LUCET | LUXEMBOURG CENTRE FOR EDUCATIONAL TESTING

uni.lu Université du Luxembourg

From Promise to Practice: A Guide to the Responsible Integration of Generative AI in Large-Scale Educational Assessment under the EU AI Act

LUCET White paper

Lisa Ripoll Y Schmitz & Philipp Sonnleitner

Luxembourg Centre for Educational Testing, University of Luxembourg

March 2026

Executive Summary

This white paper addresses the responsible integration of generative artificial intelligence (particularly large language models, LLMs) into large-scale assessments (LSA) in education against the backdrop of the European Union's Artificial Intelligence Act (AIA). The paper aims to address uncertainties among experts about how to integrate AI tools efficiently without compromising legal, ethical, or psychometric standards. We seek to provide a clear framework (including checklists and examples) to guide compliance efforts during the current transitional phase, when official technical standards for operational implementation of the AIA are still largely lacking. The strategic imperative should be to proactively create compliance structures to maintain the balance between technological progress and the protection of fundamental rights. For this purpose, informed and professional human judgement remains indispensable in strengthening public confidence in the validity and fairness of AI-supported educational assessment methods.

Promise
Integrating LLMs can offer significant efficiency gains in test development and facilitate educational workload through personalized adaptations. The ability of generative artificial intelligence (GenAI) to achieve human-level language quality promises to greatly reduce the burden of resource-intensive tasks.

Peril
The critical dangers of AI-assisted workflows range from technical errors and risks of human overreliance to ethical challenges, such as bias, discrimination and ecological strains. Unreflective use can jeopardize psychometric validity and fairness towards learners.

Principle
AI systems used in education, for instance to assess learning outcomes or to inform learning processes, are categorically classified as 'high-risk'. This forms the fundamental basis for all subsequent requirements. Organizations using AI systems in their workflow (defined as 'deployers') are subject to comprehensive compliance obligations, including the provision of human oversight, maintaining traceability through log documentation and conducting fundamental rights impact assessments (FRIA).

Practice
AI tools should be viewed as a supplement to human expertise ('human-in-the-loop' approach) and should not substitute professional judgement or psychometric validation. Human oversight as final authority is not only a matter of quality assurance but also a legal requirement under the AIA. Therefore, professionals must train in AI literacy to develop specific skills needed for identifying and efficiently mitigating AI-associated risks. Prioritizing open-source and frugal AI systems can reduce environmental costs, strengthen social justice and encourage technological independence.

Perspective
The global significance of the AIA is discussed, as current uncertainty caused by the discrepancy between the regulatory ambitions of the AIA and the actual readiness of stakeholders remains.
The practical feasibility of organizations complying with the regulations is hindered by the fact that some classifications appear static, arbitrary and potentially disproportionate, which could result in higher compliance costs.

Introduction

Educational assessment encompasses a wide spectrum of practices, ranging from classroom-based formative feedback to high-stakes examinations in higher education and national large-scale assessments. Recent advancements of generative artificial intelligence (GenAI) technologies are rapidly reshaping these established assessment practices, promising gains in efficiency, scalability, and innovation. At the same time, their responsible integration into assessment workflows remains considerably uncertain, especially for those directly involved in test development and quality assurance. This issue is particularly pertinent for subject-matter experts (SMEs), who are increasingly likely to encounter or experiment with AI-assisted tools in their professional workflows, without the support of clear operational benchmarks. However, the unverified or unreflective use of GenAI's powerful capabilities may unintentionally blur the lines between originality, authorship, and plagiarism. Beyond questions of academic integrity, professionals are therefore confronted with complex legal, ethical, and organizational challenges. These include, for instance, the lawful categorization of artificially generated material, transparency and documentation duties, data protection requirements, and accountability matters.

The European Union's newly[1] mandated Artificial Intelligence Act (AIA) (Regulation (EU) 2024/1689) aims to provide a harmonized regulatory framework for addressing these risks. However, translating the AIA's complex provisions into concrete, practical guidelines requires a thorough familiarity with its compliance expectations. Even more so, as a severe gap exists between regulatory obligations and the availability of operational compliance infrastructure. While the core obligations for most high-risk AI systems will take effect in August 2026, the technical framework intended to operationalize these requirements is critically incomplete. As the AIA relies heavily on harmonized technical standards to translate its legal provisions into concrete, auditable compliance criteria, the Joint Technical Committee 21 (JTC 21) has been tasked with developing European standards for AI, including harmonized technical standards that directly support the AIA. As of late 2025, however, only 15 of the 45 required standards have been published by the CENELEC (Comité Européen de Normalisation Électrotechnique), with the European Commission yet to issue any. Even under the most optimistic projections, nearly half of the standards (49%) are predicted to remain unavailable when the obligations take effect (Beltrame et al., 2025; Toffaletti, 2025). Thus, paradoxically, compliance will need to be demonstrated against standards that do not yet exist, creating an environment of profound legal and operational uncertainty.

Illustrative use case

In a previous exploratory study, we examined the technical feasibility of using GenAI to create reading comprehension stimuli for large-scale assessments within a hypothetical test-development scenario, focusing on perceived quality and content-related characteristics (Ripoll Y Schmitz & Sonnleitner, 2025). While the findings suggested that AI-generated texts may be on par with human-written materials, testing their actual suitability for educational assessments requires empirical psychometric validation. In order to evaluate the true measurement properties, such as reliability, validity and fairness, the items must ultimately be embedded into real testing scenarios. Doing so, however, presupposes a careful consideration of legal and ethical questions, as well as compliance with the European Union's AI Act.

The present white paper aims to address this gap by providing pragmatic guidance for SMEs and assessment practitioners who are considering the use of GenAI in educational assessment under the EU's Artificial Intelligence Act.

Through discussing the definitions and ambiguities surrounding the AIA, we propose a coherent operational framework for AI-assisted test development that aligns ethical considerations, regulatory requirements, and best-practice recommendations. However, it reflects the state of regulation and research as of early 2026. Given that AI governance is an exceptionally dynamic field, shaped by rapid technological developments, evolving standards,

[1] The European Regulation on Artificial Intelligence (EU AI Act; AIA) came into force on August 1st, 2024.

and ongoing legislative adjustments, the regulatory landscape may continue to develop beyond the scope and timeframe of this report[2]. Due to the involvement of minors and reliance on established psychometric quality criteria, educational large-scale assessment (LSA) represents a more tightly regulated and methodologically demanding assessment context under the AIA. By using LSA in Luxembourg as a detailed reference case, this white paper aims to promote responsible AI integration across the broader educational assessment landscape.

For this purpose, the remainder of the present paper is structured into five parts. After introducing various **promises** of LLM applications in the educational testing realm, we examine the real-world **perils** this approach entails by outlining the associated technical, ethical/legal, educational, and organizational/societal risks. Next, we analyze how these risks translate into concrete obligations for our use case under the **principles** of the EU's AIA. We aim to clarify the legal categorization of educational AI systems, the requirements applicable to AI-generated content, as well as the definitions and associated obligations of (primarily) deployers and providers. In the **practice** section, we propose risk-mitigation strategies and recommend best practices for the responsible integration of GenAI into test development workflows, while maintaining human oversight and compliance with EU regulations. Finally, we aim to give a **perspective** on the feasibility of integrating AI technologies in the future. We briefly discuss ongoing debates surrounding the proportionality and practical ambiguities of the AIA, all while encouraging the sustainable, human-centered use of GenAI in educational assessment.

Promise

In recent years, rapid advancements of GenAI, particularly large language models (LLMs), have drawn substantial attention to their remarkable ability to match, or in some cases even exceed, human-level performance in various language processing tasks (e.g., Ackerman & Balyan, 2023; Brown et al., 2020; OpenAI, 2023; Ripoll Y Schmitz & Sonnleitner, 2025; Tan et al., 2023; Xiao et al., 2023). LLMs are autoregressive models designed to sequentially predict the next token in a sentence, based on the initial input (prompt) and previously generated tokens. Human language inputs are transformed into high-dimensional, contextually rich representations within the GPT (Generative Pre-trained Transformer) architecture, enabling the system to produce coherent, appropriate outputs (OpenAI, 2023; Tan et al., 2024). Owing to their extensive pretraining, these models can be leveraged for a wide range of applications, without requiring task-specific architectural modifications and often with only minimal fine-tuning, if any (Bezirhan & von Davier, 2023). Beyond their technical sophistication, LLMs may also offer broader pedagogical benefits in educational contexts. By providing personalized, real-time responses, LLMs can significantly reduce teachers' workload and assist with routine, time-intensive tasks. These may include drafting instructional and differentiated learning materials, generating (multiple-choice) test questions (e.g., Lee et al., 2023; Lin & Chen, 2024; Tomikawa & Uto, 2024; Wang et al., 2022), providing feedback on student writing (e.g., Pankiewicz & Baker, 2023), and automating the scoring of assignments and essays (e.g., Farrokhnia et al., 2023; Jung et al., 2024; Latif & Zhai, 2024).

Not only do these adaptive capabilities foster the inclusion of diverse learning needs in the classroom, for instance, through translation and accessibility tools, but they could also allow teachers to focus on higher-order instructional objectives and develop deeper pedagogical and mentoring relationships with learners (Yan et al., 2024). From a learner's perspective, LLMs can serve as intelligent tutoring systems, enabling more accessible, personalized learning experiences. They can summarize complex information, support the acquisition of new knowledge, highlight linguistic or conceptual inconsistencies, and suggest personalized strategies for improvement (Kasneci et al., 2023; Farrokhnia et al., 2023). In higher education, LLMs can further support academic workflows by assisting with literature searches, producing summaries, structuring outlines, and identifying unexplored research perspectives (Kasneci et al., 2023). More broadly, this could also foster continuous self-development by challenging professionals to improve their skills and inspiring them to adopt new approaches (Ripoll Y Schmitz & Sonnleitner, 2025).

[2] This report is not intended to provide legal advice and should not be relied upon as a substitute for professional legal consultation. The authors are not acting on behalf of, nor are they endorsed by, any European Union institution or regulatory authority. All recommendations are non-binding and reflect best-practice considerations at the time of writing. The ultimate responsibility for regulatory compliance lies with the respective providers and deployers of AI systems.

