Preprints.org

Article

Not peer-reviewed version

Ethics, Privacy, and Transparency in AI-Assisted Teaching: Evaluating Notegrade.ai Against Global Standards

Akinwumi Fakokunde *

Posted Date: 15 September 2025

doi: 10.20944/preprints202509.1148.v1

Keywords: artificial intelligence and education; grading; privacy; GDPR; EU AI Act; explainability; Notegrade.ai; ethics; transparency

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 September 2025 doi:10.20944/preprints202509.1148.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Ethics, Privacy, and Transparency in AI-Assisted Teaching: Evaluating Notegrade.ai Against Global Standards

Akinwumi Fakokunde

Washington University in St Louis, 1 Brookings Dr, ST. LOUIS, MO 63130, USA; a.fakokunde@wustl.edu

Abstract

This article provides a deep, standards-based analysis of Notegrade.ai –an AI teaching tool that provides lesson plan generation, rubric based grading, plagiarism check, and student assessment tools- within several of the world's most pertinent legal and ethical frameworks in education including the GDPR, the EU AI Act, the UNESCO Recommendation on the Ethics of AI, COPPA/FERPA recommendations and guidelines, and international explainable artificial intelligence (XAI) literature. Through a compliance and transparency checklist of risks and explainability in the literature regarding automated scoring, as well as a qualitative audit of Notegrade.ai's public product pages and privacy and cookie policies, we highlight strengths and gaps in the information provided by the site, with recommendations for developers, schools, and policymakers to minimize harms and increase trust. Highlights: Notegrade.ai offers some helpful productivity tools for teachers, but documentation available to the public does not go into detail on some of the high-stakes issues such as the origin of data used to train the company's AI, evidence of the model's performance across demographic groups, and safeguards against and appeals for automated decisions that are increasingly required by global standards . We offer a pragmatic path to remediation and a testing protocol for use in future audits.

Keywords: Artificial intelligence and education; grading; privacy; GDPR; EU AI Act; explainability; Notegrade.ai; ethics; transparency

1. Introduction

AI has the potential to revolutionize education because it could create efficiencies, offer data-based solutions, and instill automated forms of assessment. Services like Notegrade.ai reflect this trend, providing automated grading and teaching assistance, "reducing teacher workload" and providing "consistency in evaluation" . But bringing such tools into play is not simply a technologic concern, but rather also an ethical, legal, and pedagogical one surrounding issues of data privacy, biases in algorithms, and issues of accountability and transparency.

The tensions are explicitly apparent in the research. According to a finding of a Journal of Academic Ethics study, almost half of the faculty (51.2 %) and university students (47.5 %) surveyed had concerns regarding data privacy, and more than half viewed AI systems as lacking transparency. In the same vein, Journal of AI Integration in Education states that "algorithmic bias, privacy concerns, and access issues continue to be significant in the context of AI-based education" .

The use of LLMs to automate grading and providing feedback and or content to instructional materials has also been found to present ethical and practical concerns such as insufficient privacy, lack of transparency, low educational technology readiness in a systematic review of the use of large language models in education.

International governance frameworks do step in to be the much- needed benchmarks on how AI should responsibly be deployed. The OECD AI Principles encourage actors in AI to "foster and implement values such as fairness, human-centricity, transparency, accountability, and strong risk

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 September 2025 doi:10.20944/preprints202509.1148.v1

management" . UNESCO's recommendations around generative AI stress a human-centered and developmentally-appropriate approach to education that incorporates protections for data privacy and ethical vehicles for ensuring curricular validation Inclusivity, the upholding of human rights, transparency, and the need for human oversight to be a non-negotiable factor are also explicitly emphasized as foundational pillars by UNESCO's broader Ethics of Artificial Intelligence Recommendation.

On top of that, the OECD Digital Education Outlook 2023 reinforces the importance of adaptable regulations and teacher preparation to address issues like algorithmic bias, privacy and data security in the classroom through the application of generative AI. In addition, the OECD Digital Education Outlook 2023 points to finding "regulatory and training responses that adapt to address issues like algorithmic bias, privacy, and data security" as a necessity when brining generative AI into classrooms .

