, educational level, and specific vulnerabilities)?
- Which digital platforms will be subject to supervision? This includes Learning Management Systems (LMS), institutional email, school-approved communication platforms, and, crucially, how students' personal devices (BYOD) and school Wi-Fi network use will be addressed.
- In what languages will the detection system operate, considering the linguistic diversity of the student body?

A precise definition of these three elements – content types, objectives, and scope – is the cornerstone for a focused and effective implementation.

12.2.2 Analysis of the Specific School Context and Institutional Preparation

Each educational institution is a unique ecosystem with its own technological infrastructure, organizational culture, human and economic resources, and a student population with particular characteristics. An exhaustive diagnosis of this reality is vital before considering the adoption of AI systems.

Existing technological infrastructure and connectivity: It is essential to evaluate the current capacity of the internal network, the speed and reliability of Internet access, and the availability, age, and type of devices used by students and staff (both school-provided and personal under BYOD policies). Learning Management Systems (LMS) such as Moodle, itslearning, Google Classroom, or Microsoft 365 Education, and Student Information Systems (SIS) in use, and their potential for integration with the new AI tool, should also be analyzed. Reports such as those from Eurydice on digital education in Europe and the "State of Digital Communications report" from Connect Europe can offer a general overview, but the specific analysis of each center is irreplaceable. The Skills Upload Jr. report from January 2025 reveals concerning data: 51% of surveyed European students experience insufficient internet connectivity in their schools and 59% report limited access to digital devices at school. This data underscores the heterogeneity of digital infrastructure in Europe and the critical importance of this contextual analysis.

Digital competencies and AI literacy of staff and students: An honest assessment should be made of the current level of preparedness of teachers, administrative staff, and students to use and understand AI technologies. This includes not only basic technical skills, but also the ability to critically interact with AI, understand its limitations, and its ethical implications. The EU AI Act, in force since August 2024 and with progressive application, establishes in its Article 4 an obligation of "AI literacy" for AI providers, requiring that personnel operating and using these systems receive adequate training, considering their technical knowledge, experience, and the context of use. In fact, AI literacy obligations and the prohibition of certain AI systems came into force on February 2, 2025. Initiatives such as the AI literacy framework jointly developed by the European Commission and the OECD for primary and secondary education, whose draft was launched in May 2025 and will be finalized in early 2026, will be valuable resources. The Skills Upload Jr. report indicates that only 46% of European students feel adequately prepared by their schools for AI, and only 44% perceive their teachers as competent in using this technology, highlighting a significant training gap.

Available resources: A detailed budget analysis that considers the Total Cost of Ownership (TCO) is crucial. This covers not only the initial acquisition of the AI

solution (licenses, hardware, software) but also implementation costs, customization, integration with existing systems, ongoing staff training, annual maintenance, updates, and technical support. Tools like CoSN's TCO Assessment Tool, though American in origin, offer applicable methodologies. The availability of technical and pedagogical staff to support implementation and continued use, as well as their professional development needs, should be evaluated. AI funding models in the EU and innovative national programs like Estonia's "AI Leap" (which contemplates a public-private foundation for its funding) can offer ideas for financial planning. AI solution costs can vary drastically, from tens of thousands of euros for basic solutions to hundreds of thousands for customized and complex systems.

Organizational culture and readiness for change: Evaluate the general willingness of the educational community (management, teachers, non-teaching staff, students, and families) to adopt new technologies, especially those involving monitoring and handling sensitive data. Identifying potential sources of resistance and planning strategies to address them is part of this preparation.

The lack of a clear AI strategy in many educational institutions (15% of surveyed institutions lack one, and an additional 42% have strategies not aligned with the institution's general objectives) and concerns about teacher preparedness (almost 49% of respondents in a CoSN survey fear that teachers are not adequately prepared to implement AI) are significant barriers that must be identified and addressed in this phase. An honest assessment of these internal capabilities is fundamental, as it directly influences the institution's ability to define realistic objectives and effectively participate in the selection and design of the implementation. Addressing these shortcomings in Phase 1, even through a pre-assessment of staff AI literacy and the formulation of a preliminary institutional strategy, is crucial before proceeding to technology selection.

12.2.3 Exhaustive Risk Assessment and Regulatory Compliance (DPIA/EIPD)

Given the nature of personal data processing, often sensitive and belonging to minors, involved in online content detection, performing a Data Protection Impact Assessment (DPIA), internationally known as DPIA, is a mandatory and unavoidable requirement before implementing any AI system for these purposes. This type of processing is considered "high risk" according to Article 35 of the GDPR and aligns with the requirements of the EU AI Act for AI systems in education.

The DPIA must be a rigorous and documented process that identifies and analyzes potential risks to the fundamental rights and freedoms of students. These risks are

not limited to privacy, but encompass non-discrimination, freedom of expression, equity, and the potential for stigmatization. For each identified risk, concrete technical and organizational measures must be established to mitigate or eliminate it. The European Data Protection Board (EDPB) and the European Union Agency for Cybersecurity (ENISA) offer general guidelines and risk management frameworks that, although not always specific to AI in education, provide a solid methodological basis.

Consultation with the educational institution's Data Protection Officer (DPO) or the competent educational authority throughout the DPIA process is essential. Furthermore, the guidelines and criteria issued by national Data Protection Authorities (DPAs) are binding and may offer specific interpretations. For example, the CNIL (France) has published guides on AI and data protection, emphasizing that DPIAs for AI systems must address specific risks such as the generation of fake content about real people (deepfakes), data poisoning, or model inference attacks. Other DPAs such as the AEPD (Spain), the Garante per la protezione dei dati personali (Italy), or the UODO (Poland) also publish relevant guidance. Experiences like those in the Netherlands and Norway, where centralized DPIAs have been carried out for commonly used educational software (e.g., Google products), can serve as a model for optimizing resources and ensuring consistent evaluation, especially for schools with less individual technical capacity.

The EU AI Act, which entered into force in August 2024 and whose provisions for high-risk systems will be fully applicable between August 2026 and August 2027, imposes rigorous requirements for AI systems in the education sector. These include risk management, quality and governance of training, validation, and test data, exhaustive technical documentation, informative transparency towards users, the need for effective human oversight, and adequate levels of accuracy, robustness, and cybersecurity. It is important to note that the prohibition of certain AI systems considered "unacceptable risk," such as emotion recognition in educational or workplace contexts and social scoring, is effective from February 2, 2025.

A well-executed DPIA should not be seen merely as a regulatory compliance exercise, but as a valuable strategic opportunity for solution co-design. By involving various stakeholders (teachers, students, families, technical staff) during the DPIA process, perceived risks can be more fully identified, especially those of concern to students and parents regarding privacy and surveillance. This early participation facilitates the definition and acceptance of mitigation measures, transforming the DPIA from a potential regulatory burden into a dynamic tool for dialogue, continuous improvement, and trust-building.

