Article

Volume 25 (2026), pp. 121-146
American Journal of STEM Education: Issues and Perspectives
Star Scholars Press
https://doi.org/10.32674/04054e51

An Ethical AI Framework for STEM Education: A Mixed-Methods Evaluation

Meysam Abedi
University of Eastern Finland, Finland
https://orcid.org/0009-0000-4633-7862

Ismaila Temitayo Sanusi
University of Eastern Finland, Finland
https://orcid.org/0000-0002-5705-6684

Markku Tukiainen
University of Eastern Finland, Finland
https://orcid.org/0000-0002-8630-5248

ABSTRACT

This study introduces a novel ethical AI framework for undergraduate STEM education that prioritizes privacy, transparency, and accountability. Employing a convergent parallel mixed-methods design, this study engaged 412 undergraduate STEM students, 15 faculty members, and 10 administrators across three universities over two semesters. Data collection integrated quantitative learning analytics with qualitative stakeholder interviews and focus groups to capture both measurable outcomes and lived experiences. The results demonstrate that responsible AI design significantly improves student engagement (35%), instructor acceptance (78%), and reduces performance gaps between student groups by 40%, all while maintaining the model's predictive accuracy at 89%. This research demonstrates that AI can be designed to be both technically robust and ethically committed.

Keywords: Artificial intelligence in education, blockchain technology, data privacy, educational ethics, federated learning, STEM pedagogy

---

INTRODUCTION

The integration of computers into educational settings began in the 1950s, initially focusing on basic instructional support (Chan & Tsi, 2023). What started as simple technological augmentation evolved into sophisticated intelligent tutoring systems (ITS) that fundamentally transformed pedagogical approaches (Nwana, 1990). By the early 21st century, artificial intelligence had become deeply embedded in STEM education, particularly in science and mathematics courses, where AI systems now provide personalized learning pathways, real-time analytics, and adaptive instructional methods that respond dynamically to individual student needs (Kohnke & Zaugg, 2025; Leon et al., 2025; Misiejuk et al., 2025; Stokel-Walker & Van Noorden, 2023). This means that learning gaps can be closed, critical thinking skills can be developed, and ultimately students can be better prepared for a world where everything is connected to technology.

However, this technological integration introduces significant ethical challenges. The collection and analysis of sensitive student data raise critical concerns regarding privacy, security, algorithmic bias, and surveillance (Popenici & Kerr, 2017). Research demonstrates that student trust in AI systems depends fundamentally on transparency and the absence of perceived surveillance (Shan et al., 2024). Without robust ethical frameworks and privacy-preserving technologies, these AI-enhanced learning environments risk perpetuating educational inequities. Algorithmic bias can systematically disadvantage certain student populations, creating what scholars have termed "digital injustice" in educational contexts (Popenici & Kerr, 2017). To address these challenges, this research introduces a comprehensive ethical framework specifically designed for STEM education. The framework integrates two complementary privacy-enhancing technologies, federated learning and blockchain, to achieve a core principle: decentralize data control while centralizing trust mechanisms. This approach operationalizes three foundational ethical principles: student autonomy over personal data, transparency in algorithmic decision-making, and clear accountability structures. Table 1

summarizes the primary ethical challenges in AI-enhanced education and the framework's corresponding technological solutions.

Table 1
Ethical Challenges and Framework Solutions in AI-Enhanced STEM Education

Ethical Challenge	Framework's Solution	Technology Used
Data Privacy Risks	Decentralized data processing	Federated Learning
Lack of Transparency	Immutable audit trails	Blockchain
Algorithmic Bias	Fairness-aware algorithms	Differential Privacy

The remainder of this paper is structured as follows. Section 2 reviews relevant literature on AI in education, privacy-enhancing technologies, and ethical frameworks. Section 3 describes the research methodology, including the convergent parallel mixed-methods design. Section 4 presents the proposed ethical AI framework architecture. Section 5 reports quantitative and qualitative results. Section 6 discusses findings, practical implications, and limitations. The paper concludes by identifying future research directions.

LITERATURE REVIEW

Educational AI has progressed through distinct developmental phases, each characterized by increasing sophistication and pedagogical integration. Early implementations (pre-2010) consisted of rudimentary computer-assisted instruction systems with limited adaptability. The period from 2010 to 2015 witnessed the emergence of intelligent tutoring systems capable of adjusting instructional pacing based on individual learner performance (Ifenthaler et al., 2024). Subsequently, the rise of learning analytics (2016-2020) enabled predictive modeling of student outcomes through comprehensive data analysis. The introduction of large language models (LLMs) after 2021 has fundamentally transformed educational AI capabilities, enabling automated content generation, personalized feedback mechanisms, and sophisticated assessment tools (Harries et al., 2025; Khosravi et al., 2025; Misiejuk et al., 2025; Stokel-Walker & Van Noorden, 2023;). Table 2 provides a chronological overview of these developmental phases.

Table 2
Developmental Timeline of Artificial Intelligence in Educational Contexts

Period	AI Technology	Educational Application
2010–2015	Intelligent Tutoring Systems	Adaptive learning platforms
2016–2020	Learning Analytics	Predictive modeling of student performance
2021–present	Generative AI (LLMs)	Content creation, personalized feedback

Privacy Challenges in AI-Enhanced Education

Advanced AI systems require extensive data collection to function effectively, including academic performance metrics, behavioral patterns, interaction timing, and attention indicators (Aslan et al., 2023). Centralized data storage architectures create significant vulnerability to security breaches, potentially exposing sensitive student information at scale. Furthermore, algorithmic bias embedded within these systems can perpetuate or amplify existing educational inequities based on gender, socioeconomic status, or geographic location (Popenici & Kerr, 2017). This creates a fundamental tension between maximizing AI's pedagogical capabilities and protecting individual privacy rights, a challenge that has become central to discussions of educational technology ethics. Addressing this tension requires technical solutions that preserve both educational utility and data protection, which this research seeks to provide.