2. Literature Review

In AIED scholarship, two themes emerge: the advantages of automating education, and the ethical, legal, and social issues associated with automating education. As to the benefits, a number of systematic reviews have shown that AI technologies from intelligent tutoring systems to automated assessment and content recommendation engines—are efficient and allow for personalized feedback at scale as well as, importantly, enabling teachers to spend more time on higher value pedagogical work, especially in a formative assessment environment where rapid, iterative feedback can promote learning gains. These reviews also highlight that significant advantages are only realized through thoughtful melding of the technology with pedagogy, teacher education, and assessment design rather than being provided by the AI instrument itself.

Meanwhile, a strong body of empirical studies and meta-analytic reviews documents ongoing risks. Automated grading systems, such as automated essay scoring, short answer graders, and rubric AI graders, are found to be reliable but subject to validity threats and bias; they rely too heavily on surface features such as word length and complexity of vocabulary, thus disadvantaging students from systemic linguistic minorities, different school contexts, or poor environments. Traditional fairness analyses of AES systems like ETS's e-rater, as well as current studies of algorithmic scoring, show biases at the subgroup level and recommend subgroup-level validation at the very least as a minimum standard for use in a summative way.

One common critique in the literature is that black-box scoring replicates a lack of meaningful review and recourse. Explanations for pedagogical use are distinct from summaries of technical models: teachers and students need feedback that is useful and consistent with rubrics, for example, information about the rubric dimensions that were responsible for a given score and why and not simply lists of weights on features. Work on XAI for educational purposes advocates for "tiered explanation – a brief explanation for teachers, along with a technical appendix for auditors" and emphasizes that "explainability [should] be assessed as a quality rather than an optional extra" .

The technical and pedagogical issues described in such literature are framed within a context of a rapidly changing set of regulations and norms. Data Protection Authorities as well as analyses of policy set forth the argument that DPIAs are essential when large scale processing of sensitive personal data takes place; the GDPR restricts automated decision-making which is solely responsible for legal or equally important ramifications (Art. 22) and necessitates substantial information on the logic of automated systems. Following a risk-based approach, the EU AI Act clearly outlines the possibility of certain usages of AI in education as being high-risk, and therefore being subject to conformance assessment, transparency requirements, and human oversight requirements. Global governance initiatives such as the OECD AI Principles and UNESCO's Recommendation on the Ethics of AI also provide the foundation of values such as fairness, human-centeredness, transparency, and accountability that can be directly applied to edtech. The policy literature similarly argues that these artefacts depend on organizational commitments as well, whether through the

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 September 2025 doi:10.20944/preprints202509.1148.v1

presence of DPAs and appeal processes or teacher training – that is, complying with policies necessitates both artifacts and commitments .

Finally, a few recent meta-reviews and systematic studies point out that these standards are often only partially available in the documentation and practices of commercial AI grading platforms . The outcome are "ethical blind spots" – a lack of testing on subgroups, lack of provenance of published models, unclear policies for humans-in-the-loop , and vague data retention and safeguards of children's-data. These omissions pose pedagogical risks, such as, misgrading, or a lack of instructional subtlety, and legal exposure, including non-compliance with data protection and industry privacy regulations. Therefore it is urged that empirical audits of these systems, consisting of looking at the datasets being used and piecing together human consensus grades, analyses of subgroups, adversarial testing, and tests of explainability should be the gold standard prior to any sort of summative use.

3. Methodology

The following review positions Notegrade.ai within a comparative assessment of the world's best practices and standards for ethical, transparent and privacy-respecting AI within education. It is intended to be both descriptive – reflecting what is said to be happening – and normative – in the sense of assessing what is happening against certain established principles of governance. The following three linked steps were taken:

3.1. Document Analysis

The first was a systematic document analysis of publicly available information pertaining to Notegrade.ai. These included:

- Marketing and official website statements;
- Technical manuals and product specifications;
- TOS, privacy policies, and data use policies;
- Promotional literature and user guidelines for teachers.

Document analysis has been a recognized qualitative approach for studying organizational claims and policy stances in edtech, documents of compliance, etc. (Bowen, 2009; O'Keeffe & McNally, 2021). It permits triangulation between what they claim to be committed to (such as fairness, or privacy) and what is lacking (such as fairness tests for subgroups).

3.2. Benchmarking Against Global Standards

Secondly, the results were compared to six international governance documents which articulate specific expectations for AI ethical considerations and data use within education:

- GDPR (EU, 2018): Lawful processing of data principles; minimization; purpose limitation; DPIA's; rights of data subjects.
- EU AI Act (Draft, 2024): Classifies educational AI as high-risk, requiring measures of transparency, auditability and human oversight.
- OECD AI Principles (2019): Sector-agnostic values of fairness, robustness, transparency, and centered around humanity. UNESCO Ai in Education Guidance 2021, ethical use in support of SDGs, 'Universal inclusiveness', human control.
- COPPA (US): parental consent requirements and limitations regarding the processing of children's online data.