The following table summarizes the key provisions of the EU AI Act and relevant DPIA requirements for sensitive content detection systems in schools:

Requirement Category	Key Provisions of the EU AI Act (for high-risk systems in education)	Key Elements of the DPIA (GDPR)	Implications for Schools
Risk Classification	Systems for determining access/admission to educational institutions, evaluating learning outcomes, or detecting prohibited student behavior are generally high-risk (Annex III, AI Act). Prohibition of AI for emotion recognition in education (Art. 5, AI Act).	Assessment of whether processing involves high risk to rights and freedoms (Art. 35 GDPR). Consider DPA/EDPB criteria for when a DPIA is mandatory (e.g., systematic large-scale monitoring of minors).	Sensitive content detection will likely require a DPIA due to the processing of minors' data and the nature of the monitoring.
Risk Management	Establish, implement, document, and maintain a continuous risk management system throughout the AI system's lifecycle (Art. 9, AI Act).	Identification and assessment of risks to data subjects' rights and freedoms. Description of planned measures to address risks.	The DPIA must be integrated with the risk management system required by the AI Act. Identify privacy, discrimination, error risks, impact on freedom of expression.
Data Governance and Quality	Training, validation, and testing data must	Assessment of the necessity and proportionality of	Detail what student data will be processed, for

	be relevant, representative, error-free, and complete. Measures to detect, prevent, and mitigate biases (Art. 10, AI Act).	processing. Sources, types, and categories of personal data processed.	what purposes, and how its quality and representativeness will be ensured, minimizing biases.
Technical Documentation and Records	Maintain detailed technical documentation (Annex IV, AI Act). Automatic logging capability (Art. 12, AI Act).	Systematic description of the planned processing. Documentation of mitigation measures.	The DPIA is part of the necessary documentation. Logs are crucial for auditing and incident investigation.
Transparency and Information Provision	Systems designed to interact with people must inform that they are interacting with AI. Users of high-risk systems must receive clear and adequate information (Art. 13, AI Act).	Information provided to data subjects (Arts. 13 and 14 GDPR). Consultation with data subjects or their representatives, if applicable.	The specific privacy policy and communication to students and parents must be clear about AI use, data collected, and rights.
Human Oversight	High-risk systems must be designed to allow for effective human oversight (Art. 14, AI Act).	Definition of roles and responsibilities for human oversight of system alerts and decisions.	Establish clear workflows for human review of AI detections before significant action is taken.

Accuracy, Robustness, and Cybersecurity	High-risk AI systems must achieve an adequate level of accuracy, robustness, and cybersecurity throughout their lifecycle (Art. 15, AI Act).	Evaluation of data security measures (technical and organizational).	Implement robust security measures (encryption, access controls) and evaluate the AI system's resilience.
Consultation with DPO and Supervisory Authority	(Not specific to AI Act, but implicit in general framework)	Mandatory consultation with the DPO. Consultation with the supervisory authority if the DPIA indicates a high residual risk (Art. 36 GDPR).	The DPO must be actively involved. If high unmitigable risks persist, the national DPA must be consulted.

12.2.4 Stakeholder Engagement and Co-design

The implementation of AI systems for sensitive content detection cannot be a top-down process; its success and legitimacy fundamentally depend on the active and early participation of the entire educational community. Involving management, educators, technical and administrative staff, students, and parents/guardians from the initial planning phases is crucial.

The benefits of this participatory approach are multiple:

Collection of diverse perspectives and needs: Each group possesses a unique understanding of the challenges, risks, and potential benefits. Students can express their concerns about privacy and the impact on freedom of expression; teachers can contribute their vision on pedagogical applicability and workload; parents can share their expectations regarding their children's safety and well-being; and technical staff can assess integration feasibility.

Fostering acceptance and reducing resistance to change: When stakeholders feel heard and involved in decision-making, they are more likely to support the initiative and collaborate in its success. Resistance to change, natural when introducing

monitoring technologies, can be significantly mitigated through open and constructive dialogue.

Ensuring legitimacy and perceived utility: A solution developed in collaboration with the educational community will be perceived as more legitimate, necessary, and beneficial, increasing the likelihood of effective and sustainable adoption.

Co-design strategies go beyond mere consultation. They involve actively engaging stakeholders in designing usage policies, alert management protocols, human oversight workflows, and communication strategies. The Fair-AIEd project, although focused on Global South contexts, offers a valuable model for multi-stakeholder participation (including industry, government, academia, and civil society) to develop algorithmic impact assessment tools and trust frameworks for AI in education. This type of approach can be adapted for the school context.

It is particularly important to conduct focus groups with students to understand their specific concerns about privacy, surveillance, and the potential impact on school culture. The Skills Upload Jr. report reveals that 46% of European students fear that AI could lead to discrimination and 49% that it could increase existing inequalities. These perceptions must be heard and addressed. Institutions like The Alan Turing Institute offer specific training on how to involve stakeholders in the development and procurement of AI systems, which can be a useful resource for educational institutions.

The mandatory requirement to conduct a DPIA should not be seen solely as a bureaucratic requirement, but as a strategic opportunity to integrate co-design. Directly involving stakeholders in the DPIA process allows for a richer and more nuanced identification of risks, especially those perceived by students and parents, and facilitates the co-creation of mitigation measures that are both effective and accepted by the community. This approach transforms the DPIA into a tool for dialogue and continuous improvement, rather than a mere compliance exercise.

12.3 Phase 2: Selecting the Right Technological Solution

Once the educational institution or body has solidified a clear and detailed understanding of its specific needs, the school context, regulatory imperatives, and the diverse perspectives of its stakeholders, it is in a position to address the crucial phase of selecting the AI technology. The choice of the sensitive content detection tool or platform is a high-impact decision, as it will directly determine the operational effectiveness, ethical robustness, and long-term sustainability of the entire initiative.

12.3.1 Establishing Detailed Selection Criteria

The criteria for selecting an AI solution must stem directly from the defined objectives and the needs analysis conducted in Phase 1. This is not a generic list, but a set of requirements adapted to the reality and aspirations of the institution. In addition to the fundamental criteria already mentioned in the original draft, it is essential to delve deeper into and expand on some key aspects for the European school context:

Accuracy and Reliability Specific to the School Context:

- The system's ability to detect with high accuracy the priority sensitive content types identified by the institution (e.g., cyberbullying with specific manifestations, hate speech adapted to local slang, CSAM, self-harm indicators, extremist proselytism, misinformation relevant to students) is paramount.
- It is essential to require providers to provide documented and auditable false positive and false negative rates, preferably validated in European educational environments or with datasets that reflect the demographic and youth language of the region. Research on the accuracy of AI detectors and the inherent problems of false positives and negatives should inform this evaluation.
- The solution must demonstrate a sophisticated ability to understand the context, nuance, and intent behind the language used by young people, including slang, neologisms, coded language, and cultural and linguistic specificities present in European diversity.

Adaptation, Customization, and Multilingual Support Capabilities:

- The possibility of fine-tuning or advanced configuration for the specific needs of the institution or organization is a critical factor. This includes the ability to add new categories of sensitive content as new threats emerge, or to adapt the model to local, regional slang or the evolution of youth language.
- Robust and demonstrated support for multiple European languages, not just majority languages, is required. Recent research warns that many AI tools are predominantly trained on English, Chinese, and Spanish data, which can compromise their security and accuracy in languages with fewer digital linguistic resources. The evaluation of multilingual accuracy and cultural nuance sensitivity is, therefore, vital.

Total Cost of Ownership (TCO) and Financial Sustainability:

- An exhaustive TCO analysis is essential, extending beyond the initial acquisition price. It must include licensing costs (recurring or perpetual), necessary infrastructure (hardware, software, cloud services), customization, integration with existing systems (LMS, SIS, networks), continuous and specialized staff training (technical, pedagogical, and review staff), annual maintenance, model and software updates, and technical support. The use of tools like CoSN's TCO Assessment Tool, adapted to the European context, can be very useful.
- Transparent and predictable pricing models should be sought, avoiding hidden costs or disproportionate increases when scaling usage or user numbers.
- The long-term financial sustainability of the solution must be a primary consideration, ensuring that recurring costs align with school budgets and potential funding sources.

Scalability and Technical Sustainability:

- The solution must be capable of efficiently scaling to adapt to the growth in the volume of data generated by students, the increasing number of users (students and staff), and the possible expansion of monitoring to new platforms or devices.
- It is important to know the provider's roadmap regarding model updates, functionality improvements, and adaptation to new emerging threats and the evolution of AI technologies.