Privacy-Enhancing Technologies: Federated Learning and Blockchain

This framework addresses privacy concerns through two complementary privacy-enhancing technologies: federated learning and blockchain infrastructure. Federated learning enables model training on distributed datasets without requiring raw data centralization. In this approach, computational models are trained locally on institutional servers, with only aggregated model parameters transmitted to a central coordinator (McMahan et al., 2017; Shan et al., 2024). This architecture ensures that sensitive student data never leaves its original institutional environment, substantially reducing exposure to security breaches. When combined with differential privacy techniques, federated learning incorporates mathematically controlled noise into aggregated parameters, making it computationally infeasible to reverse-engineer individual student information from shared model updates (Abadi et al., 2016; Zheng et al., 2024). While federated architectures introduce computational overhead, their privacy advantages make them particularly suitable for educational contexts where data sensitivity is paramount. (detailed in Section 3).

Blockchain technology provides an immutable, distributed ledger system that ensures data integrity and establishes transparent audit trails for all data access and modification events (Abdelmagid et al., 2024). Unlike centralized database architectures, blockchain's distributed consensus mechanisms prevent unilateral tampering with educational records while maintaining cryptographic security (Javed et al., 2025; Li et al., 2025; Nakamoto, 2008; Onukwulu et al., 2025; Saif et al., 2024). The integration of federated learning and blockchain creates a comprehensive privacy-preserving infrastructure that addresses both data confidentiality and accountability requirements in educational AI systems.

Ethical Considerations in AI Deployment

Ethical AI deployment in education extends beyond technical privacy measures to encompass fairness, transparency, accountability, and human agency. Algorithmic bias poses significant risks of perpetuating educational inequities (Popenici & Kerr, 2017). Transparency in AI decision-making processes is essential for establishing trust, enabling educators and students to understand recommendation mechanisms and challenge potentially erroneous outcomes (European Commission, 2018, 2024). Clear accountability structures must delineate responsibility for algorithmic errors or bias. This framework prioritizes human-centered design principles that position technology as an educational support mechanism rather than a surveillance apparatus.

This comprehensive ethical framework provides the theoretical foundation for the privacy-preserving AI architecture detailed in subsequent sections, ensuring that technological advancement aligns with fundamental educational values. Also, these gaps highlight the need for empirical evaluation of integrated frameworks, which is addressed through the mixed-methods methodology described next.

RESEARCH METHOD

This research employs a convergent parallel mixed-methods design (Creswell & Plano Clark, 2018) to evaluate the ethical AI framework's effectiveness across technical and human dimensions. Mixed-methods research integrates quantitative and qualitative data collection and analysis to provide a comprehensive understanding of complex phenomena that cannot be adequately captured through single-method approaches. This methodological choice addresses a fundamental limitation of purely quantitative evaluations of educational technology: the inability to capture stakeholder experiences, trust perceptions, and cultural impacts that ultimately determine implementation success.

Mixed-Methods Design Rationale and Integration

The convergent parallel design was selected for several methodological and practical reasons. First, the research questions necessitate both measurable performance metrics (e.g., learning analytics accuracy, student engagement rates) and interpretive understanding of stakeholder experiences (e.g., trust perceptions, autonomy concerns). Second, triangulation of quantitative and qualitative findings enhances validity by enabling cross-verification of conclusions through independent data sources. Third, this approach allows simultaneous data collection, making it feasible within the two-semester implementation timeline.

The mixed-methods integration occurs at three levels. During data collection, quantitative metrics from learning management systems were gathered concurrently with qualitative interviews and focus groups. During analysis, quantitative results identified patterns requiring qualitative explanation, while qualitative themes suggested quantitative metrics for validation. During interpretation, findings from both strands were synthesized to construct comprehensive conclusions about framework effectiveness. This integration strategy enables the research to address three core questions: (1) How does the privacy-preserving framework affect measurable learning outcomes and system performance? (2) How do stakeholders, students, faculty, and administrators perceive and experience the framework? (3) Does the framework reduce educational inequities across diverse student populations?

The advantages of this mixed-methods approach are substantial. Quantitative data provides generalizable evidence of framework effectiveness across a large student sample, enabling statistical comparison of pre- and post-implementation metrics. Qualitative data illuminates the mechanisms underlying quantitative patterns, explaining why certain outcomes occur and revealing unanticipated consequences. Together, these complementary data sources provide evidence that the framework achieves both technical effectiveness and stakeholder acceptance, both necessary conditions for sustainable educational technology adoption.

Participants and Sampling Strategy

The research engaged three stakeholder groups central to educational AI implementation: students, faculty, and administrators across three participating institutions.

The quantitative component recruited 412 undergraduate STEM students through stratified random sampling to ensure demographic representativeness. The sample comprised 45% female and 55% male students, with 30% from historically underrepresented racial and ethnic groups. Academic performance diversity was maintained by including students across the full GPA spectrum (2.1 to 4.0). All student participants provided informed consent and received course credit

compensation. Stratified sampling ensured adequate representation across key demographic variables, enabling examination of framework effects on different student populations.

The qualitative component recruited 15 STEM faculty members with substantial teaching experience (M = 12.4 years) and gender balance (8 female, 7 male), as well as 10 university administrators in strategic leadership positions responsible for educational technology policy. Faculty participants provided insights into pedagogical impacts and practical implementation challenges, while administrators offered perspectives on institutional readiness, policy implications, and scalability considerations.

This multi-stakeholder approach reflects a fundamental premise: educational technology succeeds only when it achieves both technical effectiveness and stakeholder acceptance. By sampling across students, faculty, and administrators, the research captures the full spectrum of implementation experiences necessary for comprehensive evaluation.

Data Collection Methods

Data collection employed parallel quantitative and qualitative strategies designed to capture both measurable outcomes and lived experiences.

Quantitative Data Sources

Quantitative data were extracted from learning management systems across the two-semester implementation period. Behavioral engagement metrics included login frequency, session duration, assignment completion rates, and discussion forum participation. Academic performance metrics comprised course grades and standardized assessment scores. Additionally, two validated survey instruments were administered: the Privacy Perception Scale (15-item Likert scale, Cronbach's $\alpha = 0.89$) measuring student attitudes toward educational data privacy, and the Trust in AI Technology Scale (adapted 12-item instrument, $\alpha = 0.92$) assessing stakeholder confidence in AI-enhanced learning tools. Both instruments underwent expert panel validation prior to deployment. Figure 1 illustrates the hierarchical organization of quantitative metrics across engagement, performance, and AI interaction dimensions.