In the development of standardized educational tests in particular, which is traditionally a resource-intensive and time-consuming process, LLMs offer the prospect of facilitating the creation of high-quality test items with greater flexibility and efficiency: In an exploratory study (Ripoll Y Schmitz & Sonnleitner, 2025), we investigated the suitability of OpenAI's GPT-4/ ChatGPT for generating German reading comprehension stimuli and examined its potential for integration as a support tool in the test development process for the Luxembourgish educational large-scale assessment *Épreuves standardisées* (ÉpStan).

Through a qualitative SWOT analysis (strengths, weaknesses, opportunities, and threats) with experienced SMEs from the LUCET[3], we ensured that the generated texts adhered to educational curricula and the cognitive task requirements. Using zero-shot and one-shot prompt engineering approaches[4] grounded in a template-based framework (Text Analysis Cognitive Model; TACM; cf. Sayin & Gierl, 2024), we created informative and narrative texts for Luxembourgish fifth-grade students and evaluated them using a mixed-methods design. Independent reviewers then judged the texts' readability, coherence, engagement, and content adequateness as well as their perceived authorship (AI-generated vs. human-written).

The results demonstrated the impressive potential of GenAI to emulate human-written texts in terms of linguistic style and overall quality. AI-generated texts were comparable in quality to their human-authored counterparts, with the majority of reviewers unable to consistently identify their authorship origins. One-shot prompting proved particularly effective for generating informative texts, whereas human authors still retained an advantage for narrative content, which rather involves emotional nuances and contextual subtlety. Zero-shot prompting offered considerable flexibility and creativity but exhibited the most AI-attributed characteristics and therefore still requires human refinement.

These findings suggest that GenAI could be a valuable asset in test development, efficiently providing first drafts for SMEs in areas where sourcing suitable, language-appropriate texts is costly and challenging. At the time of publication, these procedures have not yet been implemented within ÉpStan and should be understood as exploratory considerations for future development.

Using GenAI to provide new thematic content could expand the existing item pool and enable the creation of parallel test versions that adhere to the same criteria and consistency constraints. However, as emphasized in the study and reiterated here, the use of GenAI should be viewed as complementary to, rather than a replacement for, human expertise. Sustained SME involvement throughout all stages of the development process ('human-in-the-loop' or 'augmented intelligence' approach) is essential for ensuring validity, maintaining quality standards, and mitigating biases.

Peril

While LLMs offer compelling opportunities for educational innovation, their deployment in LSA also poses significant risks that must be critically examined before they are integrated into school settings. Even though our example of the Luxembourgish ÉpStan is considered a low-stakes assessment for students, this use case still requires the processing of sensitive data from vulnerable individuals (i.e., children). It necessitates targeted measures to protect minors from the consequences of improper applications. Building on the weaknesses and threats identified in our previous SWOT analysis (Ripoll Y Schmitz & Sonnleitner, 2025) and informed by the broader literature, these risks can be categorized into four intersecting domains: technical, educational, ethical/legal, and organizational/societal. The relevant regulations and risk mitigation measures will be further explained in the *Principle* and *Practice* sections.

1. **Technical Risks:** Inherent model limitations that may compromise reliability and consistency of AI-generated material

Hallucinations and inaccuracy

LLMs can sometimes generate plausible-sounding yet factually incorrect, unverifiable, or fabricated

[3] The Luxembourg Centre for Educational Testing (LUCET) is a research and transfer center at the University of Luxembourg, commissioned to implement, enhance and assure the country's school monitoring program 'Épreuves Standardisées' (ÉpStan), amongst other responsibilities (University of Luxembourg, 2025).

[4] Prompt engineering refers to the deliberate phrasing of instructions that effectively communicate tasks to LLMs (Tan et al., 2024). In zero-shot generation, the model receives only a natural language input without an explicit example. It must therefore rely solely on pre-trained knowledge, often resulting in more flexible and creative responses. In one-shot generation, an example is added to the manual instruction. This enables the model to align its response with the demonstrated structure or stylistic features, as the generated text tends to be similar to the provided reference (Bezirhan & von Davier, 2023).

information (Barberà, 2025; Schuster et al., 2025; Yan et al., 2024). These so-called hallucinations typically occur due to gaps or distortions in the training material, or as a result of the inherent complexity of language-generation processes (Barberà, 2025; Yan et al., 2024). Hallucinated inaccuracies can pose substantial risks in education and assessment, as they can mislead learning processes, generate incorrect feedback, or provide false information, thereby potentially undermining trust and reliability in the tool (Yan et al., 2024). From a regulatory perspective, hallucinations are also considered a critical risk factor under the AIA, as the European AI law explicitly requires error prevention, quality assurance, and human oversight (Schuster et al., 2025).

Lack of transparency and explainability

The opacity of LLMs, also known as their 'black-box' functioning, may prevent users (SMEs, teachers, and learners) from understanding how GenAI actually operates, making it difficult to get a feeling for its potential limitations and the origins of errors (Barberà, 2025; Schuster et al., 2025; Yan et al., 2024). This challenges not only transparency and fairness demands, but also the key objectives of trustworthiness and human oversight set out in the AIA, as a certain level of transparency and AI literacy is required to comprehend and mitigate potential risks.

Lack of robustness and recency

Despite their complexity, LLMs can exhibit unexpected brittleness when performing relatively simple tasks (Kasneci et al., 2023; Yan et al., 2024). They are susceptible to adversarial attacks (i.e., purposefully manipulated inputs) that can compromise their behavior (Barberà, 2025; Yan et al., 2024). Attackers could also introduce incorrect training data into the training dataset ('data poisoning'), causing the AI system to learn undesirable information and thereby compromising its integrity, robustness, and overall security (European Commission, High-level Expert Group on Artificial Intelligence [HLEG AI], 2020). Additionally, models based on static training data are limited in their ability to provide current or complete information, as their knowledge is confined to a specific point in time ('knowledge cut-off') (Barberà, 2025; Schuster et al., 2025).

2. **Ethical and Legal Risks:** Requirements of EU law and fundamental rights

Data protection, privacy (GDPR), and unlawful re-purposing

AI-supported systems used for performance evaluation or to generate personalized learning recommendations inevitably access sensitive information about minors. This may include academic records and behavioral or health-related data, thereby posing considerable privacy and ethical risks (Barberà, 2025; Kasneci et al., 2023). If adequate security measures are not implemented, AI tools operating via third-party services or external cloud infrastructures can increase the likelihood of data leaks and unauthorized access. Even seemingly anonymized training or test data may be traced back to individuals, thus introducing the risk of so-called re-identification (Barberà, 2025; Yan et al., 2024). Another danger pertains to personal data uploaded in the context of user queries (inputs and outputs), if it is being used for retraining purposes different from those originally intended, without having obtained the necessary formal consent (Barberà, 2025). These issues underscore the pivotal roles of data protection regulations, clear consent strategies, and the definition of responsibilities in the event of malfunctions along the AI value chain, especially for applications involving minors.

Copyright and plagiarism

This challenge is particularly pertinent because LLMs trained on large amounts of textual data may reproduce, or resemble, copyrighted materials in their outputs, posing legal risks of plagiarism and intellectual property infringements for both developers and users (Barberà, 2025; Kasneci et al., 2023; Ripoll Y Schmitz & Sonnleitner, 2025). If LLMs are used increasingly in academia for their promising capabilities without considering the risk of copying others' work, this could lead to a widespread 'democratization of plagiarism' and threaten academic integrity (Farrokhnia et al., 2023).

Bias and discrimination

GenAI models can replicate or amplify biases present in their training data, relating to social, historical, or structural factors and resulting in unfair or discriminatory outcomes (Barberà, 2025; Kasneci et al., 2023; Schuster et al., 2025; Yan et al., 2024). As the quality of the output directly depends on the quality of its inputs (training data and user input), the risk of perpetuating biases is known as the 'garbage in, garbage out' principle (Farrokhnia et al., 2023). LLMs are also highly sensitive to how inputs

are formulated, with minor variations in prompt structure potentially leading to entirely different outputs (Barberà, 2025). Accordingly, biased or vague language in prompts, formulations that suggest a specific answer, or emotionally charged inputs could compromise the output's objectivity (Bulut et al., 2024). The lack of transparency inherent to most GenAI models makes it difficult to systematically identify and address such biases, posing a direct risk to the ethical principle of beneficence (Yan et al., 2024). This issue of 'algorithmic bias' is particularly pertinent in the context of education, where biased recommendations or inaccurate assessments can disproportionately disadvantage certain student groups (Bulut et al., 2024). Continuous monitoring or automated analysis of behavior could lead to discriminatory or stigmatizing profiling. Therefore, the EU legal framework expressly prohibits the use of such AI systems in educational and vocational contexts that aim to recognize or infer emotions, social behavior, or personality traits, as this poses a significant risk of discriminatory effects (AIA, Article 5).

3. Educational and Cognitive Risks: Distortions in learning and assessment processes

Overreliance

The admirable ability of LLMs to generate high-quality, human-like responses and recent advancements in their deliberate reasoning capabilities (OpenAI, 2024) may also challenge the validity of conventional learning assessment methods (e.g., essays). It will become increasingly difficult for teachers to distinguish genuine learner work from AI-generated text (Kasneci et al., 2023; Yan et al., 2024). The ease with which information can now be obtained could also lead to increased 'laziness' among learners (Kasneci et al., 2023). Ultimately, this could have a negative impact on agency, problem-solving, and creativity, making it crucial for learners, teachers, but also SMEs to recognize the potential pitfalls of excessively depending on such models (AIA Article 14(4); Barberà, 2025; Kasneci et al., 2023; Yan et al., 2024). Overreliance on GenAI, therefore, entails the risk of developing a dependency that hinders innovation and original thinking, leading to a decline in users' higher-order cognitive skills (Yan et al., 2024; Zhai et al., 2024). Moreover, if SMEs rely on LLM-generated outputs in critical contexts, such as test development, without sufficient understanding or oversight, this overreliance could (unintentionally) compromise their autonomy and professional accountability (Barberà, 2025).