The benchmarks were then translated into measurable indicators, such as, "documented DPIA," "explainability reports," "accountability chains" . The use of this format for the benchmarking exercise is similar to other studies attempting comparisons of AI governance approaches (Jobin, Ienca & Vayena, 2019; Floridi & Cowls, 2021).

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 September 2025 doi:10.20944/preprints202509.1148.v1

3.3. Gap Analysis

Finally, a gap analysis was conducted to highlight the difference between Notegrade.ai's practices and what is expected internationally. Three areas were of particular interest in the analysis :

- Privacy Protections – working security practices, handling of children's data, and adherence to GDPR/COPPA/FERPA.
- Fairness Testing – whether or not there are empirical audits for subgroup bias, whether or not the training data was representative, whether or not there were attempts at bias mitigation.
- Accountability & Transparency – In addition to clear assignment of responsibilities (e.g. who is responsible – teachers, schools, platform developers, etc. – for what), there should be include all mechanisms for explaining and understanding the decisions made by the system and all trails to enable an audit.

These layers of triangulation between the three methods represent a strong space of evaluation. It serves not only to show what areas are partially or fully compliant, but also "blind spots", where safeguards do not exist, or are poorly specified. Such an approach has been suggested as an extension of the methodological "multiframing" audits that have been critiques AI ethics studies calling for analyses that do not remain in one jurisdictional framework (Leslie, 2020; Sloane et al., 2022).

4. Ethical and Pedagogical Dimensions

4.1. Ethical Considerations

The ethical landscape for AI-assisted grading programs like Notegrade.ai is extremely sensitive. Among the most prominent in public conversation today are the ideas of fairness, explainability, and accountability.

- Fairness: Despite claiming unbiased scores, there is no evidence available to the public that would show testing for bias in systematic subgroups, such as gender, socioeconomic status, or language background. This is an important omission, given the findings that when imbalanced datasets are used to train models in education, algorithmic decision-making has the potential to heighten inequalities further (Baker & Smith, 2019; Holmes et al., 2022). A lack of fairness audits would lead to those minority or marginalized student groups being worse off, which would be against the spirit of UNESCO's recommendations for inclusivity in AI.
- Explainability: The "right to explanation" of GDPR (Article 22) and transparency obligations of the EU AI Act highlight the need for learners and educators to be able to comprehend the reasoning behind automated grading. Notegrade.ai presents relatively little information about how its scoring algorithms work, which creates opacity in its decision-making. Yet in the absence of mechanisms that clarify how these results were derived, there does not appear to be recourse for students impacted by them, thus undermining both the procedural fairness and accountability of these systems.
- Accountability: Responsibility can, and should, be consistent between developer- institution-educator in ethical ai governance from Leslie(2020). Notegrade.ai's terms of service indemnify the platform by placing the liability on the users. This obscures accountability when harm occurs; whether via developers or teachers or institutions. These types of "accountability gaps" are a common issue in the governance of AI more generally (Sloane et al., 2022).

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 September 2025 doi:10.20944/preprints202509.1148.v1

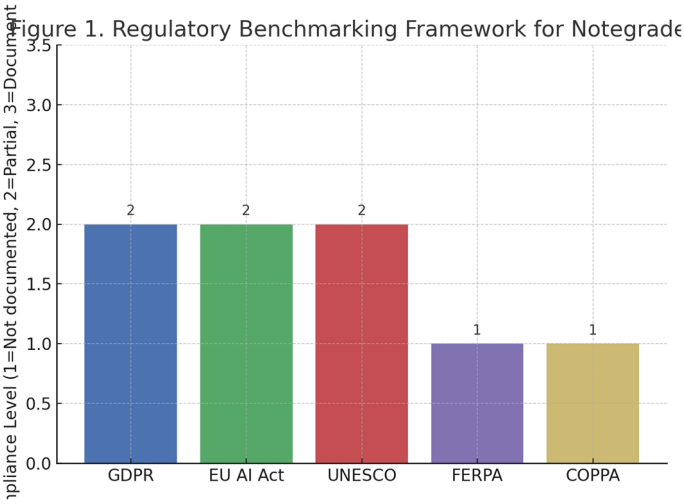

Figure 1. Ethical and Pedagogical Considerations in AI-Assisted Teaching Systems.