Regulatory Compliance and Robust Security:

- The provider must offer explicit and verifiable guarantees regarding compliance with the GDPR, the EU AI Act (especially if the solution is classified as high-risk), and other relevant online child protection laws in the European and national spheres.
- The provider's data management policies must be explicit and detailed, covering aspects such as data storage location (preferably within the EU), access protocols, retention periods, and secure data deletion procedures.

Explainability, Transparency, and Auditability:

- The degree to which the AI model's detection decisions can be understood and explained is fundamental. This is vital for accountability and for staff to trust and validate alerts. The European Digital Education Hub session on Explainable AI underscores this need.

- The ability to audit system decisions, the data that underpinned them, and any inherent model biases is a key requirement, especially under the EU AI Act.
- The provider must be transparent regarding the training data used, methodologies for bias mitigation, and known model limitations.

Technical and Vendor Support:

- The availability, quality, response times, and linguistic competence (in relevant European languages) of technical support are crucial for efficiently resolving incidents.
- Evaluate the provider's commitment to continuous improvement of their solution, their adaptability to the changing needs of the European education sector, and their proactivity in incorporating feedback from institutions.
- The clarity and thoroughness of Service Level Agreements (SLAs) that define mutual responsibilities and expectations.

12.3.2 Evaluation of Alternatives: Proprietary vs. Open Source vs. Hybrid

The choice between proprietary, open-source, or hybrid AI models is a strategic decision with profound implications for the educational institution's resources, control, and responsibility.

Proprietary Models:

- Advantages: Generally, these solutions, offered by commercial companies, come with a higher level of technical and user support, "out-of-the-box" functionalities with more polished interfaces, and often more explicit guarantees of regulatory compliance (e.g., GDPR certifications or alignment with the AI Act). Providers can bear a significant portion of the burden of risk management, platform security, and continuous updates.
- Disadvantages: The main drawback is less control and flexibility over the internal workings of the model. License costs are often high and recurring, which can be a significant burden for school budgets. There is a risk of vendor lock-in, making it difficult to migrate to other solutions in the future. Furthermore, opacity in the internal functioning of algorithms ("black box") can hinder explainability and independent auditing.

Open Source Models:

- Advantages: The main advantage lies in flexibility and customization capability. By having access to the source code, institutions with technical capacity can adapt the solution to their exact needs, audit the algorithm,

and potentially contribute to its improvement. There are no direct licensing costs, which can be attractive. The inherent transparency of open source is valued by organizations like the Open Source Initiative (OSI), which actively works to ensure that regulations like the EU AI Act do not impose disproportionate barriers to the development and use of open-source AI, successfully ensuring, for example, that acceptable use policies are optional for these models in recent drafts of the EU's General Purpose AI Code of Practice.

- Disadvantages: The implementation and maintenance of open-source models entail a high burden of development, configuration, and technical management that falls entirely on the educational institution. This includes full responsibility for regulatory compliance (GDPR, AI Act, children's data protection), system security, and vulnerability management. It requires highly qualified technical staff in AI, cybersecurity, and data protection, a scarce and costly resource for many institutions. Technical support largely depends on the developer community, which may not offer the immediacy or service level required by a school environment.

Hybrid Models:

- These models seek to combine elements of proprietary and open-source solutions. An example could be the use of an open-source content detection model as a base, complemented by proprietary alert management services, user interfaces, or reporting modules. Another modality is the use of AI for a first layer of automated moderation, escalating more complex or ambiguous cases to qualified human review.
- Potential Advantages: They can offer a balance between the flexibility and customization of open source and the support and ease of use of proprietary solutions. Costs could be optimized by combining free components with specific paid services.
- Potential Disadvantages: The complexity of integrating different components from different origins can be a technical challenge. The delineation of responsibilities regarding security, maintenance, and regulatory compliance between open-source and proprietary components can be diffuse and require careful contractual and technical management.

The final decision among these alternatives will depend critically on a realistic assessment of internal technical resources (availability of staff with experience in AI, software development, cybersecurity, and data management), financial resources (budget for licenses, development, specialized staff, and long-term maintenance), and the institution's capacity and appetite to manage the inherent

risks of each model. It is a strategic choice that must align with the institution's long-term vision for technology use.

12.3.3 Proofs of Concept (PoC) and Thorough Vendor Evaluation

Before making a final decision and committing significant resources, it is highly recommended, and in many cases essential, to conduct Proofs of Concept (PoC) with the shortlisted technological solutions. In parallel, if proprietary or hybrid solutions with commercial components are being considered, a thorough evaluation of the vendors must be carried out.

Proofs of Concept (PoC):

Strategic Importance: PoCs allow educational institutions to validate the feasibility and effectiveness of one or more AI solutions in their specific context, before large-scale implementation. This step is crucial given that, as the Council of Europe points out, there is little independent evidence on the efficacy or safety of many AI-enabled EdTech technologies. PoCs help mitigate the risk of investing in a solution that does not meet expectations or does not adapt well to the school environment. The UK's Department for Education (DfE) itself has conducted PoCs with AI tools for grading, demonstrating their usefulness for evaluating technology in practice.

Methodology for Effective PoC in School Contexts:

1. Clear Definition of the Problem and PoC Objectives: Specify what concrete aspect of sensitive content detection will be tested (e.g., cyberbullying detection on a specific platform, identification of self-harm language in student forums) and what the PoC's success criteria are.

2. Preparation of Relevant Data: Use anonymized or synthetic datasets that accurately reflect the real context of the institution. This includes the main language(s), common student slang, priority sensitive content types, and the platforms where this content is generated. The quality and representativeness of this data are crucial for a meaningful evaluation.

3. Selection of a Controlled and Representative Environment: Choose a limited scope for the PoC, such as a specific course, educational level, or particular digital platform.

4. Establishment of Clear Evaluation Metrics: Define beforehand how the solution's performance will be measured during the PoC. These metrics should go beyond simple general accuracy and include:

- Correct detection rate for each priority sensitive content type.
- False positive rate (incorrect alerts) and its impact on the workload of review staff.
- False negative rate (undetected sensitive content) and its implications for security.
- Ease of use of the interface for staff managing alerts.
- System response time for detection and notification.
- Ability to adapt to linguistic and cultural nuances of the institution.

5. Involvement of End Users: Involve the staff (teachers, counselors, IT personnel) who will be responsible for reviewing alerts and managing incidents. Their feedback on usability, alert clarity, and workflow efficiency is invaluable.

6. Rigorous Documentation: Meticulously record the PoC process, data used, system configuration, results obtained (quantitative and qualitative), and all lessons learned.

Support Resources: Although not specific to sensitive content detection, general guides for conducting PoCs in AI projects and templates, such as those offered by Microsoft for Copilot PoCs, can be adapted.

12.4 Phase 3: Detailed Implementation Design and Data Governance

Once the technological solution has been carefully selected, the next crucial phase is the detailed design of the implementation plan and the establishment of a robust framework for data governance. This stage is where the strategy and the chosen technology are translated into concrete actions, ensuring that the AI system is integrated securely, efficiently, ethically, and in full compliance with the complex European regulatory framework. A deficient design in this phase can compromise even the best selected technology.

12.4.1 Technical Implementation Plan and System Architecture

The technical plan must address all aspects of infrastructure, integration, and security necessary for the optimal and secure functioning of the AI system.