Performance-Illusion

This risk is particularly problematic as GenAI can create the illusion of improved performance in learners without them actually developing essential skills, such as self-regulated learning. A 'performance paradox' may arise when performance declines again once AI support is withdrawn (e.g., Darvishi et al., 2023). Performance illusion can also result in an overestimation of one's own understanding, a phenomenon known as the 'fluency bias' (Yan et al., 2024). Even SMEs, or item developers in this context, may be tempted to give a well-worded AI output certain credibility, momentarily forgetting about their responsibility for the ultimate content (Ripoll Y Schmitz & Sonnleitner, 2025).

Linguistic and cultural limitations, and multilingualism

Most AI algorithms are still primarily calibrated using data from English-speaking Western countries that is subsequently translated into other languages. Since GenAI platforms typically require around two million words to effectively integrate a language (The Government of the Grand Duchy of Luxembourg, 2025), the scarcity or underrepresentation of high-quality linguistic training data for low-resource or less common languages, such as Luxembourgish, can compromise their accuracy. This may result in the use of inappropriate vocabulary, cultural biases, social stereotypes, and discrimination against certain groups, as well as raising fairness concerns in multilingual educational settings (Bulut et al., 2024; Ripoll Y Schmitz & Sonnleitner, 2025). In Luxembourg, the education system's distinct multilingual orientation entails that the language of instruction at school (usually German or French) may differ from the learners' first language, leading to highly heterogeneous proficiency profiles (Ugen et al., 2023). Furthermore, the standard German spoken in countries such as Luxembourg, Austria, and Switzerland constitutes a distinct normative variety that is not fully comparable to standard German in Germany. AI systems may produce formally and stylistically correct outputs, but struggle to understand cultural nuances, connotations, and the specific meanings of language in other standard German contexts. Linguistic pragmatics, in particular, are strongly bound to culture and can thus be difficult for AI to translate. Not only does this lead to problems when transferring pragmatic

aspects across languages, but it also makes this a particularly labor-intensive task for SMEs (Ripoll Y Schmitz & Sonnleitner, 2025).

4. **Organizational and Societal Risks:** Resource-intensive nature of implementing safe and compliant AI systems

AI-Literacy: Lack of technical expertise

Many teachers and educational institutions lack the skills required to utilize LLMs in a pedagogically meaningful, technically correct, and ethically responsible manner (Kasneci et al., Yan et al., 2024). Additionally, a new ethical challenge known as the 'transparency gap' could arise if transparency requirements are comprehensible only to technical experts, thereby excluding educational stakeholders from key decision-making processes (Yan et al., 2024). Even some AI developers may be unaware of potential legal implications, and public authorities may lack the legal expertise needed to effectively regulate AI across sectors (Fedele et al., 2024).

Fair access and digital divide

In the absence of fair resource allocation, unequal access to these powerful technologies could exacerbate existing inequalities ('digital divide') in learning opportunities and educational prospects (Bulut et al., 2024; Kasneci et al., 2023; Yan et al., 2024). Rural and economically disadvantaged communities without reliable internet service may experience reduced employment opportunities and increased social division. Likewise, economic disparities between those who possess the necessary skills and resources to take advantage of (generative) AI and those who do not may perpetuate an 'AI Divide' (Bulut et al., 2024). The financial implications of training, maintenance, and compliance with new regulations, such as the comprehensive EU AIA, can pose a considerable challenge for (educational) institutions with constrained financial resources (Kasneci et al., 2023; European Commission, 2025). The predominance of English-language AI solutions has been shown to perpetuate a bias towards Western, educated, industrialized, wealthy, and democratic societies, with significant implications for the applicability and fairness of these solutions on a global scale (Kasneci et al., 2023; Yan et al., 2024).

Sustainability and resource consumption

Last but not least, as these systems continue to improve in sophistication and processing speed, they are also becoming increasingly energy-intensive. Although AI systems could be used to monitor environmental changes and inform research and political decisions, they themselves require substantial computing resources and cooling infrastructure to maintain operational efficiency. Data centers, which house most large-scale AI deployments, need (destructively mined) rare elements for microchips, produce hazardous electronic waste, and use vast quantities of water to cool electrical servers. Additionally, AI data centers have increased electricity demand, most of which still comes from fossil fuels, resulting in higher carbon dioxide emissions (e.g., UN Environment Programme, 2025). The associated risks to environmental sustainability and water scarcity, as well as the risk of exacerbating the digital divide further, must be given due consideration when discussing the ethical and moral use of AI-supported systems.

Principle

These risks and challenges highlight the urgent need for a clear regulatory and legal framework to foster the responsible development of AI in the EU. The European Union's Artificial Intelligence Act (Regulation (EU) 2024/1689) establishes comprehensive harmonized rules for the development, marketing, and use of AI systems within the EU. At the heart of the regulation is the definition and governance of AI systems through a risk-based approach. Minimal-risk applications (e.g., AI used in spam filters) or open-source systems[5] can be used freely, while limited-risk applications (e.g., chatbots, deepfakes) require transparency declarations. High-risk AI systems used in health, education, or law enforcement are subject to stringent requirements for risk management, data quality, and human oversight. The European legislator is thus responding to the uncertainties associated with the use of AI in sensitive or high-risk areas. Furthermore, certain AI practices are expressly prohibited, such as social scoring, mass surveillance, and the use of AI to predict criminal risk solely through profiling[6]. Key governance measures include the introduction of AI regulatory sandboxes and cooperation

[5] AI models released under free or open-source licenses provide information about parameters, weights, and model architecture, allowing users to freely use, modify and improve sotfware and data, if the original provider of the model is credited (AIA, Recital 102).

[6] Profiling refers to the automated processing of personal data to evaluate aspects of an individual's behaviour, performance or abilities.

with national competent authorities (market surveillance authorities) to monitor compliance, enforce regulations, and promote AI literacy. The implementation of the AIA in the EU member states will be overseen by the AI Office, established within the European Commission (2025). Finally, the regulation sets out specific transparency obligations and liability rules for general-purpose AI (GPAI) models, particularly those posing systemic risk, ensuring responsibility is fairly distributed along the entire AI value chain. Similarly to our study, the principles of the 'human-in-the-loop' approach are reflected throughout the AIA, with human oversight remaining an indispensable component in ensuring technical reliability, ethical accountability, and professional integrity of AI-supported processes.

After its entry into force on 2nd August 2024, the AIA will be implemented gradually (see Table 1), originally with full compliance expected by August 2nd, 2026. However, the European Commission has extended certain regulations by a six-month transition period (until February 2nd, 2027) in its recent amendments to the AIA (European Commission, 2025). The different AI system types with their respective obligations will be detailed further in this section.

Table 1

AIA staggered applicability timeline[7] (TP = Transition Period since 1st August 2024)

AI System Type/ Obligation	Date of Applicability
Prohibited AI Practices & AI Literacy Obligations	2nd February 2025 *TP: 6 months*
Governance Rules & Obligations for GPAI Models	2nd August 2025 *TP: 12 months*
GPAI Code of Practice (Voluntary Compliance Tool)	July 10th 2025 (Published) *TP: 9 months*
Remainder of AIA including most High-Risk AI Systems (Annex III)	2nd August 2026 *TP: 24 months*
High-Risk AI Systems Embedded in Regulated Products (Annex I)	2nd August 2027 *TP: 36 months*

The selected use case of text-based item generation leveraging an LLM is intended to demonstrate how the AIA's comprehensive provisions can be applied to specific promising scenarios and illustrate the legal and ethical requirements for using AI in educational testing. To understand the normative framework for classifying and evaluating AI-generated content, the key terms and definitions of the AIA are explained below.

Definitions under the AIA

General Purpose AI Models vs. Systems

General Purpose AI (GPAI), as the name suggests, exhibits significant generality, meaning it can competently perform a wide variety of tasks across multiple domains, possibly extending beyond its developers' intentions. These models are distinguished by their flexibility and scalability, which are typically facilitated by self-supervised or reinforcement learning with large amounts of training data (Article 3(63); Recital 97; Recital 98). This property enables them to adapt to various applications without requiring substantial modifications or fine-tuning. The AIA (Article 3(66)) and the Future of Life Institute (2024) also describe these as foundation models that underpin other, more specialized applications. A GPAI *Model* refers to the technical model itself, prior to its deployment, integration into an interface, or market placement[8].

Although AI models form the core of AI systems, they do not constitute AI systems per se. It is only when additional components that facilitate user interaction with the underlying model (e.g., interfaces, features, fine-tuned layers) are added that a functional AI system is created. A GPAI *System* is therefore the functional implementation of such a model; for instance, the operational AI system built on top of a GPAI model (Article 3(66)). While GPAI systems can perform a wide range of cognitive tasks, including classification, summarization, and reasoning, the term Generative AI (GenAI) specifically refers to systems designed to produce new content. Large-scale GenAI models (e.g., GPT, Gemini, Mistral) constitute a prominent subset of GPAI models, as they can flexibly generate diverse content, including text, audio, images, video, and code (Recital 99; Dushi, 2024). Consequently, ChatGPT is an example of both a GPAI system and a GenAI system.

GPAI models with and without systemic risk

The AIA establishes a further distinction between GPAI models with and without systemic risk (Article

[7] Full implementation guideline: https://artificialintelligenceact.eu/implementation-timeline/

[8] Research, development, and prototyping activities with AI models carried out exclusively for this purpose prior to being placed on the market do not fall within the scope of the definition.

51; Recital 112), in an effort to strictly regulate disruptions to critical sectors, public health and safety, democratic processes, or the dissemination of illegal or discriminatory content. A GPAI model is classified as being of systemic risk if it exhibits high-impact capabilities, or if it significantly impacts the internal market due to its reach (Article 3(65)). High-impact capabilities are assumed if the cumulative training computation exceeds 10^{25} floating-point operations per second (FLOPs)[9]. Developers (providers) must notify the Commission (AI Unit) immediately and no later than two weeks after this threshold is met. Alternatively, the European Commission, informed by a scientific panel of independent experts, has the authority to designate a GPAI model as posing systemic risk based on its high impact, scale, technical potential, or algorithmic improvements (Future of Life Institute, 2024).

