A Call for Fully Open and Trustworthy Generative AI for Education

Anonymous ACL submission

Abstract

The rapid integration of generative artificial intelligence (GenAI), particularly large language models (LLMs), is significantly transforming educational practices, impacting diverse areas from intelligent tutoring to curriculum design and assessment. Current GenAI applications in education predominantly rely on closed-source, commercial models, raising critical concerns related to transparency, data privacy, interpretability, and trustworthiness. In contrast, open-source models offer enhanced transparency and foster community-driven development, yet remain underutilized in educational research and applications. This position paper argues for **a strategic shift toward fully open-source and trustworthy GenAI in education**. We highlight the benefits of transparency, reproducibility, equitable accessibility, and auditability provided by open-source solutions, address practical considerations and challenges, and propose actionable recommendations and policy initiatives to encourage collaborative development among educators, researchers, and policymakers.

1 Introduction

Recent years have witnessed the rapid emergence of powerful generative artificial intelligence (GenAI), including visual generative models (e.g., Imagen (Saharia et al., 2022), Parti (Yu et al., 2022), DALL-E (Ramesh et al., 2021, 2022), and Stable Diffusion (Rombach et al., 2022)), large language models (LLMs) such as GPT-series (OpenAI, 2023), Gemini (Anil et al., 2023), LLaMA (Dubey et al., 2024), and DeepSeek (DeepSeek-AI et al., 2025), and multi-modal large language models (MLLMs) including LLaVA (Liu et al., 2023), BLIP-2 (Li et al., 2023a), and GPT-4o (Hurst et al., 2024). Among these, LLMs have demonstrated particularly impressive capabilities, including understanding textual context, processing complex prompts, reasoning over information, and generating highly coherent, contextually relevant textual outputs (Chan and Hu, 2023; Akpan et al., 2025).

The rapid development of these powerful LLM models has led to widespread adoption across numerous domains, from healthcare (Wang et al., 2023b; Goyal et al., 2024), finance (Li et al., 2023b; de Zarzà et al., 2023), scientific discoveries (Zhang et al., 2024c; Reddy and Shojaee, 2025), up to transportation (Da et al., 2025; Zheng et al., 2023), etc. In particular, education has significantly benefited from the integration of GenAI technologies (Wang et al., 2024; Jeon and Lee, 2023), driven by their potential to fundamentally transform teaching methodologies, curriculum design, student learning and educational research (Ayeni et al., 2024; Jin et al., 2024; Song et al., 2024). As educators and learners increasingly embrace GenAI models, GenAI promises transformative potential by delivering personalized, scalable, and interactive learning experiences, fundamentally reshaping both educational content delivery and accessibility.

Status Quo. Since 2023, academic research has highlighted the increasing prominence of GenAI applications across diverse educational scenarios. Recent literature identifies several primary use-cases across multiple dimensions. (1) **As a tutor/assistant**: GenAI can be used as a tutor or teaching assistant, providing tips and hints to help students solve challenging problems in lieu of the time-consuming and error-prone manually authored tips (Jia et al., 2024; Prihar et al., 2023; Urrutia and Araya, 2024). (2) **For curriculum development**: Using ChatGPT to prepare lesson plans in primary school courses is proved to be effective in academic achievement (Karaman and İdris Göksu, 2024). Teachers may consider combining ChatGPT's plans and theirs and benefiting from both in the implementation process (Malik et al., 2024; Rizos et al., 2024). (3) **For enhancing student achievement**: Empirical studies found

that GenAI use correlates with improved student achievement (Asare et al., 2023), and identified a statistically significant positive moderation effect of students' interest on the association between ChatGPT usage and academic achievement (Yunianto et al., 2024; Urban et al., 2024). (4) **For model development**: Researchers have developed models (e.g. EBERT (Duan et al., 2024)) specialized in downstream education tasks, and models like this can outperform former state-of-the-art baselines such as MathBERT (Shen et al., 2021) or GPT-3 (Brown et al., 2020) on accuracy metrics (Xu et al., 2024; Zhang et al., 2024a). (5) **Evaluation of GenAI ability**: Recent research explores methodologies aimed at fostering a synergistic relationship between GenAI and human intelligence, encouraging the students' development of robust problem-solving strategies while harnessing the computational capabilities of AI (Getenet, 2024; Rane, 2023; Urhan et al., 2024).