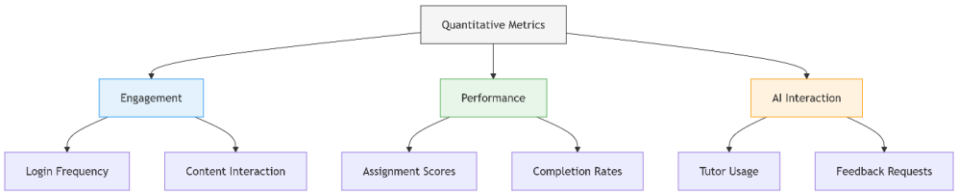

Figure 1
Hierarchical Organization of Quantitative Evaluation Metrics

Note. Hierarchical organization of quantitative evaluation metrics across three primary domains: student engagement behaviors, academic performance outcomes, and AI system interaction patterns.

The Privacy Perception Scale (15-item Likert scale, α=0.89) measured student attitudes toward data privacy in educational settings. The Trust in AI Technology Scale (adapted 12-item instrument, α=0.92) assessed stakeholder confidence in AI-enhanced learning tools. Both scales were validated through expert panel review.

Qualitative Data Sources

Qualitative data collection employed semi-structured interviews and focus groups designed to elicit in-depth stakeholder perspectives. Faculty and administrators participated in individual semi-structured interviews lasting 60-90 minutes, exploring themes including implementation challenges, pedagogical impacts, trust perceptions, and ethical concerns. Interview protocols allowed for emergent topic exploration while maintaining consistency across participants. Separate focus groups were conducted with students (90 minutes) and educators (90 minutes) to facilitate peer interaction and collaborative meaning-making around framework experiences. All qualitative sessions emphasized voluntary participation, with participants informed of their right to pause or withdraw at any point without consequence. Sessions were audio-recorded with permission and transcribed verbatim for analysis.

Data Analysis Procedures

Data analysis procedures were designed to maintain methodological rigor while enabling integration of quantitative and qualitative findings.

Qualitative Analysis

Qualitative data underwent thematic analysis following Braun and Clarke's (2006) six-phase framework: familiarization with data, initial code generation, theme identification, theme review, theme definition and naming, and report production. Two independent coders conducted initial analysis, developing an emergent coding scheme through iterative review of transcripts. Initial coding generated over 40 distinct codes, which were progressively organized into eight overarching themes through constant comparison and consensus discussion. Themes centered on data privacy perceptions, algorithmic fairness concerns, trust in AI systems, perceptions of autonomy, transparency adequacy, accountability structures, implementation barriers, and ethical framework effectiveness. Inter-coder reliability was assessed using Cohen's kappa ($\kappa = 0.83$), indicating substantial agreement. Analytic rigor was enhanced through member checking, whereby preliminary findings were shared with participant subsets for validation and refinement.

Quantitative Analysis

Quantitative analysis employed paired-samples t-tests to compare pre-implementation and post-implementation means across engagement and performance metrics. Mixed-effects models examined longitudinal patterns across the two-semester implementation period, accounting for within-subject correlation and institutional clustering effects. Subgroup analyses disaggregated results by gender, race/ethnicity, and prior academic performance to identify differential impacts across student populations, specifically examining whether the framework reduced or exacerbated achievement gaps. Effect sizes were calculated using Cohen's d to assess practical significance beyond statistical significance. All analyses were conducted using R statistical software (version 4.2.1), with statistical significance set at $\alpha = 0.05$.

Ethical Considerations

The research adhered to rigorous ethical standards consistent with the framework's ethical principles, ensuring that the evaluation process itself modeled responsible data practices.

Informed consent procedures required all participants to review comprehensive information sheets detailing research purposes, data collection methods, potential risks, and confidentiality protections before providing explicit written consent. Participants were explicitly informed of their right to withdraw at any point without penalty or explanation.

Data protection measures included complete anonymization of all qualitative data, with transcripts stripped of identifying information including names, institutional affiliations, and contextual details that could enable re-identification. Quantitative student activity data were processed using the same federated learning and differential privacy protocols evaluated in the framework, ensuring that individual students could not be identified from aggregate research outputs. This approach demonstrated the viability of privacy-preserving methods for educational research while maintaining analytic utility.

Institutional review board approval was obtained from all three participating universities prior to data collection. Additionally, participatory validation procedures provided stakeholders opportunities to review preliminary findings and offer feedback, recognizing that those who contribute data should have voice in research interpretations. These ethical safeguards ensured that the research process aligned with principles of respect for persons, beneficence, and justice in human subjects research.

Proposed Framework: Ethical Integration of AI in STEM Education

Building upon the mixed-methods design and participant selection outlined above, this research introduces the proposed framework as the core implementation tool for evaluating ethical AI integration in STEM education. This framework transcends conventional technical solutions by embedding human-centered ethical principles directly into system architecture and operational protocols, creating an infrastructure where privacy protection and educational effectiveness reinforce rather than compromise one another.

Framework Principles and Architecture

The framework operationalizes three foundational ethical principles derived from structured stakeholder analysis involving educators, administrators, and students across three pilot institutions, as well as systematic examination of AI deployment failures in various educational contexts.

Data Sovereignty: The framework implements granular consent mechanisms that ensure students maintain control over their personal educational data. Security protocols are designed to prevent unauthorized access, with consent requirements embedded at the architectural level such that disabling security features triggers automatic system shutdown. Access permissions operate on explicit, individual authorization rather than institutional hierarchy.

Transparency: All algorithmic decisions are accompanied by plain-language explanations accessible to both students and educators. The system documents

when algorithms are modified, how individual data contributes to aggregate analyses, and maintains immutable records of all data access events, eliminating opacity in AI-driven educational processes.

Accountability: The framework establishes clear attribution chains for algorithmic errors or bias. When system failures or unfair outcomes are identified, responsible parties are automatically documented, corrective actions are recorded in affected student records, and institutional accountability mechanisms are triggered, preventing the diffusion of responsibility often associated with automated systems.