General Purpose AI Code of Practice

The European Commission's original draft of the AIA did not explicitly address regulations for GPAI technologies. However, the Council later emphasized the necessity to incorporate them within the legislative framework. Experts had warned that classifying AI systems as high-risk solely on their intended purpose could result in GPAI systems being largely unregulated, since this approach would focus on applications rather than the underlying foundation models (Madiega, 2023). Therefore, the European Commission has published a General-Purpose AI Code of Practice (*The General-Purpose AI Code of Practice*, 2025) on July 10th, 2025, covering the topics (1) transparency, (2) copyright, and (3) safety and security. GPAI model providers who voluntarily sign this Code of Practice can demonstrate compliance with the AIA, thereby reducing their administrative burden and providing them with greater legal certainty. Adherence to the code is voluntary, but companies that sign it[10] will benefit from a 'presumption of conformity', meaning that EU regulators will assume compliance with the AIA's obligations for GPAI.

Providers vs. Deployers

Under the AIA, the responsibilities for those involved in the AI value chain are primarily determined by their role in developing and using the system. The Act distinguishes between provider, deployer, authorized representative, importer, distributor, and operator (Articles 3(3-8)) in order to determine legal obligations, liabilities, and compliance requirements. Providers are defined as entities (natural or legal person, public authority, agency, or other body) that develop or have developed a GPAI model or an AI system, which they either place on the market or put into service under their own name or trademark, whether for payment or free of charge (Article 3(3)). In other words, providers are the main manufacturers of the AI system that will be sold or deployed, while a downstream provider integrates an existing AI model to develop their own AI systems (Article 3(68)). Further down the AI value chain are the deployers, meaning the end users or operators that utilize an AI system under their authority and within their own professional, non-personal context (Article 3(4); Recital 13). Their obligations concern how the system is used, such as ensuring human oversight, protecting data subjects' rights, and maintaining transparency in interactions with affected individuals. Most organizations fall into the deployer category, integrating existing AI systems into their

Implications for the use case

In the context of LLM-assisted test development, the system-level is classified as a high-risk educational application under Annex III, once the GPAI model is deployed within an educational assessment pipeline (e.g., for generating reading comprehension texts or evaluating student responses). This classification is especially important for large-scale assessments, where the generated materials (such as reading texts) form the basis for evaluating learning outcomes, which could impact students' educational opportunities. Although the provider (OpenAI in the case of ChatGPT) remains responsible for ensuring that the underlying model complies with GPAI-related transparency and safety standards, the deployer must meet certain obligations associated with high-risk systems (see orange box, p. 12).

[9] FLOPs are a subset of real numbers that are usually represented on computers as an integer of fixed precision scaled by an integer exponent of a fixed base (Article 3(67)). In other words, FLOPs describe how computers perform basic calculations with real (decimal) numbers, and counting these small calculation steps shows how much computational effort an AI system requires.

[10] An official and continuously updated overview on companies that have signed the General-Purpose AI Code of Practice can be found on https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

workflows or services rather than developing their own models (Daley-Gage, 2024).

The AIA does not introduce new provisions regarding the copyright of self-uploaded materials (input data or prompt examples); rather, it supplements and builds on existing EU intellectual property law. However, it's important to note that if the uploaded text contains personal data, the processing of this data must comply with the General Data Protection Regulation (GDPR) requirements (Article 2(7); Recital 10). From a data protection perspective, the deployer typically acts as the data controller under the GDPR, as they determine the purposes and means of processing personal data through the LLM interface. Together, these roles form a shared responsibility framework in which providers carry out distinct yet complementary obligations to safeguard privacy, ensure compliance, and maintain the secure and ethical use of AI systems (Barberà, 2025). This notion is also central to the AIA, which encourages both providers and deployers to voluntarily adopt broader ethical and sustainability principles with the intent of fostering responsible innovation (Recital 165). The upcoming practice section will provide a more thorough illustration of best practice recommendations to achieve a respectful, fair, and accessible deployment of AI systems.

High-risk vs. limited-risk

Under the AIA, a GPAI system's intended purpose and context of application determine whether it is classified as high-risk or limited-risk (Schuster et al., 2025). This distinction is crucial, as it dictates the level of regulatory control, the extent of documentation, and the compliance obligations imposed on providers and deployers. While GPAI models are generally categorized as limited-risk systems, their downstream application in specific sectors can elevate them to high-risk status. In other words, a model that is low-risk in a general-purpose chatbot could become high-risk when integrated into sensitive domains such as healthcare, law enforcement, or education (Schuster et al., 2025). In this context, AI systems that may significantly impact a person's educational and professional path, affect their ability to secure a livelihood, perpetuate discriminatory patterns, or violate the right to education are categorized as high-risk (Recital 56).

According to Annex III (Referred to in Article 62) of the AIA, AI systems used in the context of education and vocational training are explicitly listed as high-risk if they are intended to be used:

- To determine access or admission or to assign natural persons to educational and vocational training institutions at all levels;
- To evaluate learning outcomes, including when those outcomes are used to steer the learning process of natural persons in educational and vocational training institutions at all levels;
- For the purpose of assessing the appropriate level of education that an individual will receive or will be able to access, in the context of or within educational and vocational training institutions at all levels;
- For monitoring and detecting prohibited behavior of students during tests in the context of or within educational or vocational training institutions at all levels.

Additionally, the AIA specifies that any system listed under Annex III that performs profiling is automatically considered high-risk (Future of Life Institute, 2024).

The AIA reinforces the central role of deployers in ensuring the safe and responsible implementation of AI systems by explicitly linking their use to a series of concrete obligations (detailed on p. 12).

Obligations for deployers of high-risk AI systems (Article 26)

1. *Proper use and compliance with instructions:*

Deployers must take appropriate technical and organizational measures to ensure that high-risk AI systems are used in accordance with the provider's instructions (Article 26(1)).

2. *Human oversight and AI Literacy:*

Deployers must also assign human oversight to individuals who have the necessary competence, training, authority, and support (Article 26(2)). In that regard, deployers must take appropriate technical and organizational measures (Article 26(3)) to ensure that their staff and other individuals involved in operating AI systems have sufficient AI literacy, which is essential for making informed decisions about the opportunities and risks of AI systems (Articles 3(56); Article 4).

3. *Handling of data:*

Deployers must ensure that the input data is relevant and sufficiently representative for the intended purpose, to the extent of their exercised control over the input data (Article 26(4)). Implementing a Target Operating Model (TOM) and repurposing efforts already invested in data management can help meet this obligation (Daley-Gage, 2024). Importantly, deployers must conduct a Data Protection Impact Assessment (DPIA) (Article 26(9)), if applicable, in accordance with the transparency requirements and information provided by the provider (Article 13).

4. *Monitoring, record-keeping, and reporting requirements:*

In order to regularly monitor the proper functioning of high-risk AI systems (Article 26(5)), such as ensuring the provider's cybersecurity and robustness measures (Article 15), deployers are required to retain the system's automatically generated logs. The retention period must be appropriate for the intended purpose. It must be at least 6 months, unless another period is required by applicable EU or national law, particularly data protection law (Article 26(6)). In the event of any serious incidents or risks, the logs allow for transparently informing the provider and the relevant market surveillance authorities, in order to take the necessary corrective actions (e.g., suspending the system) (Article 26(5)). It can be helpful to look at existing incident management systems, into which the AI elements can be incorporated as part of a holistic governance framework (Daley-Gauge, 2024).

5. *Transparency towards affected parties:*

Deployers of high-risk AI systems that make or assist in decisions must inform affected individuals about the purpose and type of decision, and their right to an explanation of the AI system's role in the decision-making process (Article 26(11)). Similarly, deployers who are employers must inform affected employees and consult employee representatives that they will be subject to the system's use in the workplace before being put into service (Article 26(7)). Deployers are obliged to disclose when individuals are exposed to AI-generated content, meaning that using LLMs for content creation, as in our use case, must be made transparent to those involved in the assessment process. These outputs must be marked in a machine-readable format and detectable as artificially generated or manipulated. This obligation does not apply where the AI system merely performs assistive or editing functions, or where it does not substantially alter the meaning or intent of the input data (Article 50(2)).

6. *Cooperation:*

Deployers must cooperate with the relevant competent authorities on all matters to implement this regulation (Article 26(12)).

Additional obligations apply to deployers that are public authorities or private entities providing public services. First, they must ensure their information, and that of their system, is registered in the EU database (Article 49, Article 71) and must not use unregistered systems. If they are using high-risk systems (such as in education), they must carry out a Fundamental Rights Impact Assessment (FRIA)[11] prior to putting the system into service, which must be updated if there are any changes to the relevant factors (Article 27). The deployer must then inform the respective market surveillance authority of the assessment results.

[11] The FRIA has to contain certain elements, including a description of the processes in which the system is used, the categories of persons likely to be affected (including vulnerable groups), the specific risks of harm and the measures to be taken if these risks materialize (e.g., internal governance and complaint mechanisms) (Barberà, 2025). In December 2025, the Danish Institute for Human Rights and the European Centre for Not-for-Profit Law have created a practical guide (including a template) to help deployers of high-risk AI systems conduct FRIAs under the EU AI Act.

Exceptions and grey areas?

There may be flexibility in how the regulations apply to our use case, allowing for exceptions or legal grey areas that are not exhaustively defined. By way of exception, an AI system listed in Annex III is not considered a high-risk system if it does not pose a significant risk to health, safety, or fundamental rights (Article 6(3)). This may be the case if the AI system does not materially influence decision-making; is intended to perform a narrow procedural task; is intended to perform a preparatory task to an assessment; or is intended to improve the result of a previously completed human activity[12]. Therefore, using an AI system merely to assist with content creation as a preparatory task in the test development procedure, or to improve existing texts written by item developers, could be argued to fall outside the high-risk scope. However, it is crucial that the educational decision itself is not determined or substantially influenced by the AI system. In the case of low-stakes LSA with no immediate consequences for students, but which inform national education policy decisions, this argument could be upheld. Meanwhile, the threshold for classifications as a high-risk system is likely to be reached quickly if the quality of the generated text directly affects the proficiency measurements. The justifications would therefore have to be carefully weighed up.