Infrastructure and Deployment:

- Deployment Model: An informed decision must be made whether the solution will be deployed in the cloud (public, private, or community), on-premise servers, or through a hybrid model. Each option has implications for costs, security, data sovereignty, scalability, and maintenance. Cloud solutions may offer greater scalability and lower initial infrastructure costs but raise considerations about data localization and vendor dependence. On-premise solutions offer greater control over data but require significant initial investment and technical expertise for management and maintenance. Hybrid models can seek a balance but add complexity to management.
- Network Capacity and Storage: It is vital to ensure that the institution's network infrastructure (bandwidth, latency) can support the volume of data that the AI system will generate and process, especially if it involves video analysis or large volumes of text in real-time. The necessary storage capacity must be planned, considering data retention and log requirements.
- Device Compatibility: Ensure system compatibility with devices used by students and staff.

Integration with Existing Systems:

- Meticulously plan how the AI solution will integrate with systems already in use at the educational institution. This is fundamental for efficient workflow and to avoid creating data silos.
- Learning Management Systems (LMS): Integration with platforms such as Moodle, Google Classroom, Microsoft Teams for Education, itslearning, Fronter, etc., to monitor activity in forums, assignments, or internal messaging, if defined in the scope. Some LMS are already incorporating their own AI tools or facilitating integrations.
- Communication Platforms: Integration with school email systems, approved instant messaging applications, or collaborative platforms.
- Student Information Systems (SIS): To cross-reference relevant information (e.g., to contextualize alerts) securely and with data minimization, always under strict privacy and need-to-know controls.

- Wi-Fi Networks and Mobile Device Management (MDM): If monitoring includes activity on the school network or on school-managed devices, integration with these systems will be necessary.
- Interoperability standards should be considered to facilitate these integrations.

Data and System Security (Cybersecurity):

- Define and implement robust technical security measures to protect personal data (especially minors' data) processed and stored by the AI system. This is a requirement of both the GDPR and the EU AI Act.
- Encryption: Encrypt data both in transit (e.g., TLS/SSL) and at rest (e.g., AES-256) for all sensitive information handled by the system.
- Access Controls: Implement a role-based access control (RBAC) system to ensure that only authorized personnel can access AI system data and functionalities, and only to the extent necessary for their duties (principle of least privilege). This is especially critical for access to alerts and student data.
- Multi-Factor Authentication (MFA): Require MFA for access to AI system administration consoles and for personnel reviewing sensitive alerts.
- Underlying Infrastructure Security: Whether in the cloud or on-premise, the infrastructure supporting the AI system must be adequately secured (firewalls, intrusion detection/prevention systems, vulnerability management, regular security patching).
- Audit Logs: The system must generate detailed activity logs (accesses, queries, modifications, alerts generated and managed) to allow for traceability and incident investigation. The AI Act requires logging capabilities for high-risk systems.
- Anonymization and Pseudonymization: Apply anonymization or pseudonymization techniques to data whenever possible, especially for model training or improvement, and for generating statistical reports. However, the EDPB warns that true anonymization in AI models is complex and requires rigorous case-by-case evaluation.
- Security Incident Response Plan: Have a plan for managing potential security breaches or privacy incidents related to the AI system, including notification to authorities (DPAs) and affected parties, as per GDPR requirements.

12.4.2 Development of Clear and Detailed Policies and Protocols

Parallel to the technical design, it is fundamental to develop a set of policies and protocols that regulate the use of the AI system, the management of the information it generates, and the resulting actions. These documents must be clear, accessible, aligned with current regulations, and effectively communicated to the entire educational community.

Acceptable Use Policy (AUP) for AI Technology:

- Aimed at students and staff, establishing the permitted and prohibited uses of the AI system and monitored platforms. It must be specific about how AI for sensitive content detection integrates into the school's digital environment.
- Must include guidelines on creating and sharing online content, respecting intellectual property, and the consequences of misuse.
- Monsha.ai's guide and SWGfL's template can serve as references for developing these policies.

Alert Management Protocol:

- Reception and Prioritization: Define who or which roles (e.g., DSL, OSL, well-being team) receive AI-generated alerts. Establish clear criteria for prioritizing alerts based on their severity and urgency.
- Initial Investigation: Detail the steps for initial alert investigation, including information verification, contextualization, and assessment of AI detection credibility. This is crucial for minimizing the impact of false positives.
- Escalation Criteria: Establish clear thresholds and criteria for deciding when an alert should be escalated to higher levels of intervention (e.g., management, social services, law enforcement). Charterhouse School's protocol offers an example of how to structure these procedures.
- Logging and Documentation: All alerts, investigations, and actions taken must be meticulously recorded in a secure system, complying with GDPR and AI Act documentation requirements.

Intervention and Support Protocol:

- Steps to Follow: Define concrete steps to follow once a sensitive content incident is confirmed. This should include differentiated actions based on the nature and severity of the content (e.g., cyberbullying, self-harm, exposure to CSAM, radicalization).
- Victim Support: Establish clear procedures for offering immediate and long-term support to affected students or victims, including psychological support, protection measures, and accompaniment.

- Measures for the Offender: Define educational, restorative, or disciplinary measures for students who generate or distribute sensitive content, in accordance with the school's internal regulations and applicable law.
- Collaboration with Families and External Agencies: Protocols for communication and collaboration with parents/guardians and, when necessary, with social services, health services, or law enforcement, always respecting data protection regulations.

Specific Privacy Policy for the AI System:

- Draft or update the school's privacy policy to specifically detail the personal data processing carried out by the AI system.
- Must inform clearly and accessibly (especially for minors) about: what data is collected, for what specific purposes, who accesses it, how long it is retained, implemented security measures, and data subjects' rights.
- This policy must be easily accessible to students, parents, and staff.

Procedures for Exercising Data Subject Rights:

- Establish clear and simple procedures for students and their parents/guardians to exercise their rights under the GDPR: access to their data, rectification of inaccurate data, erasure of data (right to be forgotten), restriction of processing, data portability, and objection to processing, including automated decisions.
- Designate a clear point of contact (usually the DPO) for managing these requests.
- Akira AI describes how AI agents can help automate some GDPR compliance monitoring, although ultimate responsibility lies with the institution.

12.4.3 Designing the Workflow with Essential Human Oversight

As repeatedly emphasized in previous chapters, and in line with the requirements of the EU AI Act for high-risk systems, human oversight is a non-negotiable component in the implementation of AI systems for sensitive content detection in school environments. Exclusive reliance on automated decisions is unacceptable given the risks of errors, biases, and potential impact on students.

Workflow Design:

- A clear workflow must be designed to ensure that all significant AI-generated alerts are reviewed and validated by trained human personnel before any disciplinary, support intervention, or external party notification (e.g., families, authorities) action is taken.
- This workflow must define:
    - How alerts are received and logged.
    - Who performs the first review and triage (e.g., to discard obvious false positives).
    - How the alert is investigated and contextualized (e.g., by reviewing communication history if relevant and legally permitted, consulting with the homeroom teacher).
    - Who makes the final decision on the validity of the alert and the actions to follow.
    - How the entire process is documented.

The goal is a human-in-the-loop model, where AI acts as a supporting tool to identify potential risks, but final interpretation and decision-making rest with human professionals.

Identification and Training of Human Reviewers:

- Staff Selection: Clearly define which professional profiles will be responsible for alert review. These may include members of the management team, school counselors, psychologists, social support staff, or teachers with specific training in safeguarding and student well-being (e.g., the DSL or OSL as in the Charterhouse School model).
- Specific Training: These reviewers will need comprehensive training covering:
    - Basic technical operation of the AI tool and its limitations (understanding false positives and negatives).
    - The alert management and intervention protocols defined by the institution.
    - The ethical and privacy implications of monitoring and reviewing student data.

    - Applicable regulations (GDPR, AI Act, child protection laws).
    - Identification and handling of different types of sensitive content.
    - Strategies for addressing difficult conversations with students about sensitive content.
    - Awareness and mitigation of their own biases in interpreting alerts.
    - Referral procedures to internal or external support services.