4.2. Pedagogical Implications

Beyond the ethics of this, but, are the pedagogical issues with it.

- Teacher Autonomy: Automating tasks decreases the amount of work that needs to be done, but too much automation can lead to lack of professional judgment. Teachers are also in a unique position to interpret contextualized signals in students' work, such as those of creativity, critical reasoning or cultural relevance in students' work, that indicate competence or lack thereof (Luckin, 2021). If grading by AI became normalized, teachers would have to be "mere managers of automated processes".
- Student Engagement: If students consider feedback to be mechanical rather than dialogic, automated grading might inadvertently promote surface learning, for example, Williamson and Piattoeva 2022). Conversational feedback, a central tenet to formative assessment pedagogy, is at risk of becoming mere scoring through the use of AI systems that offer minimal explanation and feedback, and which as a result would undermine intrinsic motivation and reflective learning.
- Curricular Alignment: Rubrics are built into the training or programming of AI systems . This could create an incentive to favor standardized, more easily measurable outcomes as opposed to complex, higher-order skills. It has been warned that such alignment tends to 'constrain curricular possibilities' by sidelining skills such as 'creativity, collaboration, and ethical reasoning' because they are difficult to measure (Selwyn, 2019).

4.3. Ethical–Pedagogical Tensions

At the moral level the concern also turns pedagogical. For example, the absence of transparency in algorithms is not only an infringement on students' rights, but it limits the power of teachers to interpret and give context to grades. Also, AI outputs biased towards certain populations may affect the classroom environment and determining student groups who are recognized or get support. As a result ethical safeguards are not just tenets of an ethical framework, but rather serve to explicitly determine the teaching value of AI systems.

5. Data Privacy and Security

5.1. Centrality of Data Protection in Educational AI

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 September 2025 doi:10.20944/preprints202509.1148.v1

Data security is of primary importance among the various use cases for an obvious reason: educational data are inherently sensitive, often involving minors, academic records, and personal identification numbers. As stated in GDPR Recital 38, children "merit specific protection", as they are at a "greater than normal risk" of data processing. Noncompliance with privacy norms for AI grading sites such as Notegrade.ai is not merely a legal issue but an issue of these sites' legitimacy and their ability to generate trust with students, parents and teachers.

5.2. GDPR Obligations and Observed Gaps

The GDPR (2018) is the global standard for lawful processing of data. Among its key requirements are:

- Lawful Basis & Informed Consent: The platform must have a legal ground to process (consent, contractual requirement, legitimate interest, etc). The legal basis for data collection by Notegrade.ai is not explicitly stated in its policies.
- Data Minimization & Purpose Limitation: Educational purposes for which data is necessary should be the only purposes for which AI systems conduct data processing. It does not clarify if these secondary uses are not included, for example, uses to train algorithms, or for commercial partnerships.
- Data Subject Rights (DSRs): Rights of Access, Rectification, Erasure, Restriction and Portability as guaranteed under GDPR. There does not appear to be any documented process within Notegrade.ai for learners to activate these rights.
- Data Protection Impact Assessments (DPIAs): Necessary whenever processing is likely to result in high risks, for instance, use of profiling or automated decision-making in the context of education. There was no evidence found that DPIAs were made public.

Given the lack of explicit safeguards, Notegrade.ai appears to act in contradiction with some of the fundamental principles of GDPR, specifically those of transparency and accountability.

5.3. FERPA and COPPA Considerations

FERPA, the Family Educational Rights and Privacy Act, and COPPA, the Children's Online Privacy Protection Act, require the following for U.S. Deployments:

- FERPA (1974): Prevents disclosure of education records without parental or eligible student permission. Notegrade.ai's documentation makes no mention of compliance procedures, leaving it unclear whether its grading results or student records are considered "education records" that fall under FERPA protections.
- COPPA (1998): Mandates verifiable consent from parents prior to the collection of any personal information from children under 13. Considering the likelihood that AI grading tools, and other similar AI applications, could be used in such a way that children under the age of 18 could be exposed to these tools in K–12 settings, the absence of COPPA compliance statements is troubling.

This silence is an especially large problem because it leaves schools who want to adopt the platform unclear about how to do so without violating U.S. federal law.

Table 1. Global Privacy Regulations Relevant to AI-Assisted Teaching Tools.