Furthermore, AI systems or models, as well as their outputs, which are developed and put into service solely for scientific research and development purposes, are exempt from the aforementioned obligations (Article 2(6)). In line with Recital 25, the EU aims to ensure that the AIA does not hinder scientific progress or innovation, provided that the AI system is not subsequently marketed or deployed in a way that affects end users, and that research activities remain ethically and professionally conducted. Accordingly, our AI-assisted content could be included in national large-scale assessment tests to analyze its psychometric properties and item validity, so long as it does not inform or make educational decisions. Another potentially relevant exemption concerns open-source AI systems: according to Article 2(12) of the AIA, open-source AI models are not subject to the Act unless they are placed on the market[13] or used as part of a system that falls under the categories of high risk, prohibition, or transparency obligation. In other words, as long as the model is openly accessible, not commercialized, and not integrated into a regulated downstream application, it is unlikely to trigger any obligations under the AIA. This exclusion may therefore apply to research-driven, non-commercial item-generation tools, particularly those shared openly for academic collaboration.

In summary, the EU AI Act does not prohibit the deployment of GenAI to assist in the development of educational LSA, but it does require documented human oversight and awareness of the risks, along with the respective transparency measures. These compliance obligations are detailed in the orange box (p. 12) as well as in the form of a checklist (Table 2, p. 14) and will be explained further in the following section on best practices.

Practice

Prior to the implementation of the AIA, the European Commission appointed an independent high-level expert group on AI (HLEG AI) to develop the Ethics Guidelines for Trustworthy AI, and, on this basis, a self-assessment guide called the Assessment List for Trustworthy Artificial Intelligence (ALTAI)[14]. This comprehensive checklist sets out seven ethical principles (European Commission, High-level Expert Group on Artificial Intelligence [HLEG AI], 2020):

(1) human agency and oversight
(2) technical robustness and safety
(3) privacy and data governance
(4) transparency
(5) diversity, non-discrimination, and fairness
(6) societal and environmental well-being
(7) accountability

[12] This exception does not apply if the AI system involves profiling of natural persons (Article 6(3)).

[13] Defined as the first making available of an AI system or GPAI model on the Union market (Article 3(9)).

[14] For a more comprehensive and practical insight into how the ALTAI checklist can be used as an interdisciplinary tool to develop trustworthy AI based on a practical example, refer to the academic article by Fedele and colleagues (2024).

Table 2

Checklist: Deployer obligations for AI systems in high-risk contexts according to the AIA

Before implementation	☐ Is the AI system going to be used in a high-risk context in accordance with Annex III AIA?
	If the system is listed under Annex III but only performs a preparatory function that does not pose a significant risk, the 'non-high-risk' classification must be documented prior to commissioning.
	☐ Is there a written agreement with the (GPAI) system provider to supply the necessary information, skills and technical access?
	☐ Are there instructions for use available from the provider?
	For example accuracy, robustness, cyber-security, capabilities, limitations. This information must be obtained to meet own compliance requirements as a deployer.
	☐ Has a Fundamental Rights Impact Assessment (FRIA) been carried out?
	This is required for deployers that are public service bodies or private companies providing public services.
	☐ If applicable, have the results of the FRIA (including risk mitigation measures) been reported to the market surveillance authority?
	☐ Has a Data Protection Impact Assessment (DPIA) been carried out in accordance with the General Data Protection Regulation (Article 35 GDPR)?
	This is particularly important if sensitive data is being processed, as there is a higher risk of a violation of rights.
	☐ If deployers exercise control over the data (e.g., specific training datasets for fine-tuning or RAG sources), has it been ensured that this data is relevant and representative for the intended purpose?
	☐ Was explicit consent from parents or guardians obtained?
	This is relevant, if children's personal (e.g., academic records), or health-related/behavioral data (e.g., indications of mental health conditions or special assistance needs), is processed.
	☐ Have workers' representatives been consulted and affected workers been informed about the integration of the AI system in the workplace?
	☐ Has a quality management system (QMS) been established to ensure compliance?
During implementation	☐ Is the AI system being used in accordance with the instructions and for its intended purpose?
	☐ Is the operation of the AI system continuously monitored in accordance with the instructions for use?
	☐ Is the AI system being supervised by a natural person?
	☐ Do supervisory staff have the necessary expertise, training and authority?
	Prerequisites include sufficient AI literacy to be aware of potential bias and discrimination.
	☐ Have precautions been taken to avoid relying too heavily on AI output (automation bias)?
	This is particularly important when AI is used to provide recommendations or information for human decisions.
	☐ Can operators ignore, override or safely interrupt the system's output?
	☐ In the event of serious incidents or justified suspicion of risk, have the necessary steps been taken to inform the provider and the relevant market authority immediately?
After implementation	☐ Has the content been labelled in a machine-readable format (watermark, metadata) as artificially generated or manipulated, if technically feasible and relevant?
	☐ Are automatically generated logs kept and stored for at least 6 months?
	Logging serves the purpose of traceability.
	☐ Have individuals been informed about how the AI system is used to make or assist decisions affecting them?
	This includes the purpose, type of decision and their right to an explanation.

Although the ALTAI checklist served as a basis for the development of the AIA, it has now been replaced by the AIA's legally binding, authoritative regulations, which establish harmonized rules for the development and use of AI systems. Nevertheless, the checklist retains its value as a self-assessment tool for identifying risks, and the principles set out therein could inform the design of coherent, trustworthy, and human-centered codes of conduct under the AIA.

Attentive readers will notice that the seven principles essentially comprise the mitigation strategies for the aforementioned risks and that the concepts themselves are not new: In a way, they formalize governance dimensions already embedded in psychometric quality assurance.
(1) Human agency and oversight correspond to SME review processes, panel discussions, and item approval procedures.
(2) Technical robustness, LLM stability and safety reflect reliability and item validation analyses.
(3) Privacy and data governance include anonymization measures and secure data storage that are already implemented.
(4) Transparency parallels technical documentation and (construct) validation.
(5) Diversity, non-discrimination, and fairness align with established statistical and qualitative fairness tests, such as bias detection and differential item functioning (DIF) or subgroup analyses.
(6) Societal and environmental well-being relates to educational objectives and informing national policies in a broader sense.
(7) Accountability mirrors long-standing governance structures in test development, including defined roles and audit trails.

Responsible AI integration can therefore be viewed as an extension of established professional norms.

The following section will propose best-practice recommendations for AI-supported test development processes and workflows. They are formulated as commitments that deployers should adopt to preserve professional integrity, informed by recent literature and the previously detailed AIA requirements for high-risk systems in education.

Commitment to (1) Human Oversight and (7) Accountability

Establish multidisciplinary development or peer review systems to identify potential areas of concern.

This commitment is already implemented at LUCET through regular internal reviews and should similarly be adopted for AI-generated content. By bringing together expertise in psychometrics, education, subject-specific knowledge, ethics, and AI development, accountability can be enhanced (Fedele et al., 2024). This interdisciplinary collaboration also helps to ensure that the technical model outputs are meaningfully aligned with educational and assessment standards. Human involvement is also crucial in the post-processing phase, where AI-generated content must be systematically evaluated, refined, and corrected where necessary. This includes identifying and replacing biased, inappropriate, or harmful language, a task usually carried out by SMEs. Similarly, Bulut and colleagues (2024) emphasize that robust human oversight is essential when integrating AI systems into assessment workflows, as it ensures the reliability and defensibility of the resulting test materials. The 'human-in-the-loop' governance mechanism enables human intervention in every decision cycle of the AI system (HLEG AI, 2020), meaning that all AI-generated content is supervised and approved by an expert. This review structure should incorporate formal item quality verification, such as independent plagiarism checks and sensitivity reviews. It ensures that the generated content adheres to psychometric, ethical, and educational standards before it is adopted for use in assessment contexts (e.g., Chuang & Yan, 2025).

Cultivate AI literacy (on a national scale) by providing comprehensive training and professional development opportunities.

To counteract overreliance on GenAI and prevent a performance illusion, it is crucial to critically evaluate AI-generated content, question its plausibility, and verify information against authoritative external sources. In their national framework for integrating AI literacy into schools (KI Kompass), the Luxembourgish Ministry of Education aptly formulated it as a way to learn to think with AI, while preserving the ability to think without it (Ministère de l'Éducation nationale, de l'Enfance et de la Jeunesse, 2026). Therefore, it is important to enhance technical expertise and awareness about GenAI's limitations, biases, and potential errors among all those involved in the educational (assessment) context (Chuang & Yan, 2025; Kasneci et al., 2023; Yan et al., 2024). In application areas such as test-related item or content development, as in the present use case, experts must adopt a critical

examination of AI-generated content, comparing it with valid sources and recognizing typical error patterns, such as hallucinations.

Implement traceability and logging mechanisms and clearly define responsibilities.

To ensure auditability, internal or external auditors must be able to evaluate AI decision-making processes and outcomes using measures of reproducibility, traceability, and explainability. This includes clear protocols for human oversight and auditing (Dumas et al., 2025), as well as written disclaimers that highlight the probabilistic nature of LLMs and their susceptibility to producing inaccurate outputs (Fedele et al., 2024). In complex value chains, such as those involving GPAI, the somewhat entangled responsibilities between providers and deployers should be clearly delineated. Since deployers often rely on third-party models, these responsibilities (e.g., between the data controller and processor under the GDPR) should be precisely defined in written agreements (Barberà, 2025).

Commitment to (4) transparency and explainability

Effectively and transparently communicate the methods, objectives, and outcomes of the AI tool used.

In their input, practitioners should define precise objectives, employ an appropriate tone, provide contextual information, indicate the expected response format, and supply examples or references (Chuang & Yan, 2025) (see one-shot or few-shot prompting; Ripoll Y Schmitz & Sonnleitner, 2025). A thorough framework, such as the TACM approach demonstrated in our pilot study, can further clarify the system's purpose, detail the most effective prompt engineering strategies, and illustrate corresponding outputs. Such traceable documentation supports consistent and responsible use and could increase transparency around the otherwise opaque LLM mechanisms. Furthermore, the use of AI in the testing process must be disclosed and explained to participants (e.g., students) and stakeholders (e.g., parents) in a comprehensible, meaningful manner beforehand (Dumas et al., 2025). This is especially important if the AI system plays a role in a decision-making process that has legal consequences (see AIA).

Consider using eXplainable AI (XAI) methods, if applicable.