Training must be continuous and updated as threats and technology evolve.

Ethical Considerations in Human Oversight:

- Establish an ethical framework for decision-making by human reviewers.
- Ensure the confidentiality and respectful treatment of sensitive information accessed by reviewers.
- Implement control and audit mechanisms over reviewers' access and actions to prevent abuse or overreach.
- Provide psychological support and supervision to reviewers, given that repeated exposure to sensitive content can impact their own well-being.

The effective integration of human oversight is what can transform an AI system from a potentially intrusive surveillance tool into an early warning system that truly supports student safety and well-being, always within a framework of respect for their fundamental rights.

12.5 Phase 4: Change Management, Strategic Communication, and Training

The introduction of AI technologies for sensitive content detection in educational institutions represents a significant change, not only technological, but also cultural and operational. Planned, sensitive, and proactive change management, accompanied by transparent communication and comprehensive training, is indispensable for successful implementation and to foster the trust and collaboration of the entire educational community.

12.5.1 Transparent and Continuous Communication Strategy

Open, honest, and continuous communication is the cornerstone for building the trust and acceptance necessary for this type of initiative. It is vital that all members of the educational community – students, parents/guardians, teaching and non-teaching staff, and governing bodies – understand the purpose, general operation, and safeguards of the system.

Developing a Multifaceted Communication Plan:

This plan must identify the different target audiences and adapt messages and channels to each.

Communication Content:

- The purpose and expected benefits: Clearly explain why the system is being implemented (e.g., to improve online safety, prevent cyberbullying, support student well-being) and how it is expected to benefit students and the school environment in general.
- The general operation of the system: Describe in an understandable way how AI operates to detect sensitive content, without revealing technical details that could be exploited to circumvent the system. It is important to be honest about the capabilities and limitations of AI, including the possibility of false positives and negatives.
- What data is collected, how it is used, and how it is protected: Specify what type of data will be monitored (e.g., text on school platforms, images on school devices), who will have access to it, for what purposes it will be used (exclusively for sensitive content detection and student support), how long it will be stored, and the security measures implemented to protect privacy and confidentiality, in line with the GDPR and the system's specific privacy policy.
- Associated policies and protocols: Inform about the existence of the Acceptable Use Policy, the Alert Management Protocol, the Intervention Protocol, and the specific Privacy Policy, and how to access them.
- The role of human oversight: Emphasize that important decisions are not made autonomously by AI, but that there is always review and validation by qualified personnel.
- The rights of students and parents: Explain how they can exercise their rights to access, rectify, and delete data, and whom to contact for this.
- Channels for expressing concerns or making inquiries: Establish clear and accessible channels for any community member to ask questions, express concerns, or request clarifications.

Communication Channels: Use a variety of channels to ensure that information reaches all audiences: informational meetings (in-person and online), circulars, newsletters, school website, school communication platforms, parent workshops, and specific sessions with students adapted to their age and maturity.

Timing: Communication should not be a one-time event, but a continuous process that begins before implementation, intensifies during the pilot phase and deployment, and is maintained over time with periodic updates.

12.5.2 Comprehensive Training and Continuous Professional Development

Adequate staff training is a fundamental pillar for the successful and sustainable implementation of AI. The EU AI Act, in fact, establishes the mandatory nature of AI literacy for personnel operating and using AI systems, effective from February 2025. This training must go beyond simple technical handling of the tool.

Target Audience for Training:

- Personnel directly involved in alert review and incident management: (e.g., management team, counselors, DSL/OSL, school psychologists). This training should be the most intensive and specialized.
- General teaching staff: Need to understand how the system works, how it can impact their classrooms, how to educate students about responsible technology use, and how to act if a student reports an incident or if an alert involves them.
- Technical and IT staff: Responsible for system maintenance, integration, and security.
- Governing bodies and DPO: To understand the strategic, legal, and ethical implications.

Key Training Content:

AI Literacy:

- Basic concepts of AI and machine learning: what it is, how it works at a conceptual level, types of AI.
- Specific operation of the implemented content detection tool.
- Capabilities and, crucially, limitations of AI: deep understanding of false positives and false negatives, and how AI can misinterpret context, sarcasm, or slang.
- Risks of algorithmic biases (cultural, linguistic, gender, etc.) and how they can affect detections.

Tool Use and Protocols:

- Practical handling of the AI system interface (for those who will use it).
- Interpretation of alerts and reports generated by the system.
- Detailed knowledge of the alert management and intervention protocols defined by the institution: who does what, when, and how.
- Procedures for documenting and logging incidents.

Ethical, Legal, and Privacy Aspects:

- Ethical principles for AI use in education (transparency, fairness, non-maleficence, responsibility, human-centered).
- GDPR requirements applicable to student data processing, especially sensitive data and AI-based decisions. Data subjects' rights.
- Implications of the EU AI Act, especially if the system is considered high-risk.
- School privacy and acceptable use policies.

Intervention and Communication Skills:

- Strategies for addressing sensitive and supportive conversations with students who have been victims or perpetrators of problematic content.
- How to communicate decisions and actions taken to students and families clearly, respectfully, and constructively.
- Knowledge of available internal and external support resources for students.
- Continuous Professional Development: Training should not be an isolated event. Regular update sessions should be planned to address new threats, technology changes, lessons learned, and refresh knowledge. Foster a community of practice among staff to share experiences and best practices.

Training Modalities:

Combine different modalities: practical workshops, online sessions, self-learning modules, case studies, simulations, and peer mentoring.

Resources such as those offered by European Schoolnet Academy, the European Digital Education Hub, or TeachAI can be useful.

The lack of adequate staff preparation not only reduces the effectiveness of the AI tool but also increases the risk of errors, misinterpretations, privacy violations, and harm to student well-being. A solid investment in training is an investment in responsible and effective AI implementation.

12.5.3 Proactively Addressing Resistances and Concerns

It is natural and expected that the introduction of AI systems for content monitoring will generate resistance and concerns among various members of the educational community. These concerns often center on student privacy, fear of excessive surveillance ("Big Brother effect"), the reliability and fairness of AI algorithms, and the potential impact on school culture and the trust relationship between students and educators. Ignoring or minimizing these concerns can severely undermine the project.

Create Spaces for Open Dialogue and Active Listening:

- Organize forums, meetings, and specific Q&A sessions for each stakeholder group (students, parents, teachers, non-teaching staff).
- Ensure these spaces are safe and that all voices can express themselves freely without fear of reprisal.
- Practice active listening: not just hearing, but genuinely understanding concerns and validating them, even if not shared.

Provide Clear and Honest Information:

Many concerns arise from misinformation or lack of understanding. Reinforce transparent communication about the system's objectives, its limited operation (without revealing details that could be circumvented), privacy safeguards, the crucial role of human oversight, and action protocols.

Be explicit about AI limitations, including the possibility of errors, and how these will be managed.

Involve Constructive Critics:

Identify community members who, though skeptical, raise legitimate and constructive concerns. Involving them in working groups or oversight committees can transform their skepticism into valuable contributions to improve the system and protocols.

Adapt Implementation in Response to Feedback:

Change management is not a one-way process. Being willing to make adjustments to technology configuration, policies, or protocols based on received feedback demonstrates that concerns are taken seriously and increases acceptance. For example, if students express strong concern about monitoring certain platforms or communication types, the scope can be re-evaluated.

Highlight Benefits and Protective Purpose:

Focus the narrative on how AI will be used to create a safer and more supportive learning environment, and how it will help protect students from serious harm such as cyberbullying or exposure to harmful content.