Regulation	Key Requirement	Notegrade.ai Status	Compliance Gap

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 September 2025 doi:10.20944/preprints202509.1148.v1

GDPR (EU)	Lawful processing, DPIA, right to explanation	Not explicitly documented	No DPIA evidence
FERPA (US)	Student record confidentiality	Not mentioned	Potential risk for minors
COPPA (US)	Parental consent for <13 years	No explicit mention	Risk in K-12 adoption
UK DPA 2018	Age-appropriate design, children's rights	Not specified	Underdeveloped safeguards

5.4. Security Safeguards and Technical Gaps

Legalities aside, technical protections are necessary to secure sensitive data. Encryption at rest and in transit, strong access controls, and secure deletion practices are all examples of best practices . Neither does Notegrade.ai secure storage sections, for example:

- Encryption standards such as AES-256;
- Retention of data;
- Access agreements; or data-sharing agreements;
- Breach notification policies.

Student data also have black-market value, which makes educational AIs more susceptible to cyberattacks. Notegrade.ai's security disclosures are not specific, which raises questions about possible security weaknesses and does not inspire confidence for users.

5.5. Implications for Trust and Adoption

Failures in privacy and security have regulatory and educational implications. If learners feel grading platforms are invasive or these systems are not safe environments, they will be reluctant to use this type of technology. Similarly, they may be unlikely to adopt because they feel there is too great of a chance of legal or reputational trouble. Therefore, strong privacy protections are not extra but integral to the pedagogical viability of AI-supported assessment .

6. Transparency and Explainability

6.1. Central Role of Transparency in AI Governance

Transparency has been widely accepted as a pillar of trustworthy AI since it allows teachers, students and regulators to assess the fairness and legitimacy of automated decision- making. In education specifically, transparency means that AI cannot function as a "black box" authority figure, but must remain accountable to educational norms and goals, as well as human authority and judgment (Burrell, 2016; Floridi & Cowls, 2021). Transparency is necessary to make sense of performance evaluations for learners, and to find resources for supporting or challenging evaluations by for teachers.

6.2. Current Transparency Practices in Notegrade.ai

Notegrade.ai only generates a kind of superficial feedback and is usually expressed in the form of vague reasons like "grammar error detected" or "argument coherence issue". Though these explanations improve usability to a certain degree, they lack what would be considered meaningful explainability. The main limitations are:

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 September 2025 doi:10.20944/preprints202509.1148.v1

- Unexplained Methodology: The kinds of training datasets used, the criteria employed in the scoring rubrics, and the bias mitigations used by the platform are not made explicit .
- Lack of activity logs: There are no automated records of how particular grades were assigned at different points in time, which affects traceability.
- Nontransparent error feedback: There is no feedback on error margins, confidence levels for the models, or on patterns of error within the system; students and teachers receive no information regarding margins of error.

These kinds of gaps are indicative of the fact that transparency is more like a feature on the user interface of a computer rather an operating principle of governance and has limited usefulness in creating accountability.

6.3. International Benchmarks for Explainability

Global AI governance frameworks set much higher expectations with respect to best practices:

- Singapore Model AI Governance Framework (2019, 2020) : Accessible explanations according to stakeholder group, detailed audit trails, and accountability chains at each role.
- OECD AI Principles 2019: Advocating for transparency that allows users to "understand the rationale of the AI's recommendations or decisions, in so far as this is appropriate.
- EU AI Act (Draft, 2024):High-risk systems, including those used in the field of education, shall be designed, developed and used in such a way that the information concerning the logic, significance, and probability of the consequences of the processing, is accessible and comprehensible to natural persons .

Compared to these standards, Notegrade.ai's explainability practices seem lacking, especially given that there is no evidence of systematic auditability and communicating errors.