Since the black-box nature of these models makes it difficult to enforce the individual's right to an explanation, eXplainable AI (XAI) methods could be considered (Bulut et al., 2024). These transparency mechanisms can make the LLM's decision-making process more comprehensible and easier to identify potential sources of bias (Barberà, 2025). Other knowledge-grounding techniques, such as retrieval-augmented generation (RAG), in which LLMs are 'fed' trusted external information, could improve response accuracy, timeliness, and traceability (Gao et al., 2024; Schuster et al., 2025).

Commitment to (2) technical robustness and safety

Adhere to cybersecurity measures and instructions given by the provider of the AI system.

Most of this category's risk mitigations falling under the deployer's responsibility are already covered by the aforementioned commitments to human oversight and AI literacy training for staff. Others primarily concern the provider side and relate mainly to technical design and security as part of the development process. For instance, this involves so-called fail-safe plans, which developers should test and keep on standby in case interruptions to operations or a switch to alternative procedures become necessary.

Additionally, providers should take appropriate measures against cyber threats, such as data poisoning (manipulation of training data) and adversarial attacks (provoking incorrect decisions through inputs) (AIA, Recitals 75-76). They are encouraged to use AI Regulatory Sandboxes[15], protected and controlled testing environments that enable innovators to develop and test new AI systems while mitigating risks to fundamental rights and security. These sandboxes ensure that innovations are carried out under the close supervision of a competent authority and in compliance with legal requirements (AIA Article 3(55); The Government of the Grand Duchy of Luxembourg, 2025). Deployers are then responsible for maintaining these safety measures by using the system in accordance with the provider's instructions and ensuring that

[15] By August 2026, each EU member is required to set up at least one national AI sandbox (AIA Article 57(1)). In Luxembourg, the data protection commission (CNPD) has already launched a data protection sandbox in May 2024 (The Government of the Grand Duchy of Luxembourg, 2025).

the secured system is not compromised by improper handling (AIA, Recitals 91; Article 26(1)).

Commitment to (5) diversity, non-discrimination, and fairness

Take interdisciplinary measures to prevent misuse of and unfair outcomes from AI tools.

In order to promote diversity, non-discrimination, and fairness, those responsible for deploying the GenAI system must adopt best practices that extend beyond purely technical safeguards. Especially in contexts of high societal relevance, such as educational assessment, it is crucial to ensure that the use of AI tools does not inadvertently reinforce existing social inequalities. In line with the aforementioned AI Act obligations, deployers must conduct Fundamental Rights Impact Assessments (FRIAs) to evaluate the potential adverse effects on fundamental rights and identify risks specific to marginalized or vulnerable groups, including children.
To support this process, deployers can use external analytical tools, such as LLM observatories[16], which compare popular language models with regard to their potential gender, age, ethnicity, and other implicit biases (The Government of the Grand Duchy of Luxembourg, 2025). Alongside strengthened AI literacy, the aforementioned XAI approaches can further support the identification of discriminatory decision patterns by making the model behavior more interpretable (Barberà, 2025; Fedele et al., 2024). Fostering critical reflection can enable SMEs to detect such algorithmic biases rather than viewing AI outputs as neutral or objective.

To the extent possible, ensure training and input data reflect the diversity of society.

Many bias mitigation strategies fall under the provider's responsibility, such as using high-quality, diverse, and non-discriminatory training datasets. However, the quality of an AI system's output is directly dependent on the quality of its input data, known as the 'garbage in – garbage out' principle (Farrokhnia et al., 2023; Dumas et al., 2025). Consequently, if a system is 'fed' flawed, biased, or incomplete data, these shortcomings will inevitably be reflected or even amplified in its outputs or downstream applications (Barberà, 2025; Bulut et al., 2024). The same applies to prompt quality: even highly sophisticated models cannot deliver meaningful results if the input is unclear, poorly formulated, or inaccurate (Barberà, 2025). As deliberately specified in our pilot study, clear communication with the model through carefully crafted prompts and based on a theoretical framework is therefore crucial for mitigating algorithmic bias (Ripoll Y Schmitz & Sonnleitner, 2025).
Since the AIA only requires GPAI providers to supply a high-level summary of their training data (Article 53(1); Article 11(1) and Annex IV for high-risk systems), verifying if it contains copyright infringements or systematic distortions becomes difficult for deployers. They can, however, ensure that input data, prompts, and fine-tuning datasets represent the intended target population, thereby mitigating downstream bias. A commendable example with regard to copyright protection is Adobe Firefly's intellectual property indemnification policy, which guarantees that its AI system is trained using only licensed Adobe Stock, public domain, and open-licensed content (*Our approach to generative AI with Adobe Firefly,* 2026). In fact, Adobe will defend its users and even pay potential damages in the event of copyright, trademark, or privacy infringement claims. Generated images used within tests from LUCET are created on this basis. Such open-source approaches have the potential to reduce costs for deployers, mitigate plagiarism risks, and alleviate certain regulatory and compliance burdens, while increasing transparency and adaptability in educational contexts.

These considerations also directly relate to the linguistic dimension of fairness. To ensure equitable access, GenAI systems must be able to function across multiple languages and standard varieties. In Luxembourg's multilingual context, integrating Luxembourgish into the equation represents a technical, cultural, and political commitment linked to the country's digital sovereignty and linguistic justice. In this regard, sustainable AI development requires shared, high-quality public datasets and tools that serve the long-term interests of entire linguistic communities. Beyond accessibility, AI systems should contribute to the preservation, evolution, and enrichment of cultural identity through language rather than enforcing dominant linguistic norms (The Government of the Grand Duchy of Luxembourg, 2025).

[16] LIST (Luxembourg Institute of Science and Technology) LLM observatory: https://ai-sandbox.list.lu/llm-leaderboard/

Select AI systems that have documented fairness metrics and empirical evidence of their performance.

Deployers should not assume that GenAI models generalize across cultural contexts. Instead, they must empirically and qualitatively test fairness across different cultural and linguistic groups to avoid disadvantaging specific learner populations (Dumas et al., 2025). If possible, the statistical accuracy of the model for its intended purpose should be evaluated (Barberà, 2025), although assigning this responsibility between the provider and the deployer is more difficult.

Continuously review AI-generated content before using it in critical applications.

Although this best practice recommendation is already covered in the commitment to human oversight and may appear redundant, it is important to reiterate this point: Due to the mentioned limitations in controlling the input data for non-discrimination and fairness, SMEs must continuously monitor performance and have the final decision-making authority when reviewing AI-generated materials, particularly those used for evaluation or assessment (Kasneci et al., 2023; Ripoll Y Schmitz & Sonnleitner, 2025).

Commitment to (3) privacy and data governance

Consult a data protection officer (DPO) in the procurement and planning phase.

Involving a data privacy expert as a strategic consultant early on fosters "privacy by design" (Fedele et al., 2024). It facilitates the voluntary, yet recommended, data protection impact assessment (DPIA; Article 34 of the GDPR), as well as the mandatory fundamental rights impact assessment (FRIA; see AIA). Designating a data protection officer can also be beneficial for obtaining parental consent for minors and overseeing individuals' rights to withdraw consent or to be forgotten (right to erasure) (Fedele et al., 2024; HLEG AI, 2020).

Minimize input data and maintain control over its storage.

Establish written agreements that providers will only process data for specific purposes, that no unauthorized or impermissible disclosure or transfer of data to third countries will occur, and that the data will not be used for the provider's own purposes (e.g., training the underlying model) (Barberà, 2025). No personal or sensitive data, nor protected materials, should be uploaded in AI tools (The Government of the Grand Duchy of Luxembourg, 2025), which is why anonymization or pseudonymization techniques should be applied. Implement multi-factor authentication (MFA) or role-based access controls (RBAC) to limit system access and encrypt stored data (Barberà, 2025; *The General-Purpose AI Code of Practice*, 2025). Best practices for complying with AIA requirements also include creating an incident response plan for potential data breaches that specifies who is notified, how they are notified, and the corrective measures to be taken.

Commitment to (6) societal and environmental well-being

Take into account environmental, social, and societal perspectives.

To support societal and environmental well-being in the deployment of GenAI systems, it is essential that deployers adopt best practices that address both equitable access to AI technologies and their environmental sustainability. In educational settings, proper technological infrastructure and equitable resource allocation help all learners benefit from AI-supported tools without being excluded due to socio-economic, gender-related, or other structural disparities. As Bulut and colleagues (2024) emphasize, advancements in AI should be inclusive of all members of society, ensuring that no group is systematically privileged over another. In a democratic society, inclusive access to AI should not be a technical issue alone but a societal one, ensuring that all members have a voice in the development of the services these systems provide. To address this, deployers must proactively seek ways to help mitigate the digital divide.

Prioritize open-source and energy-efficient (frugal) AI models.

From a practical standpoint, a more equitable access can be supported through using open-source AI models released under suitable licenses for educational applications (The Government of the Grand Duchy of Luxembourg, 2025). Models such as Llama 3[17] or Mistral[18] have the potential to lower financial barriers, foster collaborative research, and reduce dependency on proprietary systems. This, in

[17] https://www.llama.com/models/llama-3/

[18] https://mistral.ai

turn, could contribute to a more competitive and inclusive AI ecosystem (Bulut et al., 2024).

In addition to promoting social inclusion, deployers are also responsible for addressing the environmental impacts of AI systems, especially given the significant computational demands of training and deploying large models. Sustainable AI usage requires energy-efficient hardware, shared infrastructure, and continued research into reducing the computational costs of training and maintenance (Kasneci et al., 2023). Favoring "frugal AI practices" in Luxembourg (The Government of the Grand Duchy of Luxembourg, 2025), which require less computational, memory, and energy consumption per query, therefore also means distancing oneself from highly resource-intensive GPAI models. As it is our shared ethical responsibility as global citizens, no industry, including the AI sector, should be exempt from addressing climate change and its environmental consequences (Bulut et al., 2024). Hence, deployers should also ensure that they select an AI system powered by renewable energy sources, where feasible. This includes choosing cloud computing providers that rely on renewable energy or pursue carbon-neutral strategies (Bulut et al., 2024). In its AI Strategy, Luxembourg aims to adopt several environmental cost mitigation strategies, including renewable-powered data centers, optimized algorithms, sustainable hardware life-cycle management, and increased transparency regarding the carbon footprint of AI-based services (The Government of the Grand Duchy of Luxembourg, 2025).