Build Change Champions:

Identify and support respected members of the educational community (teachers, students, parents) who understand and support the initiative. They can act as ambassadors, helping to explain benefits and address peers' concerns.

Monitor School Climate and Trust:

Once the system is implemented, even in a pilot phase, it is important to monitor its impact on the school climate and trust levels. Anonymous surveys or periodic focus groups can help identify if initial concerns persist or if new ones arise.

Overcoming resistance to change and effectively addressing concerns is an ongoing process that requires empathy, transparency, and a genuine willingness to collaborate on the part of school leadership. The perception that AI is implemented with the community, and not on the community, is fundamental to its long-term success. A key strategy can be the creation of an ethics committee or a diverse oversight group, including representation from all stakeholders, to periodically review the system's operation and its implications.

12.6 Phase 5: Pilot Implementation and Gradual Rollout

A hasty and large-scale deployment of a technology as sensitive as AI models for content detection in school environments carries significant risks. Unforeseen technical problems, the inadequacy of protocols to the school's reality, or a negative response from the educational community can be magnified if implementation is massive from the outset. Therefore, a pilot implementation approach followed by a gradual and staggered rollout is the most prudent and effective strategy.

12.6.1 Critical Importance of Pilot Testing

Conducting a pilot implementation in a controlled environment and on a reduced scale is an indispensable step before a full deployment. This pilot acts as a "test bench" under real conditions, allowing for the identification and correction of problems before they affect the entire institution.

Pilot Scope: The pilot could be limited to:

- A specific course or educational level.
- A particular department or subject.

- A concrete digital platform (e.g., the LMS or email system).
- A small, voluntary group of teachers and students.

Pilot Objectives:

- Test the technology under real conditions: Evaluate the performance of the AI model with the institution's authentic data and communication patterns. This includes verifying accuracy, false positive/negative rates in the real environment, and suitability for the slang and languages used by students.
- Identify technical or usability issues: Detect integration failures with existing systems, network bottlenecks, device compatibility issues, or difficulties in using the interface by assigned staff.
- Evaluate protocol effectiveness: Test the alert management and intervention protocols. Are they clear? Are they practical? Do they allow for a quick and appropriate response? Are roles and responsibilities well-defined?
- Measure workload: Estimate the time and resources that staff will need to dedicate to reviewing alerts and managing incidents.
- Collect early and specific feedback: Obtain initial impressions and suggestions from direct users (students aware of participating in a pilot and staff managing alerts). This feedback is invaluable for making adjustments.
- Validate the communication and training strategy: Observe how information about the pilot is received and whether the initial training has been adequate.

The Croatian case study on AI integration in K-12, though focused on programming instruction, offers a three-year pilot methodology with theoretical instruction and project-based learning phases, along with teacher training and continuous support, which can inspire the design of pilots for security tools.

12.6.2 Systematic Feedback Collection and Iterative Adjustments

During the pilot phase and immediately after its conclusion, it is crucial to implement a robust system for collecting feedback from all involved participants and observers. This feedback is the driving force for making necessary adjustments before considering a wider rollout.

Feedback Collection Methods:

- Anonymous surveys: For students and staff, asking about their experience, concerns, and suggestions for improvement. The Skills Upload Jr. report demonstrates the usefulness of surveys for capturing student perspectives.
- Focus groups: With small groups of students, teachers, and review staff for deeper, more nuanced discussions.

- Individual interviews: With key participants to obtain detailed information.
- Dedicated communication channels: A specific email or form for reporting problems or suggestions during the pilot.
- Direct observation: By the implementation team to identify difficulties or friction points in system and protocol use.
- System log analysis: To identify usage patterns, technical errors, or recurring alerts.

Types of Feedback to Collect:

- On technology: Detection accuracy, number and type of false positives/negatives, interface usability, system performance.
- On protocols: Clarity, efficiency, suitability for real situations, generated workload.
- On communication and training: Effectiveness of messages, clarity of information, usefulness of received training.
- On perceived impact: Sense of security, privacy concerns, impact on trust climate.

Iterative Adjustment Process:

- Systematically analyze all collected feedback.
- Prioritize identified problems and areas for improvement.
- Make necessary adjustments to:
    - AI tool configuration (e.g., sensitivity thresholds, keyword lists).
    - Policies and protocols (e.g., clarifying roles, improving intervention steps).
    - Training materials and methods.
    - Communication strategy.

If adjustments are significant, consider the possibility of a second, more limited pilot phase to validate changes before general deployment.

Restack.io's guide on feedback loops in AI pilot programs, while enterprise-oriented, offers metrics and an iterative improvement approach (data collection, analysis, refinement) that are directly applicable to the educational context.

12.6.3 Progressive and Planned Scaling

Once the pilot phase has successfully concluded, results have been analyzed, feedback collected and processed, and necessary adjustments made, the educational institution can begin planning the scaling of the implementation. This scaling must be progressive and carefully managed, not a massive overnight deployment.

Scaling Strategies:

- Phased Approach: Expand implementation gradually to other educational levels, departments, or digital platforms, rather than all at once. Each phase can have its own mini-pilots or testing periods.
- Maturity and Preparedness-Based: Prioritize deployment in areas or groups that demonstrate greater preparedness and willingness, using lessons learned to address more complex contexts later.
- Accompanied by Continuous Training: Ensure that each new group of users receives adequate training before the system is activated for them.
- Intensive Monitoring During Scaling: Maintain close vigilance over system performance and community response as the scope expands.

Considerations for District-Level or Network-Level Scaling:

- If implementation is carried out at the school district level or a network of institutions, scaling will require even greater coordination.
- Lessons learned in initial pilot institutions must be shared and adapted to the specificities of each new institution.
- Variations in technological infrastructure, staff competencies, and organizational culture among different institutions must be considered.
- Establish mechanisms for support and sharing of best practices between institutions that have already implemented the solution and those that are in process.

The challenges of scaling EdTech solutions, especially those involving AI, are considerable and include the need to adapt the solution to different contexts, ensure interoperability, manage data privacy at a larger scale, and maintain the quality of support and training. A progressive approach allows these challenges to be addressed more manageably and ensures a smoother transition to full adoption.

12.7 Phase 6: Monitoring, Continuous Evaluation, Maintenance, and Adaptation

The implementation of an AI system for sensitive content detection is not a project with a defined end, but a dynamic and continuous process that requires constant attention and adaptation. Once the system is deployed, whether partially or fully, a crucial phase of monitoring its performance, evaluating its impact, technical maintenance, and adaptation to a perpetually evolving digital and social environment begins.

12.7.1 Establishment of Key Success Metrics (KPIs) and Their Monitoring

To measure the effectiveness of the AI solution and justify its continuation, it is fundamental to define clear, measurable Key Performance Indicators (KPIs) from the outset, aligned with the objectives established in Phase 1. These KPIs must be monitored regularly.

KPIs Oriented Towards Detection Effectiveness:

- Reduction of specific incidents: Percentage decrease in prioritized sensitive content types (e.g., cyberbullying, hate speech) on monitored platforms.
- Average response time to alerts: From when the AI generates an alert until it is reviewed by a human and, if necessary, an intervention is initiated.
- False positive rate: Percentage of AI-generated alerts that, after human review, are determined not to correspond to actual sensitive content. It is crucial to monitor their evolution and work to reduce them, as a high volume can lead to reviewer fatigue and distrust in the system.
- False negative rate: Estimation of the percentage of actual sensitive content that the system fails to detect. This metric is more difficult to measure directly but can be inferred through random manual audits, user reports, or analysis of incidents not alerted by AI.
- Accuracy by content category: Evaluate whether the system is more effective for certain content types than others.