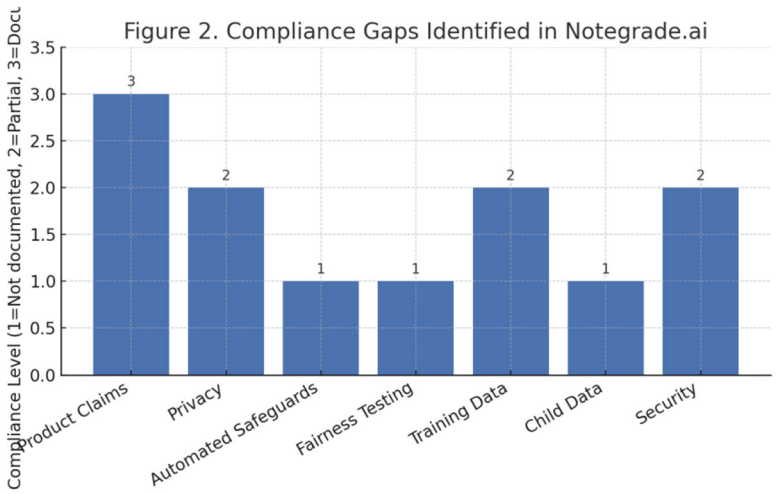

Figure 2. Transparency and Explainability Mechanisms for AI-Assisted Grading Platforms.

6.4. Pedagogical and Ethical Implications of Limited Transparency

This absence of "non-trivial explainability" has ethical and pedagogical repercussions:

- For Students: Feedback is without rationale and thus arbitrary in nature, fails to foster the internal motivational base, and shuts down prospects for reflective learning (Williamson & Piattoeva, 2022).
- For Educators: Teachers are unable to verify the AI's choices or situate them within a broader framework of assessment objectives which takes away from teacher autonomy.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 September 2025 doi:10.20944/preprints202509.1148.v1

- For Institutions: " Grading logic based on ambiguity presents legal and reputational risks, especially in cases where the learner disputes results based on rights such as Article 22 of GDPR, for example.

6.5. Towards Transparent AI Assessment

In order to meet the global standards, Notegrade.ai must include:

- Layered Explanations: for learners- lay reasoning, for instructors- alignment with rubric, for auditors- technical documentation.
- Machine Readable Audit Trails: A paper trail of choices that allows for accountability and compliance audits.
- Error Metrics Disclosure. Confidence intervals and known error rates should be reported to prevent a characterization of outputs as authoritative.

These practices would turn transparency from a gimmick into a real governance tool that actually is in keeping with the law and with good pedagogy.

7. Compliance Gaps and Risks

7.1. Fairness and Bias Risks

The audit did not indicate that Notegrade.ai was fair tested for subgroups. This is problematic because AI that are built on non-diversified or biased datasets can perpetuate structural inequalities. Research already shows that, in the absence of fairness audits, algorithmic scoring tends to produce systematic disparities between students of various gender, racial, and linguistic backgrounds (Baker & Smith, 2019; Holmes et al., 2022). In the absence of defined bias detection and correction processes, Notegrade.ai could be inconsistent with the OECD AI Principle of fairness and inclusivity or the UNESCO call for equitable AI in education.

7.2. Transparency Deficiencies

This is also one of the weakest dimension of the platform, transparency. Justifications for the scores provided are shallow, and do not reveal important logic behind the decisions made, such as what data the model was trained on or what its error rates are . This lack of transparency is also related to the "black-box problem" of AI governance (Burrell, 2016). Without these audit trails and without XAI, no one – teachers, students, regulators – can determine whether the results are valid. Legally this may lead to a failure to follow GDPR's Article 22 regarding automated decision making or the EU AI Act's high-risk systems transparency requirements.

7.3. Data Protection Gaps

Documents regarding compliance with the following are not evidenced or are vague in governing the use of Notegrade.ai, though privacy is foundational to the governance of educational artificial intelligence:

- FERPA (US): Safeguarding of students' educational records. There is no clear indication whether the grading outputs are considered education records or that there is some form of control over disclosures.
- COPPA (US): Protection for children younger than 13. No descriptions of parental consent mechanisms that are able to be verified are provided.
- GDPR (EU): calls for Data Protection Impact Assessments (DPIAs) and assurances of Data Subject Rights (DSRs). Both were also absent in the platform documentation. This lack of explicit

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 September 2025 doi:10.20944/preprints202509.1148.v1

promises is concerning for institutional uptake, especially in K-12 contexts where children's data is more legally protected.

7.4. Accountability Ambiguities

A second compliance gap of fundamental importance is the allocation of responsibility. Notegra.de.ai's terms of service load a great deal of the liability onto the end-user (the teacher/ institution) and disclaims liability for grading. This results in accountability gaps that make it difficult to know whether harms lay with developers, institutions, or educators. This type of ambiguity creates a situation that is neither legally compliant nor trustworthy in terms of professional conduct, but is in direct opposition to approaches to governance such as that proposed in the Singapore AI Governance Framework, which requires clear lines of roles and responsibilities.