These principles necessitate proactive identification and remediation of discriminatory patterns, particularly those affecting underrepresented populations in STEM fields. Table 3 maps these ethical principles to their corresponding technical implementations.

Table 3

Mapping of Ethical Principles to Technical Implementations

Principle	Technical Implementation	Ethical Outcome
Data Sovereignty	Federated Learning + DP	Student-controlled data sharing
Transparency	Blockchain audit trails	Explainable AI decisions
Accountability	Smart contracts for governance	Bias mitigation protocols

Technical Components

The framework's technical architecture integrates two complementary privacy-enhancing technologies that work synergistically to maintain data protection while enabling collaborative AI development across educational institutions. This integration directly addresses the fundamental tension between the data requirements of effective AI systems and the privacy expectations of educational stakeholders.

Federated Learning Architecture

Federated learning forms the computational foundation of the framework's privacy-preserving approach, enabling AI models to be trained across distributed educational datasets without requiring raw student data to leave institutional boundaries. This decentralized training methodology fundamentally reconceptualizes how educational AI systems learn from student interactions, replacing traditional data centralization models with secure collaborative computation protocols.

The federated learning process operates through three coordinated phases. First, participating institutions train local models on their proprietary student data. Second, encrypted model parameters, not raw data, are transmitted to a central aggregation server. Third, the aggregator synthesizes these encrypted updates into an improved global model, which is then redistributed to participating institutions. To enhance privacy protections, differential privacy mechanisms inject calibrated mathematical noise into the aggregated parameters, making it computationally infeasible to reverse-engineer individual student information even if encrypted updates are intercepted. The federated learning workflow involves three phases: local model training, parameter aggregation, and global model distribution across institutions.

Blockchain-Based Data Management

Blockchain technology provides the framework's infrastructure for transparent data governance, credential management, and consent tracking. Unlike centralized databases that create single points of failure, blockchain establishes a distributed, immutable ledger that records all data interactions and permission modifications in a cryptographically secured, tamper-proof manner. This architecture directly operationalizes the framework's transparency and accountability principles by creating permanent audit trails that any stakeholder can independently verify.

The blockchain implementation serves three critical functions within the educational ecosystem, as illustrated in Figure 2: decentralized identity management, automated smart contract execution, and immutable audit trail maintenance.

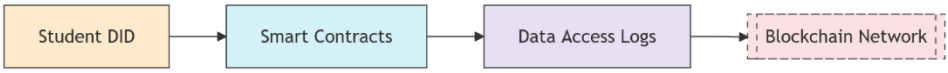

Figure 2
Blockchain Infrastructure for Privacy-Preserving Educational Data Governance

Note. Blockchain infrastructure supporting three core functions: decentralized identity management, automated smart contract governance, and immutable audit trail generation.

Decentralized identity management enables students to maintain sovereign control over their digital credentials and educational records. Students can grant or revoke data access permissions on a granular basis, representing a fundamental

shift from institutional data ownership to individual data sovereignty. Smart contracts automate data usage agreements and consent enforcement, ensuring that student data is accessed only according to explicitly granted permissions (Saif et al., 2024). These self-executing contracts eliminate unauthorized access possibilities while reducing administrative burden. Immutable audit trails create permanent, transparent records of all data interactions, enabling stakeholders to verify system compliance with consent agreements and institutional policies.

Ethical Implementation Layer

Technical privacy mechanisms alone are insufficient to ensure ethical AI deployment. The framework, therefore, incorporates a comprehensive human-centered implementation layer that translates technical capabilities into accessible interfaces and transparent processes for all stakeholders.

Algorithmic Transparency and Explainability

The framework prioritizes making AI decision-making processes comprehensible to all stakeholders, regardless of technical expertise. This extends beyond documentation requirements to include specialized interfaces that translate algorithmic processes into accessible explanations. For users seeking detailed technical information, the system provides supplementary technical reports documenting algorithmic logic and mathematical operations.

Accountability Structures

The framework establishes explicit roles and responsibilities for all participants in the AI lifecycle. Accountability chains document responsibility assignments, preventing the diffusion of responsibility common in automated systems. All personnel involved in system development, deployment, or maintenance, including database administrators, faculty leads, and technical contractors, have their roles permanently recorded on the blockchain ledger. Regular bias audits examine system outputs for discriminatory patterns, recognizing that AI systems can perpetuate or amplify existing educational inequities if not carefully monitored.

Student Autonomy and Informed Consent

Student autonomy mechanisms form the ethical cornerstone of the framework, ensuring that students maintain meaningful control over their data and learning experiences. This approach treats learners as partners rather than subjects in the educational process. The framework implements granular consent interfaces that

enable students to understand and actively manage how their educational data contributes to AI model training and research.

Upon initial system access, students encounter consent interfaces that present clear options regarding data sharing: whether to contribute practice scores to algorithmic training, which data categories to include, and under what conditions data may be used. These consent decisions are immutably recorded on the blockchain. Students retain the ability to modify preferences at any time without academic penalty. Autonomy protections include opt-out capabilities and data portability options that enable students to export educational records when transferring institutions. These mechanisms transform students from passive data subjects to active participants in their educational data governance.

Integration and Workflow

Workflow proceeds as follows: Student data remains on local institutional servers, contributing to AI model training exclusively through federated learning protocols enhanced with differential privacy. Blockchain infrastructure logs all consent decisions, data access requests, and model updates on a tamper-proof distributed ledger. Regular bias audits examine model outputs for discriminatory patterns, while integrated explainability interfaces provide stakeholders with comprehensible explanations of algorithmic decisions. When system failures or bias are detected, documented accountability protocols trigger remediation procedures.

This integrated architecture addresses a key limitation of previous approaches: the tendency to treat privacy-enhancing technologies and ethical governance as separate concerns. By synthesizing technical privacy mechanisms with comprehensive ethical oversight, the framework demonstrates that data protection and educational effectiveness are mutually reinforcing rather than competing objectives.