Ultimately, securing societal and environmental well-being requires that deployers treat access, inclusion, and sustainability as interdependent dimensions of responsible use. Active measures to prevent the widening of digital inequalities regarding access to AI technologies are essential for vulnerable groups at risk of further marginalization (Bulut et al., 2024). Embedding social justice and environmental sustainability into deployment decisions helps develop AI systems that prioritize long-term societal benefits over short-term technological advancements.

Perspective

While the EU's AIA is widely regarded as a landmark regulatory framework with global relevance, in practice, it has not resonated much with the general public and key stakeholders. Consistent with our own observations in academic and applied assessment contexts, it appears that the AIA has not yet meaningfully reached many of the actors who are expected to implement it in professional practice.

A representative survey conducted by Forsa on behalf of TÜV Germany in 2024, involving 1,001 respondents, found that 72% had never heard of the AIA (Scheurenbrand, 2024). Additionally, there was a noticeable lack of confidence in government AI policy: 45% of respondents expressed some degree of uncertainty, and 23% stated a complete lack of confidence. Although awareness may have improved since the AIA's formal adoption, the continued absence of concrete, application-oriented guidance suggests that uncertainty persists, particularly among non-specialist audiences. Similarly, a Deloitte survey of 500 private-sector AI decision-makers (who were already familiar with the AIA) found mixed opinions regarding preparedness and perception of the regulations. Only around one-third of respondents (35.7%) felt well prepared to implement the AIA's requirements, while more than half (53.8%) reported not having taken any preparatory measures. Notably, despite the AIA's aspirations to increase legal certainty, nearly half of the respondents (47.7%) viewed it primarily as an obstacle to AI-based applications in their company, compared to only 24.1% who considered it a facilitating factor (Becker & Contzen, 2024).

At the same time, the AIA was intentionally designed with a broad jurisdictional scope to leverage what is commonly referred to as the 'Brussels effect'. This strategy builds on the EU's regulatory capacity and market size to shape global standards, as companies often find it economically advantageous to align their global operations with EU rules to access hundreds of millions of consumers (Almada & Radu, 2024). Importantly, the AIA applies not only to public and private entities established within the EU, but also to providers and deployers outside the Union whose AI systems are placed on the EU market or affect individuals within it. Due to this extra-territorial reach and geopolitical implications, the AIA has faced mounting pressure from the United States, where the regulation has been widely framed as a strategic constraint on major US Big Tech companies. Most prominently, the Computer and Communications Industry Association (CCIA), whose members include Apple, Meta, and Amazon, has publicly campaigned for a relaxation of the AIA and the EU's broader digital rulebook (Besliu, 2025). These industry efforts coincide with ongoing

political discussions between EU officials and the current Trump Administration, reflecting concerns that the AIA could disadvantage US firms in global AI competition.

Against this backdrop, and in response to both external lobbying and internal concerns from member states, the European Commission very recently published the so-called 'Digital Omnibus on AI'[19] in November 2025. The stated objective of this comprehensive reform package is to strengthen Europe's competitiveness by simplifying certain aspects of the AIA, though the planned changes primarily focus on governance structures and implementation timelines (Besliu, 2025). Meanwhile, core obligations for high-risk AI systems remain largely unchanged. For deployers of such systems, the requirement to ensure AI literacy among staff will continue to apply, while responsibility for AI literacy in lower-risk contexts is expected to fall under public authorities. From a practical standpoint, AI literacy remains a strategic priority for ensuring reliable human oversight and responsible AI integration, regardless of formal risk classification.

In that regard, critics have argued that the risk-based classification system is arbitrary because it lacks a transparent, systematic methodology (Edwards, 2022). Rather than deriving risk categories from a structured assessment of potential harms, the AIA relies on a predefined, static list of high-risk application areas, for instance, educational and vocational training (see blue box). This challenge is further amplified when GPAI models with systemic risks are integrated into high-risk contexts. As the boundaries between responsibilities get blurred, effective compliance now depends on close cooperation between GPAI providers and deployers. In working through the AIA for this report, we repeatedly encountered difficulties in clearly distinguishing these responsibilities between providers and deployers. For instance, deployers who want to fine-tune their AI system may suddenly assume full provider-level responsibilities if they undertake "substantial modifications" to the underlying model (Article 25(1)). Meanwhile, there are no specified quantitative criteria or concrete thresholds for when a modification becomes substantial (e.g., the number of parameters or the percentage of retraining). In such cases, deployers are responsible for conformity assessments, even though they did not originate the underlying model and may lack access to training or test data used in its development.

Implications for the use case

Under the AIA, AI systems used in education and vocational training are classified as high-risk (Annex III), regardless of their actual function in this context. For our use case, using GenAI to support content creation for assessment items automatically (and disproportionately) triggers high-risk obligations. Since the AIA currently does not differentiate between functional categories, the fact that the impact on students would only be indirect (through exposure to AI-assisted texts), not used for making decisions about learners or scoring responses, and mediated through thorough human oversight, is irrelevant. Additionally, if a GPAI model with systemic risk is used as such an assistance tool, responsibilities may overlap or become unclear.

However, the high-risk classification also carries economic implications. Compliance requires substantial investments in technical documentation, quality management systems, and cybersecurity measures, which disproportionately affect startups and small to medium-sized enterprises with limited budgets (Bignami et al., 2025). An early impact assessment from 2021 estimated certification costs between €16,800 and €23,000, potentially reducing profits by around 40% for a European small or medium-sized enterprise with €10 million in turnover (e.g., Mueller, 2021). The same holds true in principle for national or private testing agencies. Since the AIA entered into force only in August 2024 and many high-risk obligations will not apply until 2026-2027, there is currently no comprehensive empirical evidence to confirm or update these projections. Nevertheless, concerns persist that innovative AI firms may still find it more advantageous to set up operations outside the EU. This could contribute to a gradual 'brain-drain' towards markets with less restrictive or more innovation-friendly regulations (Bignami et al., 2025). Larger, well-funded companies are better positioned to absorb compliance costs and maintain dedicated legal and technical teams. The AIA does attempt to mitigate these effects by permitting small and medium-sized enterprises to submit simplified technical

[19] https://digital-strategy.ec.europa.eu/en/library/digital-omnibus-ai-regulation-proposal

documentation for high-risk systems (AIA, Article 11(1)), and the Commission has announced plans to extend similar relief to mid-cap companies through dedicated templates that reduce administrative burden (European Commission, 2025). Priority access to AI regulatory sandboxes is also intended to allow smaller actors to develop, test, and validate their systems free of charge and without fear of immediate sanctions. However, the reality is that such sandboxes are still largely unavailable across member states, as the European Commission has officially acknowledged (Toffaletti, 2025; European Commission, 2025). The EU-level AI regulatory sandbox set up by the official AI office is expected to be operational as of 2028, hence only after the regulations apply (European Commission, 2025).

This delayed development of the technical and institutional infrastructure required to operationalize the AIA is one of the most pressing challenges. Due to slow progress in developing harmonized technical standards for high-risk AI systems, several key obligations have been postponed, giving companies more time to implement compliance measures. According to the Digital Omnibus, high-risk rules will now only take effect once the European Commission confirms that adequate support measures, including guidance documents, are in place. After this decision is made, the rules will apply six months later, but not later than December 2nd, 2027 (European Commission, 2025). In the meantime, compliance largely relies on self-assessment, given the well-documented shortage of AI experts capable of conducting complex risk analyses, evaluating social impact, and determining effective governance (e.g., Lee et al., 2024; Mueller, 2021).

Concluding Remarks

We argue that the AIA will only fulfil its promise of trustworthy, human-centric AI if it is accompanied by actionable guidance that enables practitioners to translate abstract regulatory principles into responsible, real-world practice. Our efforts in deciphering the AIA have revealed a discrepancy between its regulatory ambition and the EU's preparedness to incorporate these requirements into practical workflows. For assessment practitioners and SMEs, the question is not whether to follow the AIA, but rather how to apply it proportionately in the absence of relevant guidelines. Even with tools such as official compliance checkers or AI help desks, many operational questions remain, especially where professional standards, ethical responsibilities, and legal expectations intersect. Using large-scale assessment as a reference case, we wanted to illustrate how regulatory principles can align with quality criteria such as validity, reliability, fairness, and human oversight. Many of the considerations discussed, particularly those concerning transparency, documentation, human-in-the-loop processes, and AI literacy, are relevant to educational assessment in general. The guidelines and explanations presented in this white paper are intended to address the lack of official implementation frameworks to navigate the rapidly evolving regulatory landscape. Grounded in existing psychometric and ethical standards, we aim to offer best-practice interpretations and support informed, reflective professional judgment in real assessment workflows, as well as help practitioners reflect on the responsible use of GenAI. Ultimately, while GenAI can support preliminary phases of item development, it cannot and should not replace expert review or psychometric validation. Responsible AI integration in assessment is not a one-time compliance exercise but an ongoing institutional commitment to validity, fairness, and public trust.

Reference List

Ackerman, R., & Balyan, R. (2023). *Automatic Multilingual question generation for health data using LLMs*. Springer Nature Singapore.