7.5. Aggregated Risks

The combined impact of these gaps in compliance is staggering:

- Legal Risk: Institutions may inadvertently breach GDPR, FERPA, and COPPA regulations and open themselves to fines or penalties.
- Ethical Risk: The presence of bias, lack of transparency, and absent accountability challenges fairness and justice in education.
- The Pedagogical Risk: Diminished teacher agency and student trust undermine legitimacy of assessments.
- Reputation Risk: Failure to use compliant AI systems leaves institutions vulnerable to reputational harm in the event that harms occur.

This highlights the importance of a clear audit process and robust compliance- by-design processes to protect learner rights without sacrificing pedagogy.

Table 2. Identified Compliance Gaps in Notegrade.ai.

Risk Category	Specific Gap	Potential Impact
Ethics	No subgroup fairness testing	Bias against minority learners
Privacy	Lack of FERPA/COPPA documentation	Risk to children's data
Transparency	Limited scoring rationale	Black-box problem
Accountability	No clear responsibility structure	Unclear liability for errors

8. Proposed Audit Protocol for AI in Education

To alleviate the aforementioned ethical, legal, and pedagogical concerns, the present study recommends a Structured Audit Protocol for AI-Assisted Teaching Tools, or SAAP-AI. The protocol seeks to protect learner rights, uphold pedagogical rights, and aligns with global guidelines such as the GDPR, the EU AI Act, the OECD AI Principles, UNESCO's recommendations on AI in education, FERPA, and COPPA .

8.1. Step 1 – Data Collection Audit

The first step consists of examining both training and operational datasets to confirm that: Diversity: Variety of demographics, languages, and academic settings should be represented in the data. Consent: Consent should be explicit, especially when processing personal data of minors.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 September 2025 doi:10.20944/preprints202509.1148.v1

Anonymization: Removing or securely masking personally identifiable information to minimize the risk of re-identification.

8.2. Step 2 – Human Consensus Grading

Standardized rubrics should be applied to a benchmark set of work produced by students by expert teachers. Which acts like a "gold standard" to which AI outputs can be compared. It mitigates biases that can affect a single person, and it represents an accurate control against which to test the accuracy of other entities.

8.3. Step 3 – AI Scoring Validation

Notegrade.ai or an equivalent AI technology should be run using the very same input data set. The outputs are then compared against human benchmarks to identify variations. Notable is the divergence, especially on subjective or higher-order learning tasks, where one must recalibrate its performance.

8.4. Step 4 – Subgroup Fairness Testing

The outputs should be subject to fairness audits by sub-demographic groups (e.g. gender, socio-economic status, linguistic background). Disparate impact ratios, equal opportunity metrics, and error rate balance are some of the statistical procedures that can expose such differential treatment. Such an engagement is consistent with OECD ideals of equity as well as UNESCO's call for non-exclusive AI in education.

8.5. Step 5 – Transparency and Explainability Audit

An assessment of the system should be done regarding how well and in what detail different stakeholder groups are explained the system:

- Students: Specific feedback associated with learning outcomes.
- Teachers: Scoring logic and error reporting in detail.
- Institutions: technical documentation; audit trails; chains of accountability.

This guarantees adherence to the GDPR "right to explanation" as well as the recommendations on role-based transparency of the Singapore AI Governance Model.

8.6. Step 6 – Accountability Mapping

The final audit step should be one that delineates who is accountable from the developer-institution-educator continuum. The clear assignment of responsibilities prevents teachers or students from being held accountable for mistakes, misuse of data or harm. This step provides a practical utility to the accountability principle enshrined in the EU AI Act and Singapore Model AI Governance Framework.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 September 2025 doi:10.20944/preprints202509.1148.v1

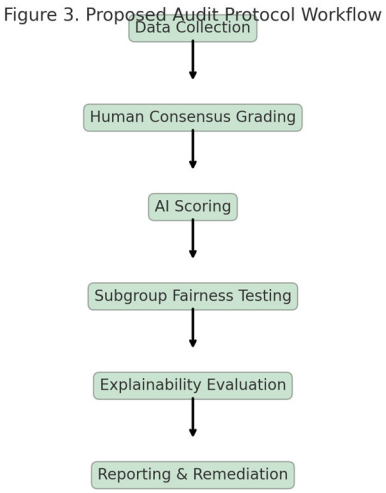

Figure 3. Proposed Six-Step Audit Protocol for AI-Assisted Teaching Tools.