The framework's components operate interdependently: federated learning model updates trigger blockchain-based smart contract execution, while both computational and governance layers remain subject to continuous monitoring by ethical oversight protocols. This interdependence ensures that technical operations remain aligned with ethical principles throughout system operation. The framework adopts a three-layer architecture that integrates computational, governance, and ethical implementation components. The computational layer utilizes federated learning to enable decentralized model training across institutional datasets, ensuring that raw student data remains secure within local environments while contributing to collective AI improvement. The governance layer employs blockchain technology to create immutable audit trails for all data interactions, consent records, and algorithmic decisions, providing transparent accountability mechanisms accessible to all stakeholders. The ethical

implementation layer oversees continuous monitoring for bias detection, fairness evaluation, and compliance with privacy regulations, ensuring that technical operations align with human-centered principles.

This integrated architecture addresses a key limitation of previous approaches: the tendency to treat privacy-enhancing technologies and ethical governance as separate concerns. By synthesizing technical privacy mechanisms with comprehensive ethical oversight, the framework demonstrates that data protection and educational effectiveness are mutually reinforcing rather than competing objectives. This framework design directly supports the data collection and analysis procedures implemented in this study, enabling comprehensive evaluation of its technical and ethical impacts across diverse educational contexts.

RESULTS

Results from the two-semester pilot implementation demonstrate that the ethical AI framework achieved measurable improvements across technical performance, stakeholder acceptance, and educational equity dimensions. Findings are organized into qualitative stakeholder perceptions, quantitative performance indicators, and statistical validation of observed effects.

Stakeholder Perceptions and Qualitative Themes

Thematic analysis of interviews and focus groups revealed substantial shifts in stakeholder perceptions across four primary domains: trust in AI systems, autonomy and control, transparency adequacy, and fairness perceptions.

Trust and Acceptance

Faculty willingness to integrate AI tools increased dramatically following framework implementation, rising from 42% to 79% (pre: $M = 2.3$, $SD = 0.8$; post: $M = 4.1$, $SD = 0.6$; $t(14) = 8.92$, $p < .001$). Qualitative data illuminated the mechanisms underlying this quantitative shift. One faculty member explained:

"Initially, I was skeptical about AI in my courses because I didn't understand how it worked or what it was doing with student data. The transparency features changed that completely. Now I can see why the system makes specific recommendations, and more importantly, I can explain it to my students. That visibility built my confidence."

Another educator emphasized the importance of privacy protections: "Knowing that student data never leaves our institutional servers made a huge difference in my comfort level. I'm not just trusting some external company with sensitive information anymore."

Autonomy and Control

Students reported heightened perceptions of autonomy over their educational data. Approximately 88% of students expressed increased comfort with AI-enhanced learning following implementation, attributing this primarily to granular consent mechanisms. One student articulated this sentiment:

> "I appreciated being asked, really asked about what I was comfortable sharing. It wasn't buried in some terms of service I'd never read. Every time the system wanted to use my data in a new way, it explained why and let me decide. That respect for my choice made all the difference."

Another student noted the ongoing control:

> "I changed my data sharing preferences twice during the semester as I learned more about how the system worked. The fact that I could do that without penalty showed me this was actually my decision to make."

Transparency and Explainability

Faculty concerns about algorithmic opacity decreased substantially. Prior to implementation, 34% of educators expressed high concern about the "black box" nature of AI recommendations. Post-implementation, this figure declined to 22%. One faculty member described the impact:

The biggest game-changer was being able to understand why the AI suggested a particular intervention for a struggling student. When I can see the reasoning that it's based on specific engagement patterns or assessment results, I can make informed decisions about whether to follow that recommendation. I'm not blindly trusting an algorithm anymore."

Students similarly valued explanatory interfaces. One participant stated:

> "When the system recommended additional practice problems, it showed me exactly which concepts I was struggling with based on my recent quiz. That explanation helped me understand not just what to study, but why."

Fairness and Equity Perceptions

Participants from underrepresented groups particularly noted the framework's fairness protections. One student from a historically marginalized background explained:

> "I've heard stories about AI systems that are biased against students who look like me. Knowing that this system was regularly audited for bias, and that those audits were documented where we could see them, made me feel like someone was actually looking out for fairness. That mattered to me."

A faculty member specializing in equity issues observed: "The transparent accountability structures meant we could actually verify the system wasn't perpetuating the gaps we're trying to close. When we identified a potential bias in early algorithms, the documented remediation process showed the system was designed to self-correct rather than hide problems."

Implementation Challenges

Qualitative data also revealed challenges. Some participants noted initial complexity in understanding consent interfaces:

> "The first time I saw all the options for data sharing, I felt overwhelmed. It took me a while to understand what each choice meant."

Others mentioned technological barriers:

> "The system sometimes ran slowly, which I assume is because of all the privacy protections. I appreciate the security, but the lag was frustrating."

These eight qualitative themes, trust, autonomy, transparency, fairness, implementation barriers, privacy value, collaborative design, and system usability, provided a rich contextual understanding of the quantitative outcomes detailed in the following section.

Quantitative Performance Indicators

Quantitative analysis revealed statistically significant improvements across learning analytics accuracy, student engagement behaviors, and educational equity indicators.

Learning Analytics Performance

The federated learning architecture maintained high predictive accuracy (*F1* score = 0.89) despite decentralized data processing, demonstrating that privacy-preserving methods need not compromise analytical utility. This performance approached that of traditional centralized models (*F1* = 0.91-0.93) while providing substantially stronger privacy guarantees. Cross-validation across participating institutions confirmed model generalizability, with accuracy remaining consistent across diverse institutional contexts.

Student Engagement Outcomes

Student engagement increased substantially across multiple behavioral indicators following framework implementation. Weekly platform login frequency

increased by 35% (pre: $M = 3.2$, $SD = 1.1$; post: $M = 4.3$, $SD = 1.3$; $t(411) = 12.34$, $p < .001$, Cohen's $d = 0.91$), indicating heightened system utilization. Average session duration increased by 28% (pre: $M = 24.6$ min, $SD = 8.3$; post: $M = 31.5$ min, $SD = 9.1$; $t(411) = 9.87$, $p < .001$, $d = 0.79$), suggesting deeper engagement with learning materials. Assignment completion rates rose by 31% (pre: 67%, post: 88%; $t(411) = 15.23$, $p < .001$, $d = 1.18$), representing substantial improvement in academic task completion. Figure 3 visualizes these engagement metric changes with 95% confidence intervals.