KPIs Oriented Towards Operational Efficiency:

- Number of alerts managed per reviewer/hour: To evaluate workload and efficiency of the review team.
- Cost per incident prevented or managed: Although complex to calculate, it can help assess the return on investment.

KPIs Oriented Towards Satisfaction and Trust:

- User satisfaction level (students, teaching staff, parents) with the system and its impact, measured through periodic surveys.
- Perception of safety and trust in the school digital environment.

KPIs Oriented Towards Equity and Absence of Bias:

- Disparity analysis in alert/intervention rates among different student groups (e.g., by gender, ethnic origin, special educational needs) to detect possible algorithmic biases or biases in protocol application.
- Feedback from minority groups on their experience with the system.
- Estonia's "AI Leap 2025" program, for example, plans to measure success based on the incorporation of schools and teachers, the improvement of

student and teacher skills, and the efficiency and effectiveness of education. These KPIs must be reviewed and adjusted periodically to ensure their continued relevance.

12.7.2 Periodic Comprehensive Impact Evaluation

Beyond specific system KPIs, it is crucial to periodically evaluate the broader impact of AI implementation on the life of the educational institution. This evaluation must be holistic and consider multiple dimensions:

Impact on School Climate and Sense of Security:

- How has the implementation affected the general atmosphere of the institution? Is a safer and more respectful environment perceived?
- Have there been changes in the prevalence or nature of cyberbullying or other online risk behaviors?
- How do students and staff value the system's contribution to their safety?
- Studies like the WHO's on cyberbullying can offer a framework for evaluating changes in prevalence.

Impact on Students' Psychological Well-being:

- Is there any indication that monitoring is generating anxiety, stress, or excessive self-censorship among students?
- Conversely, has the system facilitated early identification of at-risk students and the provision of support, improving their well-being?
- It is vital to consider students' own perceptions of how AI affects their trust and their relationship with the institution. Studies on student trust in educational AI can be relevant.

Impact on Workload and Role of Educators:

- How has the system affected the workload of personnel involved in alert review and incident management?
- Has it freed up teachers' time for other pedagogical or support tasks, or has it added a new layer of responsibilities?
- How do educators perceive that AI is changing their role and their relationship with students?

Impact on Equity and Inclusion:

- Continuously review whether the system is operating equitably for all students, without discriminating against or disproportionately affecting particular groups (e.g., by socioeconomic origin, culture, language, gender, or students with special educational needs).

- Research on AI biases and UNESCO's ethical guidelines are fundamental here.
- Impact evaluation should use a combination of quantitative methods (KPI analysis, surveys) and qualitative methods (interviews, focus groups with students, staff, and families) to obtain a complete and nuanced picture.

12.7.3 Technical Maintenance, Model Updates, and Continuous Security

The reliable and secure functioning of the AI system over time depends on proactive technical maintenance and continuous updating of its components.

System Maintenance:

- Ensure the proper functioning of the underlying hardware and software infrastructure (servers, networks, databases).
- Apply security patches and operating system and application updates regularly to protect against known vulnerabilities.
- Monitor system performance to detect and resolve bottlenecks or failures.

AI Model Retraining and Updating:

- AI models are not static. Their accuracy can degrade over time as language evolves (new slang, emojis, memes), new types of sensitive content emerge, or users find ways to circumvent detection.
- If custom models or continuously learning models (online learning) are used, it is crucial to plan their periodic retraining with new relevant and verified data from the institution itself (anonymized when possible) or with updated datasets provided by the vendor.
- This retraining should include strategies for continuous bias mitigation and adaptation to new evasion tactics.
- For proprietary models, it is important to understand the vendor's policy on the frequency and method of updating their models and how these changes are validated.

AI-Specific Security Updates:

- Keep the AI system and its components updated to protect against vulnerabilities specific to machine learning models (e.g., adversarial attacks, data poisoning).
- Follow security recommendations from ENISA and other cybersecurity bodies.

Post-Market Monitoring (EU AI Act):

- Providers of high-risk AI systems have post-market monitoring obligations under Article 72 of the EU AI Act. This involves actively collecting and analyzing usage data, continuously assessing compliance, and documenting any changes.
- Educational institutions must collaborate with providers in this monitoring and report serious incidents (Art. 73 AI Act).

The long-term sustainability of these solutions depends on the continuous allocation of resources for these maintenance and update tasks.

12.7.4 Continuous Adaptation to a Changing Environment

The digital environment, young people's online behaviors, types of sensitive content, and the AI landscape itself are constantly and rapidly evolving. Therefore, educational institutions must be prepared to continuously adapt their strategies, policies, and AI technology configurations to address new risks and challenges.

Monitoring New Threats and Trends:

- Stay informed about new forms of sensitive content, emerging platforms popular among young people, and the evolution of tactics by those seeking to cause harm online. Reports from Europol (such as IOCTA and specific notifications), ENISA, and child safety organizations are valuable sources.
- Pay attention to the evolution of youth language, including new slang, memes, and coded forms of communication that may not be detected by AI models without updating.

Periodic Review and Update of Policies and Protocols:

- Acceptable use policies, alert management and intervention protocols, and privacy policies should be reviewed at least annually, or more frequently if new risks or significant legislative changes arise.
- Ensure that these reviews involve relevant stakeholders.

Flexibility in Technological Configuration:

- The AI solution should allow for adjustments in its configuration to respond to new threats or to refine its accuracy (e.g., updating keyword lists, adjusting sensitivity thresholds, incorporating new detection modules if offered by the vendor).

Fostering a Culture of Learning and Adaptation:

- Promote a culture within the institution that recognizes that online safety and AI technology are dynamic fields.
- Encourage staff to share observations and propose improvements in processes and tools.
- Integrate education on new risks and ethical AI use continuously into the curriculum and staff training.

Frameworks like the European Training Foundation's "Digital Education Reform Framework" or the Council of Europe's guidelines on AI regulation in education emphasize the need for flexibility and continuous adaptation of educational policies and practices in the digital age. The capacity for adaptation is not just a technical matter, but an institutional mindset necessary to navigate the complexity of AI in education.

12.8 Conclusion: Towards Responsible and Sustainable AI Implementation for School Safety

The practical implementation of artificial intelligence models for sensitive content detection in educational institutions is, without a doubt, a complex and multifaceted journey. It is not a simple technological acquisition, but a transformation process that demands visionary and committed leadership, meticulous strategic planning encompassing all the described phases, and a profoundly people-centered approach involving the educational community. The success of this endeavor does not lie in the search for a "magic solution," but in the recognition of AI as a powerful tool that, when designed, implemented, and managed responsibly, ethically, and legally, can contribute significantly to creating safer, more inclusive learning environments conducive to the holistic development of all students.

The path to effective implementation requires a deep and contextualized understanding of each institution's specific needs, including the nature of online risks its students face and the institutional capacity to manage such advanced technology. The clear definition of objectives and a well-defined scope are crucial from the outset, as is a rigorous impact assessment (DPIA/EIPD) that not only complies with the GDPR and the EU AI Act but also serves as a catalyst for dialogue and co-design with all stakeholders. The active participation of management, educators, technical staff, students, and families throughout the entire implementation lifecycle is not optional, but a fundamental pillar for ensuring the legitimacy, acceptance, and suitability of the solution to the school reality.

Technology selection must be an informed and critical process, based on criteria that prioritize contextualized accuracy for the multilingual European environment, adaptability, transparency, security robustness, and regulatory compliance, without forgetting financial sustainability through total cost of ownership analysis. Proofs of concept (PoCs) and a thorough vendor evaluation are essential steps to validate technological promises in the real world.

Implementation design must translate into a solid technical plan, but, equally importantly, into the development of clear, detailed, and effectively communicated policies and protocols. These must regulate technology use, alert management, and intervention strategies, always with human oversight as an essential and indispensable component of the workflow. No critical decision affecting a student should be made solely by an algorithm.