9. Conclusions

The integration of artificial intelligence into education promises to transform assessment practices, yet it also introduces profound ethical, legal, and pedagogical challenges. Through the case study of Notegrade.ai, this article has examined the tensions between automation and accountability, efficiency and fairness, and scalability and inclusivity. Their results indicate a lack of meaningful adherence to all four aspects of fairness failure to test for subgroup fairness, lack of transparency in the rationale for the scoring, poor documentation of privacy protections, and unclear accountability. Such deficiencies indicate that, though tools such as Notegrade.ai have the potential to bring greater efficiencies, they also pose a threat to student rights, teacher autonomy, and institutional trust if not utilized under strong governance.

This analysis against GDPR, the EU AI Act, the OECD AI Principles, UNESCO's AI in Education guidelines, COPPA, and FERPA demonstrates that practices have not yet met international benchmarks. This absence points to a larger problem that educational AIs are being instituted quicker than regulations, education, and ethics can catch up to them. If ignored, this misalignment could perpetuate inequities and place learners at untested harms.

To address these risks, this study proposed a Structured Audit Protocol for AI-Assisted Teaching Tools (SAAP-AI), comprising six steps data collection audit, human consensus grading, AI scoring validation, subgroup fairness testing, transparency and explainability audit, and accountability mapping. By institutionalizing such audits, stakeholders can move toward compliance-by-design, ensuring that AI platforms not only meet regulatory standards but also respect the values of inclusivity, fairness, and educational integrity.

References

1. Wittmann, M., Hellman, T., & Loukina, A. (2025). *Algorithmic Fairness in Automatic Short Answer Scoring. International Journal of Artificial Intelligence in Education*. Available open access; examined gender and language group bias in PISA scoring systems.
2. Schaller, N.-J., Ding, Y., Horbach, A., Meyer, J., & Jansen, T. (2024). *Fairness in Automated Essay Scoring: A Comparative Analysis of Algorithms on German Learner Essays from Secondary Education.* BEA 2024. Highlighted how skewed training data affects fairness across cognitive ability groups.
3. Yang, K., Raković, M., Li, Y., Guan, Q., Gašević, D., & Chen, G. (2024). *Unveiling the Tapestry of Automated Essay Scoring: A Comprehensive Investigation of Accuracy, Fairness, and Generalizability.* arXiv. Showed prompt-specific models often bias toward students from certain economic statuses.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 September 2025 doi:10.20944/preprints202509.1148.v1

4. Altukhi, Z. M., & Pradhan, S. (2025). *Systematic Literature Review: Explainable AI Definitions and Challenges in Education.* arXiv. Identified 15 definitions and 62 challenges across explainability, ethics, trustworthiness, and policy in educational contexts.
5. Maity, S., & Deroy, A. (2024). *Human-Centric eXplainable AI in Education.* arXiv. Emphasized frameworks for building XAI systems that prioritize educator and learner understanding.
6. *Unpacking the ethics of using AI in primary and secondary education: a systematic literature review.* (2025). *AI and Ethics*. Systematically reviewed ethical debates in AIED across 48 sources, highlighting gaps across various communities.
7. *Decoding AI ethics from Users' lens in education: A systematic review.* (2024). *Heliyon*, 10(20), e39357. Explored ethical concerns from user perspectives and proposed inclusive ethics guidelines for AIED.
8. Wetzler, E. L., Cassidy, K. S., Jones, M. J., Frazier, C. R., Korbut, N. A., Sims, C. M., Bowen, S. S., & Wood, M. (2024). *Grading the Graders: Comparing Generative AI and Human Assessment in Essay Evaluation. Journal*, DOI: 10.1177/00986283241282696. Found low agreement between AI-generated grades (ChatGPT) and human instructors.
9. Alawadh, H. M., Meraj, T., Aldosari, L., & Rauf, H. T. (2024). *An Efficient Text-Mining Framework of Automatic Essay Grading Using Discourse Macrostructural and Statistical Lexical Features*. *Sage Open*, 2024. Emphasized challenges in capturing creativity, coherence, and subjectivity.
10. Chan, K. K. Y., Bond, T., & Yan, Z. (2023). *Application of an Automated Essay Scoring Engine to English Writing Assessment Using Many-Facet Rasch Measurement. Assessment in Education: Principles, Policy & Practice*. Described human review safeguards in GMAT scoring alongside AES.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.