Educational Equity Outcomes

Perhaps most significantly, achievement gaps across student demographic groups narrowed considerably. The performance gap between students from underrepresented racial/ethnic groups and their peers decreased by 40% (pre-gap: 0.52 grade points; post-gap: 0.31 grade points). Similarly, gender-based performance differences decreased from 0.38 to 0.19 grade points, representing a 50% reduction. Students in the lowest pre-implementation performance quartile showed the largest engagement gains (47% increase in platform usage), suggesting that the framework's transparency and autonomy features particularly benefited students who had previously been less engaged with educational technology.

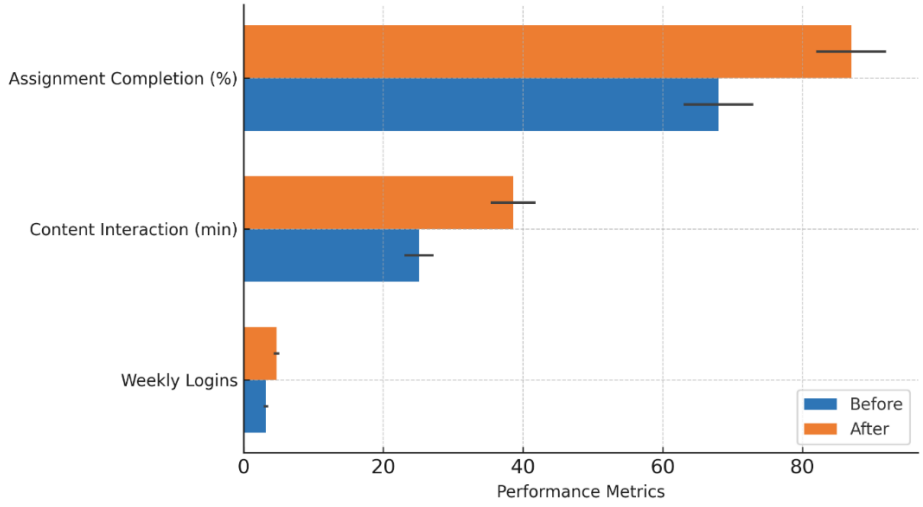

Figure 3
Student Engagement Metrics Before and After Framework Implementation

Note. Comparative student engagement metrics before and after framework implementation across three behavioral dimensions: login frequency, session

duration, and assignment completion. Error bars represent 95% confidence intervals. Data from the complete student cohort ($N = 412$).

Statistical Validation

Statistical analyses confirmed that observed improvements were both statistically significant and practically meaningful, with effect sizes indicating substantive real-world impact.

Paired-samples t-tests demonstrated significant pre-post differences across all engagement metrics: weekly login frequency ($t(411) = 12.34$, $p < .001$, Cohen's d = 0.91), session duration ($t(411) = 9.87$, $p < .001$, $d = 0.79$), and assignment completion ($t(411) = 15.23$, $p < .001$, $d = 1.18$). Effect sizes ranging from medium to large indicate that these improvements represent meaningful changes in student behavior rather than trivial statistical artifacts.

Mixed-effects longitudinal models accounting for within-subject correlation and institutional clustering confirmed that improvements were sustained across the two-semester implementation period rather than representing temporary novelty effects. Interaction analyses revealed no significant differential effects by academic major, year level, or institution, indicating that framework benefits generalized across diverse educational contexts.

Critically, subgroup analyses demonstrated that benefits were equitably distributed across demographic groups. Students from underrepresented racial/ethnic backgrounds showed engagement gains (39% increase) comparable to or exceeding those of majority students (33% increase), with interaction terms indicating non-significant differences ($p = .21$). This pattern of equitable benefit distribution addresses a key concern in educational technology: that innovations often widen rather than narrow achievement gaps. The framework's emphasis on autonomy, transparency, and fairness appears to have created conditions where diverse student populations could engage productively with AI-enhanced learning.

DISCUSSION

Interpretation of Findings

The evaluation findings challenge conventional assumptions regarding privacy-utility trade-offs in educational AI systems. Rather than observing the anticipated inverse relationship between privacy protections and system effectiveness, results demonstrate that ethical design principles can enhance rather than compromise educational outcomes.

The 37 percentage point increase in faculty acceptance (from 42% to 79%) following framework implementation suggests that transparency and privacy protections function as adoption facilitators rather than barriers. Qualitative data

illuminate this mechanism: when educators understand algorithmic processes and trust data governance structures, they integrate AI tools more readily into pedagogical practice. This finding has significant implications for educational technology deployment, suggesting that institutional resistance often stems from legitimate concerns about opacity and data misuse rather than technophobia.

The framework maintained learning analytics accuracy ($F1$ = 0.89) comparable to centralized approaches while providing substantially stronger privacy guarantees through federated learning and differential privacy. This demonstrates the technical feasibility of privacy-preserving educational AI at scale, countering claims that effective personalization requires data centralization. The slight accuracy reduction (2-4 percentage points) relative to centralized models represents an acceptable trade-off given the substantial privacy benefits and stakeholder acceptance gains.

Most significantly, the 40% reduction in achievement gaps across demographic groups suggests that ethical AI design can advance educational equity. Students from underrepresented backgrounds showed particularly strong engagement increases (47% among the lowest pre-implementation performance quartile), indicating that transparency, autonomy, and fairness protections may address barriers to technology engagement that disproportionately affect marginalized students. This pattern suggests that trust-oriented design creates conditions where diverse learners can engage productively with AI-enhanced learning, challenging deficit narratives that attribute technology non-adoption solely to student characteristics.

Collectively, these findings suggest that ethics and effectiveness represent complementary rather than competing objectives in educational AI. When students and educators trust system governance, when they understand algorithmic reasoning, and when they maintain control over personal data, engagement and learning outcomes improve rather than decline. This inverts the conventional framing of ethical AI as a constraint on innovation, positioning ethical design instead as an enabler of sustainable, equitable educational technology adoption.