Change management is, perhaps, the most delicate yet most determining component. A transparent and continuous communication strategy, which openly addresses benefits and risks, and which invites dialogue, is vital for building trust and mitigating resistance. Comprehensive and continuous staff training, especially for those who will review alerts and manage incidents, must go beyond technical handling, encompassing ethical, legal, pedagogical, and communication dimensions. AI literacy, as required by the EU AI Act, must become a fundamental competence for 21st-century educators.

Adopting a pilot implementation and gradual rollout approach allows for learning, adjusting, and refining the solution and processes in a controlled environment before larger-scale expansion, minimizing disruptions and maximizing success possibilities. Finally, implementation does not conclude with deployment. Continuous performance monitoring through relevant KPIs, periodic impact evaluation on school climate and student well-being, proactive technical maintenance, and constant updating of AI models to address language evolution and new threats are permanent tasks. The continuous adaptability of strategies, policies, and technological configurations is what will ensure the system's long-term relevance and effectiveness.

Ultimately, the goal of integrating AI for sensitive content detection is not the technology itself, but the creation of a safer digital school ecosystem where students can learn, explore, and socialize while minimizing risks. To achieve this, educational institutions must embrace this challenge with a strategic vision, an unwavering ethical commitment, and a constant dedication to improvement and adaptation. Only then can they leverage AI's potential to complement, and never supplant, the irreplaceable value of human interaction and professional judgment in the noble task of educating.

13. General Conclusions and Final Recommendations

This report has addressed the complex task of analyzing Artificial Intelligence (AI) models for sensitive content detection in European school environments. Throughout its chapters, it has explored the context of this need, the applicable regulatory framework, the capabilities and limitations of various AI technologies, the psychosocial and pedagogical impacts, and strategies for practical and responsible implementation. This final chapter synthesizes the general conclusions derived from this analysis and offers key recommendations.

13.1 Recapitulation of Challenges and Opportunities

The ubiquity of digital technologies in students' lives has brought both unprecedented educational opportunities and increased exposure to risks stemming from sensitive content, including cyberbullying, violence, misinformation, and radicalization. In this context, artificial intelligence emerges as a tool with the potential to assist in identifying and mitigating these risks, contributing to the creation of safer learning environments.

However, the implementation of AI systems for sensitive content detection is intrinsically linked to significant challenges. These include the protection of minors' privacy and personal data, the risk of perpetuating or amplifying algorithmic biases, the potential psychological impact on students due to the perception of surveillance, and the need for a contextual and nuanced understanding of youth language and interactions that often exceeds current AI capabilities.

The European regulatory framework, primarily through the General Data Protection Regulation (GDPR) and the EU Artificial Intelligence Act (EU AI Act), sets a rigorous standard for the adoption of these technologies. The AI Act, in particular, classifies many AI systems applicable to education, especially those related to student monitoring or evaluation, as "high-risk," imposing strict obligations in terms of transparency, robustness, human oversight, and risk management.

13.2 Synthesis of AI Model Evaluation

The analysis of various AI models – proprietary and open-source, for image, text, and multimodal recognition – reveals a landscape of evolving capabilities, but also inherent limitations:

Proprietary Models: Generally, offer a higher level of support, more developed interfaces, and, in some cases, "ready-to-use" functionalities or more explicit compliance frameworks. However, they usually involve significant licensing costs, less transparency in the internal functioning of algorithms ("black box"), and vendor dependency.

Open Source Models: Provide greater flexibility, customization potential, and code transparency. Nevertheless, they shift the entire burden of development, adaptation, maintenance, and, crucially, responsibility for regulatory and ethical compliance to the implementing organization. This requires high internal technical capacity and considerable investment in obtaining and managing training data.

Adaptation and Multilingualism: Regardless of their origin, most AI models require significant adaptation (fine-tuning) to be effective in detecting the specific and contextual forms of sensitive content relevant to school environments. The acquisition of high-quality, representative, and ethically obtained datasets for this purpose, especially in a multilingual context, constitutes one of the greatest challenges.

Contextual Understanding: AI's ability to interpret context, intent, sarcasm, youth slang, and cultural nuances remains limited. This can lead to errors, both false positives (flagging innocuous content as problematic) and false negatives (missing genuinely harmful content).

Multimodal Models: Although promising due to their ability to analyze different types of data together (e.g., image and text), they also present challenges in terms of training complexity, the interpretability of their decisions, and data management.

13.3 Ethical, Legal, and Pedagogical Imperatives

The implementation of AI for sensitive content detection must be guided by unshakeable ethical and legal principles:

Centrality of Fundamental Rights: The protection of students' dignity, privacy, freedom of expression, and right to non-discrimination must be prioritized.

Indispensable Human Oversight: No critical decision affecting a student (disciplinary, intervention, or well-being) should be made exclusively by automated

means. Qualified human oversight is essential to validate AI alerts, interpret context, and ensure fairness.

Psychological and Pedagogical Impact: Potential negative effects on students' emotional well-being (anxiety, inhibitory effect) and on educators' roles and workloads must be actively considered and mitigated.

Transparency and Explainability: It is fundamental that AI decision-making processes are as transparent and explainable as possible, both for educators managing the system and for students and families affected by its results.

Bias Mitigation: Proactive measures must be implemented to identify, evaluate, and mitigate algorithmic biases that could lead to discriminatory outcomes.

13.4 Towards Responsible Implementation: Key Recommendations

To navigate these challenges and leverage AI's potential responsibly, the following is recommended:

1. Adopt a Strategic and Gradual Approach: Implementation must be the result of careful planning, starting with a clear definition of objectives, a thorough needs analysis, and pilot testing in controlled environments before large-scale deployment.

2. Rigorous Regulatory Compliance: Conducting Data Protection Impact Assessments (DPIA/EIPD) is mandatory. A clear legal basis for data processing must be ensured, data minimization principles and privacy by design must be applied, and all provisions of the EU AI Act must be complied with.

3. Educational Community Participation: Involving management, teachers, technical staff, students, and families from the initial planning and design phases is crucial to foster acceptance, trust, and the suitability of the solution.

4. Comprehensive and Continuous Training: Provide educational staff, especially those who will manage alerts and interventions, with comprehensive training that covers not only the technical handling of the tool but also ethical, legal, pedagogical, and communication aspects. Promote AI literacy throughout the community.

5. Prioritize Preventive Education: Complement technological detection with robust digital citizenship programs, critical media literacy, and socio-emotional learning to empower students.

6. Continuous Monitoring, Evaluation, and Adaptation: Establish clear metrics to evaluate the system's effectiveness and impact. Collect feedback regularly and be

prepared to adapt strategies, policies, and technology as threats and knowledge evolve.

13.5 Future Perspectives and Final Considerations

Artificial intelligence will continue to evolve, presenting new possibilities and challenges for the education sector. It is foreseeable that the capabilities of AI models to understand language, images, and context will improve over time. However, exclusive reliance on technological solutions for complex human safety and well-being issues will remain a limited strategy.

The future of AI in education, especially in sensitive areas like harmful content detection, will depend on society's ability to foster development and implementation that are firmly anchored in humanistic values and respect for fundamental rights. This will require continuous collaboration among educational institutions, technology developers, researchers, lawmakers, and civil society.

The ultimate goal should not be the creation of an infallible surveillance system, but the fostering of school environments where technology serves as a support for safety and well-being, allowing learning, curiosity, and the holistic development of each student to flourish. Artificial intelligence can be an ally in this task, but only if guided with wisdom, prudence, and an unwavering commitment to the ethical principles that must govern education in a democratic and digital society.