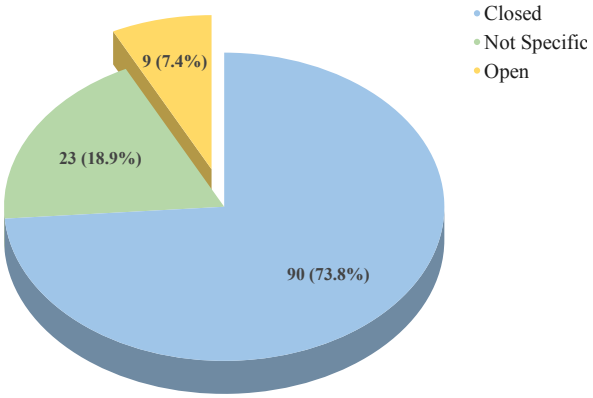


Figure 1: Open vs. Closed-Source LLMs in Education

Motivation. Despite these advancements, critical limitations remain prevalent in the current GenAI-for-education landscape. Primarily, the sector is heavily dominated by proprietary (closed-source) models such as ChatGPT, developed and controlled by large commercial entities. Although commercial models benefit from centralized resources, rapid iterative cycles, and widespread market penetration, they inherently lack transparency and openness (Liesenfeld et al., 2023; Aljanabi et al., 2023). This opacity introduces significant challenges regarding model *interpretability*, *data privacy*, *equitable accessibility*, intellectual property rights, and accountability in educational contexts (Kohnke and Zaugg, 2025). In contrast, open-source GenAI models—such as the LLaMA series and the recently emerged DeepSeek-R1 (DeepSeek-AI et al., 2025)—provide greater transparency, reproducibility, and community-driven innovation, democratizing access to advanced educational AI tools. Indeed, recent breakthroughs such as DeepSeek-R1 illustrate that open-source alternatives can now match or exceed the performance and capabilities of commercial models (DeepSeek-AI et al., 2025). However, our comprehensive review of recent educational GenAI literature (detailed in Sec. 2) reveals that, in the experimental studies, approximately 73.8% of closed source compared to only 7.4% employing open-source in education field(See Figure 1). Furthermore, substantially fewer research initiatives have pursued the development of specialized, education-centric open-source GenAI models, with even fewer openly releasing their models, code-bases, or data, highlighting a critical gap and missed opportunity within the educational research community (Zhao et al., 2024).

Our Position. In this position paper, **we advocate that future generative AI models and platforms for education should focus more on fully open-source and trustworthy resources.** We argue that adopting fully open-source GenAI will enhance transparency, encourage equitable access, and significantly improve accountability, addressing many existing challenges faced by proprietary solutions. Open-source GenAI fosters extensive collaboration, accelerating innovation, community engagement, reproducibility, and rigorous auditing—critical components for achieving trustworthiness in educational settings. In the following sections, we first provide a comprehensive review of prior works relevant to GenAI for education in Sec. 2, and present our position that open-source GenAI platform for education is promising in Sec. 3, then present our initial explorations of such an open source platform in Sec. 4. Finally, we conclude with actionable recommendations and policy proposals in Sec. 5, aiming to catalyze a shift toward open and trustworthy GenAI in education.