Almada, M., & Radu, A. (2024). The Brussels Side-Effect: How the AI Act can reduce the global reach of EU policy. *German Law Journal*, *25*(4), 646–663. https://doi.org/10.1017/glj.2023.108

Barberá, I. (2025). AI Privacy Risks & Mitigations—Large Language Models (LLMs). *European Data Protection Board.* https://www.edpb.europa.eu/system/files/2025-04/ai-privacy-risks-and-mitigations-in-llms.pdf

Becker, S. J., & Contzen, T. (2024). *AI ACT Survey 2024: Survey on the Impact of the AI Act*. Deloitte. https://www2.deloitte.com/content/dam/Deloitte/dl/Documents/legal/Deloitte%20AI%20Act%20Survey_english.pdf

Beltrame, B., Grasso, A., & Schuck, F. (2025). A state of play of AI Act enforcement: shortcomings in standards, authorities, and sandboxes. *European Standardization*, 1–6. https://www.digitalsme.eu/digital/uploads/DIGITAL-SME-Report-A-state-of-play-of-the-AI-Act-enforcement.pdf

Besliu, R. (2025, November 13). *What's driving the EU's AI Act Shake-Up?* Tech Policy Press. https://www.techpolicy.press/whats-driving-the-eus-ai-act-shakeup/

Bezirhan, U., & Von Davier, M. (2023). Automated reading passage generation with OpenAI's large language model. *Computers and Education. Artificial Intelligence, 5*, Article 100161. https://doi.org/10.1016/j.caeai.2023.100161

Bignami, E. G., Russo, M., Semeraro, F., & Bellini, V. (2025). Balancing Innovation and Control: The European Union AI Act in an Era of global uncertainty. *JMIR AI*, *4*, e75527. https://doi.org/10.2196/75527

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agar- wal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., & Amodei, D. (2020). Language models are few-shot learners. *arXiv*. https://doi.org/10.48550/arxiv.2005.14165

Bulut, O., Beiting-Parrish, M., Casabianca, J. M., Slater, S. C., Jiao, H., Song, D., Ormerod, C. M., Fabiyi, D. G., Ivan, R., Walsh, C., Rios, O., Wilson, J., Yildirim-Erbasli, S. N., Wongvorachan, T., Liu, J. X., Tan, B., & Morilova, P. (2024). The rise of artificial intelligence in educational measurement: opportunities and ethical challenges. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.18900

Chuang, P., & Yan, X. (2025). Language assessment in the era of generative artificial intelligence: Opportunities, challenges, and future directions. *System*, *134*, 103846. https://doi.org/10.1016/j.system.2025.103846

Daley-Gage, B. (2024, November 14). The EU AI Act: What are the obligations for deployers? *DataGuard*. https://www.dataguard.com/blog/the-eu-ai-act-obligations-for-deployer/

Darvishi, A., Khosravi, H., Sadiq, S., Gašević, D. & Siemens, G. (2023). Impact of AI assistance on student agency. *Computers & Education*, *210*, 104967. https://doi.org/10.1016/j.compedu.2023.104967

Dumas, D., Greiff, S., & Wetzel, E. (2025). Ten Guidelines for scoring psychological Assessments using Artificial Intelligence. *European Journal of Psychological Assessment*, *41*(3), 169–173. https://doi.org/10.1027/1015-5759/a000904

Dushi, D. (2024, September 5). *How is ChatGPT regulated by the EU AI Act: Reflections on higher education*. Global Campus of Human Rights. https://www.gchumanrights.org/preparedness/how-is-chatgpt-regulated-by-the-eu-ai-act-reflections-on-higher-education/

Edwards, L. (2022, March 31). *Expert opinion: Regulating AI in Europe: Four problems and four solutions*. Ada Lovelace Institute. https://www.adalovelaceinstitute.org/report/regulating-ai-in-europe/

European Commission, High-Level Expert Group on Artificial Intelligence [HLEG AI]. (2020). *The assessment list for trustworthy artificial intelligence (ALTAI)*. Publications Office of the European Union. https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai

European Commission. (2025). *Proposal for a regulation of the European Parliament and of the Council amending Regulation (EU) 2021/694 with regard to AI governance and implementation support measures (Digital Omnibus on AI)*. *EUR-Lex*. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52025PC0836

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2023). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International*, *61*(3), 460–474. https://doi.org/10.1080/14703297.2023.2195846

Fedele, A., Punzi, C., & Tramacere, S. (2024). The ALTAI checklist as a tool to assess ethical and legal implications for a trustworthy AI development in education. *Computer Law & Security Review*, *53*, 105986. https://doi.org/10.1016/j.clsr.2024.105986

Future of Life Institute. (2024, February 27). *High-level summary of the AI Act | EU Artificial Intelligence Act*. https://artificialintelligenceact.eu/high-level-summary/

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). Retrieval-Augmented Generation for Large Language Models: A survey. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2312.10997

Jung, J. Y., Tyack, L., & Von Davier, M. (2024). Combining machine translation and automated scoring in international large-scale assessments. *Large-Scale Assessments in Education*. https://doi.org/10.1186/s40536-024-00199-7

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., . . . Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education. Artificial Intelligence, 6*, Article 100210. https://doi.org/10.1016/j.caeai.2024.100210

Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J., & Kim, H. (2023). Few-shot is enough: Exploring ChatGPT prompt engineering method for automatic question generation in english education.

Education and Information Technologies. https://doi.org/10.1007/s10639-023-12249-8

Lee, D., Todorova, C., & Dehghani, A. (2024). Ethical Risks and Future Direction in Building Trust for Large Language Models Application under the EU AI Act. *HCAIep '24: Proceedings of the 2024 Conference on Human Centred Artificial Intelligence - Education and Practice*, 41–46. https://doi.org/10.1145/3701268.3701272

Lin, Z., & Chen, H. (2024). Investigating the capability of ChatGPT for generating multiple-choice reading comprehension items. *System*. https://doi.org/10.1016/j.system.2024.103344

Madiega, T. (2023). General-purpose artificial intelligence. In *EPRS | European Parliamentary Research Service* (Report PE 745.708). https://www.europarl.europa.eu/RegData/etudes/ATAG/2023/745708/EPRS_ATA(2023)745708_EN.pdf

Ministère de l'Éducation nationale, de l'Enfance et de la Jeunesse. (2026). Strategischer Rahmen zum Einsatz von Künstlicher Intelligenz in der Schule. In *ki-kompass.lu*. https://ki-kompass.lu/wp-content/uploads/2026/02/Digital_DE_Strategischer-Rahmen-zum-Einsatz-von-Kuenstlicher-Intelligenz-in-der-Schule-_-V5.pdf

Mueller, B. (2021). How Much Will the Artificial Intelligence Act Cost Europe? In *Center for Data Innovation*. https://www2.datainnovation.org/2021-aia-costs.pdf

OpenAI. (2023). GPT-4 Technical Report. In *arXiv: Vol. 2303.08774v6* [Technical report]. https://arxiv.org/pdf/2303.08774.pdf

OpenAI. (2024). *Hello GPT-4o: We're announcing GPT-4o, our new flagship model that can reason across audio, vision, and text in real time.* openai.com. https://openai.com/index/hello-gpt-4o/

Our approach to generative AI with Adobe Firefly. (2026). Adobe. https://www.adobe.com/hk_en/ai/overview/firefly/gen-ai-approach.html

Pankiewicz, M., & Baker, R. S. (2023). Large Language Models (GPT) for automating feedback on programming assignments. *arXiv*. https://doi.org/10.48550/arxiv.2307.00150

Regulation (EU) 2024/1689 (Artificial Intelligence Act), Official Journal L 2024/1689 (2024). https://eurlex.europa.eu/eli/reg/2024/1689/oj/eng

Ripoll Y Schmitz, L. M., & Sonnleitner, P. (2025). Evaluating AI-generated vs. human-written reading comprehension passages: an expert SWOT analysis and comparative study for an educational large-scale assessment. *Large-scale Assessments in Education*, *13*(1). https://doi.org/10.1186/s40536-025-00255-w

Sayin, A., & Gierl, M. (2024). Using OpenAI GPT to generate reading comprehension items. *Educational Measurement Issues and Practice*, *43*(1), 5–18. https://doi.org/10.1111/emip.12590

Scheurenbrand, K. M. (2024, November 16). Umfrage: Mehrheit der Deutschen misstraut der KI-Politik. *The Decoder*. https://the-decoder.de/umfrage-mehrheit-der-deutschen-misstraut-der-ki-politik/

Schuster, T., Waidelich, L. F., Schneider, A., & Lambert, M. (2025). Risk Classification and Compliance of AI Systems under the EU AI Act. *AMCIS 2025 Proceedings*, 17. https://aisel.aisnet.org/amcis2025/sig_sec/sig_sec/17

Tan, B., Armoush, N., Mazzullo, E., Bulut, O., & Gierl, M. J. (2024). A Review of Automatic Item Generation Techniques Leveraging Large Language Models. [Preprint]. https://doi.org/10.35542/osf.io/6d8tj

The General-Purpose AI Code of practice. (2025, December 10). Shaping Europe's Digital Future. https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

The Government of the Grand Duchy of Luxembourg. (2025). *Luxembourg's AI Strategy: Accelerating Digital Sovereignty 2030* [Book]. https://gouvernement.lu/dam-assets/images-documents/actualites/2025/05/16-strategies-ai-donnees-quantum/2024115332-ministere-etat-strategy-ai-en-bat-acc-ua.pdf

Toffaletti, S. (2025, October 23). *Why delaying AI Act enforcement is essential for European SMEs*. European DIGITAL SME Alliance. https://www.digitalsme.eu/why-delaying-ai-act-enforcement-is-essential-for-european-smes/

Tomikawa, Y., Uto, M. (2024). Difficulty-controllable reading comprehension question generation considering the difficulty of reading passages. *International Conference On Computers in Education*. https://doi.org/10.58459/icce.2024.4931.

Ugen, S., Schiltz, C., Fischbach, A., & Cate, I. P. (2021). *Lernstörungen im multilingualen Kontext: Diagnose und Hilfestellungen.* https://doi.org/10.26298/bw1j-9202

UN Environment Programme. (2025, November 13). *AI has an environmental problem. Here's what the world can do about that.* https://www.unep.org/news-and-stories/story/ai-has-environmental-problem-heres-what-world-can-do-about

University of Luxembourg. (2025). *Luxembourg Centre for Educational Testing (LUCET)*. uni.lu. https://www.uni.lu/fhse-en/research-groups/luxembourg-centre-for-educational-testing-lucet/

Wang, Z., Valdez, J., Mallick, D. B., & Baraniuk, R. G. (2022). Towards human-like educational question generation with large language models. *Lecture Notes in Computer Science*. https://doi.org/10.1007/978-3-031-11644-5_13

Xiao, C., Xu, S. X., Zhang, K., Wang, Y., & Xia, L. Evaluating Reading Comprehension Exercises Generated by LLMs: A Showcase of ChatGPT in Education Applications. *Conference: Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. 2023 https://doi.org/10.18653/v1/2023.bea-1.52

Yan, L., Greiff, S., Teuber, Z., & Gašević, D. (2024). Promises and challenges of generative artificial intelligence for human learning. *Nature Human Behaviour*, *8*(10), 1839–1850. https://doi.org/10.1038/s41562-024-02004-5

Zhai, C., Wibowo, S., & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments*, *11*(1). https://doi.org/10.1186/s40561-024-00316-7