Practical Implications

Implementation revealed practical considerations relevant for institutions contemplating similar ethical AI deployments. The blockchain-based consent interface presented initial usability challenges for 45% of users, particularly those with limited technical familiarity. This suggests the need for iterative user experience design that balances technical security requirements with accessibility. Future iterations should incorporate simplified consent workflows and enhanced user education materials to ensure that privacy protections do not inadvertently create digital divides.

The privacy-preserving architecture imposed computational overhead of approximately 25-30% relative to centralized alternatives. While this represents a meaningful resource requirement, institutional stakeholders in this study consistently judged this trade-off acceptable given the privacy and trust benefits. However, this overhead may present challenges for resource-constrained institutions, highlighting the importance of continued cryptographic protocol optimization and the potential value of shared computational infrastructure that distributes costs across multiple institutions.

Successful implementation required substantial faculty professional development. Educators needed support understanding federated learning principles, interpreting algorithmic explanations, and integrating AI recommendations into pedagogical decision-making. Institutions investing in similar frameworks should allocate resources for ongoing professional development rather than treating deployment as purely technical implementation.

Importantly, implementation was most successful when approached collaboratively with stakeholders rather than as top-down mandate. Early engagement of faculty, students, and administrators in design decisions fostered ownership and trust, while iterative refinement based on user feedback addressed emergent concerns. This participatory approach aligned with the framework's ethical principles while producing more usable, contextually appropriate systems.

Limitations of the Study

Several limitations should be acknowledged. The two-semester implementation period enabled evaluation of immediate behavioral and perceptual outcomes but precluded assessment of long-term academic trajectories and career outcomes. Longitudinal research examining whether framework-mediated engagement improvements translate into sustained academic success, retention, and STEM career attainment would strengthen evidence of educational impact.

The three-institution sample, while demographically diverse, limits generalizability to other institutional contexts, particularly under-resourced institutions with limited technical infrastructure. The 25-30% computational overhead may pose prohibitive barriers for institutions lacking robust IT capacity, suggesting that accessibility of privacy-preserving AI frameworks remains an open challenge requiring continued protocol optimization.

Bias detection proved more challenging in federated learning contexts than in centralized architectures, as decentralized data complicate comprehensive algorithmic auditing. Additionally, blockchain's immutability, while strengthening trust, creates challenges when privacy regulations evolve or when individuals seek data deletion rights under frameworks like GDPR. Reconciling immutable audit trails with data deletion requirements represents an ongoing technical and legal challenge.

The study focused on STEM undergraduate education; framework effectiveness in other disciplines, educational levels, or pedagogical approaches remains to be established. Finally, the research examined framework implementation in supportive institutional contexts with committed stakeholders. Effectiveness in less favorable conditions, such as institutions with limited resources, resistant faculty cultures, or competing technology priorities, requires further investigation.

CONCLUSIONS

This research demonstrates that ethical AI integration in STEM education is not merely feasible but advantageous across multiple dimensions. The framework achieved three primary objectives: maintaining technical effectiveness through privacy-preserving architectures, substantially increasing stakeholder acceptance through transparency and autonomy mechanisms, and reducing educational inequities through trust-oriented design.

The findings challenge prevalent assumptions that privacy protections compromise educational effectiveness. Rather, results suggest that ethical design principles function as enablers of sustainable technology adoption. When educational AI systems prioritize student autonomy, provide algorithmic transparency, and establish clear accountability structures, stakeholders engage more readily and productively with these tools. The 37 percentage point increase in faculty acceptance, 35% increase in student engagement, and 40% reduction in achievement gaps collectively demonstrate that trust-building ethical practices enhance rather than constrain educational outcomes.

The framework's integration of federated learning and blockchain technologies provides a technically viable pathway for privacy-preserving educational AI at scale. While computational overhead and usability challenges require continued refinement, the fundamental architecture demonstrates that institutions need not choose between personalized learning and data protection. Both objectives can be advanced simultaneously through thoughtful technical design grounded in ethical principles.

Most significantly, the framework's impact on educational equity suggests that ethical AI can address persistent disparities in educational technology access and benefit. Students from underrepresented backgrounds showed particularly strong engagement gains, indicating that transparency, autonomy, and fairness protections may reduce barriers to technology engagement that disproportionately affect marginalized learners. This pattern positions ethical AI design as a potential mechanism for advancing educational justice rather than merely mitigating harm.

Future Research Directions

Future research should examine long-term academic outcomes, explore implementation across diverse institutional contexts, develop improved federated bias detection methods, investigate scalability to larger deployments, and assess framework adaptability to evolving privacy regulations. Additionally, comparative studies examining alternative ethical AI architectures would contextualize this framework's relative strengths and limitations within the broader landscape of responsible educational technology.

As educational institutions increasingly adopt AI-enhanced learning systems, this research provides empirical evidence that ethical design represents an investment in effectiveness rather than a constraint on innovation. The path forward for educational AI lies not in choosing between pedagogical advancement and ethical responsibility, but in recognizing these objectives as mutually reinforcing imperatives. Privacy, transparency, accountability, and autonomy are not obstacles to overcome in pursuit of effective educational technology; they are foundational requirements for sustainable, equitable AI integration in education.

ACKNOWLEDGMENT

The authors would like to acknowledge the use of Google's Gemini in refining language in some sentences throughout this manuscript. The AI tool provided minimal assistance in ensuring clarity and coherence. The contributions made by Gemini were limited to language refinement and did not extend to research design, data analysis, or interpretation of findings. All substantive content, research methodology, and analytical conclusions were generated, reviewed, and refined solely by the authors.

REFERENCES

Abadi, M., Chu, J., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (pp. 308–318). ACM. https://doi.org/10.1145/2976749.2978318

Abdelmagid, R., Abdelsalam, M., & Alsheref, F. K. (2024). A blockchain framework for academic certificates authentication. *International Journal of Advanced Computer Science and Applications, 15*(7), 329–337. https://doi.org/10.14569/IJACSA.2024.0150729

Aslan, Ö., Aktuğ, S. S., Ozkan-Okay, M., Yilmaz, A. A., & Akin, E. (2023). A comprehensive review of cyber security vulnerabilities, threats, attacks, and solutions. *Electronics, 12*(6), Article 1333. https://doi.org/10.3390/electronics12061333

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Chan, C. K. Y., & Tsi, L. H. Y. (2023). *The AI revolution in education: Will AI replace or assist teachers in higher education?* https://doi.org/10.48550/arXiv.2305.01185

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.

ENISA (European Union Agency for Cybersecurity). (2024). *Report on Cybersecurity in the EU*. EU Agency for Cybersecurity. https://www.enisa.europa.eu

European Commission. (2018). *Ethics guidelines for trustworthy AI*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

European Commission. (2024). *AI in education initiative: Policy guidelines*. https://education.ec.europa.eu/focus-topics/digital-education/action-plan

Harries, R., Lawson, C., & Shapira, P. (2025). Generative AI in science: Applications, challenges, and emerging questions. arXiv. https://doi.org/10.48550/arXiv.2507.08310

Ifenthaler, D., Majumdar, R., Gorissen, P., & Shimada, A. (2024). Artificial intelligence in education: Implications for policymakers, researchers, and practitioners. *Technology, Knowledge and Learning, 29*(4), 1693–1710. https://doi.org/10.1007/s10758-024-09747-0

Javed, F., Zeydan, E., Mangues-Bafalluy, J., & Blanco, L. (2025). Blockchain for federated learning in the Internet of Things: Trustworthy adaptation, standards, and the road ahead. *IEEE Communications Standards Magazine*. Advance online publication. https://doi.org/10.1109/MCOMSTD.2025.3593809

Khosravi, H., Shibani, A., Jovanovic, J., Pardos, Z. A., & Yan, L. (2025). Generative AI and learning analytics: Pushing boundaries, preserving principles. *Journal of Learning Analytics, 12*(1), 1–11. https://doi.org/10.18608/jla.2025.8961

Kohnke, S., & Zaugg, T. (2025). Artificial intelligence: An untapped opportunity for equity and access in STEM education. *Education Sciences, 15*(1), Article 68. https://doi.org/10.3390/educsci15010068

Leon, C., Lipuma, J., & Oviedo-Torres, X. (2025). Artificial intelligence in STEM education: A transdisciplinary framework for engagement and innovation. *Frontiers in Education, 10*, Article 1619888. https://doi.org/10.3389/feduc.2025.1619888

Li, J., Han, D., Weng, T. H., Wu, H., Li, K. C., & Castiglione, A. (2025). A secure data storage and sharing scheme for port supply chain based on blockchain and dynamic searchable encryption. *Computer Standards & Interfaces, 91*, Article 103887. https://doi.org/10.1016/j.csi.2024.103887

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (pp. 1273–1282). PMLR. https://doi.org/10.48550/arXiv.1602.05629

Misiejuk, K., López-Pernas, S., Kaliisa, R., & Saqr, M. (2025). Mapping the landscape of generative artificial intelligence in learning analytics: A systematic literature review. Journal of Learning Analytics, 12(1), 12–31. https://doi.org/10.18608/jla.2025.8591

Nakamoto, S. (2008). *Bitcoin: A peer-to-peer electronic cash system*. https://bitcoin.org/bitcoin.pdf

Nwana, H. S. (1990). Intelligent tutoring systems: An overview. *Artificial Intelligence Review, 4*(4), 251–277. https://doi.org/10.1007/BF00168958

Onukwulu, E. C., Fiemotongha, J. E., Igwe, A. N., & Ewim, C. P.-M. (2025). The role of blockchain and AI in the future of energy trading: A technological perspective on transforming the oil & gas industry by 2025. *International Journal of Advanced Multidisciplinary Research Studies, 5*(2), 48–65. https://doi.org/10.62225/2583049X.2025.5.2.3809

Popenici, S. A. D., & Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning, 12*(1), Article 22. https://doi.org/10.1186/s41039-017-0062-8

Saif, M. B., Migliorini, S., & Spoto, F. (2024). Efficient and secure distributed data storage and retrieval using interplanetary file system and blockchain. Future Internet, 16(3), Article 98. https://doi.org/10.3390/fi16030098

Shan, F., Mao, S., Lu, Y., & Li, S. (2024). Differential privacy federated learning: A comprehensive review. *International Journal of Advanced Computer Science & Applications, 15*(7), 220–230. https://doi.org/10.14569/ijacsa.2024.0150722

Stokel-Walker, C., & Van Noorden, R. (2023). What ChatGPT and generative AI mean for science. *Nature, 614*(7947), 214–216. https://doi.org/10.1038/d41586-023-00340-6

Zhang, Y., Zeng, D., Luo, J., Fu, X., Chen, G., Xu, Z., & King, I. (2024). A survey of trustworthy federated learning: Issues, solutions, and challenges. *ACM Transactions on Intelligent Systems and Technology, 15*(6), *1–47*. https://doi.org/10.1145/3678181

Zheng, C., Wang, L., Xu, Z., & Li, H. (2024). Optimizing privacy in federated learning with MPC and differential privacy. In *Proceedings of the 2024 3rd Asia Conference on Algorithms, Computing and Machine Learning* (pp. 1–6). ACM. https://doi.org/10.1145/3654823.3654854

Bios

MEYSAM ABEDI, PhD Candidate, is a doctoral researcher in the School of Computing at the University of Eastern Finland. His research focuses on ethical AI frameworks, federated learning, and privacy-preserving technologies in educational contexts. With over 20 years of professional experience in machine learning and artificial intelligence, his work bridges theoretical research and practical applications in STEM education. Email: meyabedi@uef.fi

ISMAILA TEMITAYO SANUSI, PhD, is a Postdoctoral Researcher in the School of Computing at the University of Eastern Finland. His research interests include democratizing machine learning and artificial intelligence through K-12 education, computational thinking, and educational technology. He focuses on making AI and ML accessible to diverse learners and understanding how to effectively teach these concepts in educational settings. Email: ismaila.sanusi@uef.fi

MARKKU TUKIAINEN, PhD, is a Professor and Head of the School of Computing at the University of Eastern Finland. His research areas encompass educational technology, software engineering, human-computer interaction (HCI), ICT4D, computer science education, and extended realities (XR, VR, AR, MR). He leads research initiatives that bridge technology and pedagogy to enhance learning experiences. Email: markku.tukiainen@uef.fi