2 A Review of Prior Works

GenAI for Education. We conducted a systematic and comprehensive review of recent literature on GenAI and LLMs in educational contexts, following a structured methodology adapted from established guidelines outlined by Chandler et al. (Chandler et al., 2019). Our review encompasses a total of 210 peer-reviewed articles published in English from 2023 to May 2025, comprising 54 review articles and 122 experimental studies. These articles

were identified using targeted keyword searches (i.e., "LLM in education" and "GenAI education"), followed by rigorous screening and duplicate removal.

Our analysis indicates that, among the 122 experimental studies, only nine explicitly utilize fully open-source models (Jia et al., 2024; Duan et al., 2024; Zhang et al., 2024a,b; Anh and Hung, 2024; Dehbozorgi et al., 2024; Latif et al., 2024; Morris et al., 2024; Nakamoto et al., 2023) (see Figure 1). The predominance of closed-source, commercial GenAI models underscores a significant reliance on proprietary technologies such as ChatGPT, Gemini, and DeepSeek (DeepSeek-AI et al., 2025), which are predominantly developed in technologically advanced nations including the United States, the United Kingdom, and China (see Figure 2). This geographic concentration of GenAI research and development creates barriers for researchers and educators from lower- and middle-income countries, who often face limited access to costly commercial models and proprietary resources. Consequently, such disparities exacerbate global inequities in educational innovation and access to advanced educational technologies.

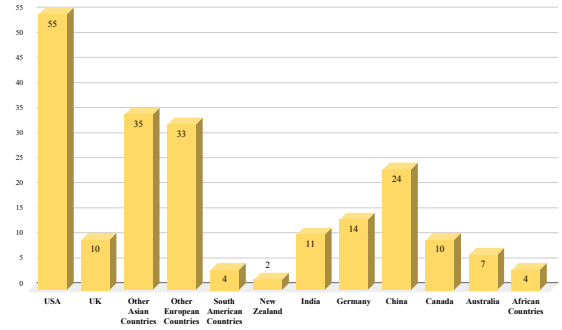


Figure 2: Geographical locations of published papers

Open-Source vs. Closed-Source LLMs. Existing LLMs can be categorized based on their openness into three primary types (see Figure 3): (1) **Closed-source (proprietary)** models, which are developed by commercial entities and restrict public access to their architectures, training data, and internal parameters; (2) **Open-weight** models, where trained model parameters (weights) are publicly available but the training code or data remains inaccessible; (3) **Fully open-source** models, characterized by publicly accessible architectures, training scripts, datasets, and model parameters; and (4) **Open-development** models whose contributable process is also shared.

Closed-source models, exemplified by GPT-4 (Hurst et al., 2024) and Gemini (Reid et al., 2024), offer advanced capabilities and rapid development cycles facilitated by significant commercial resources. However, their proprietary nature inherently restricts transparency, limits interpretability, and complicates customization for specialized educational needs. These "black-box" limitations pose substantial challenges for educational stakeholders in terms of ethical accountability, data privacy, fairness, and equity.

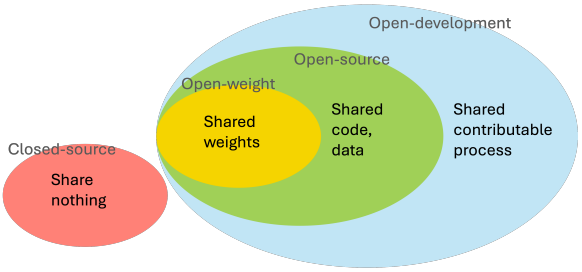


Figure 3: Definition of Open-Sourceness

Furthermore, the reliance on closed-source GenAI introduces significant economic and operational barriers for educational institutions in resource-constrained regions, restricting their ability to fully harness the benefits of advanced AI tools. Proprietary licensing and costly access fees often exclude educators in developing nations, perpetuating educational disparities. Moreover, the lack of transparency inherent in closed-source models complicates thorough assessments of model reliability, bias, and fairness, critical factors in educational settings.

Given these challenges, a strategic pivot toward fully open-source GenAI models is necessary to democratize access, foster transparency, and enable rigorous community-driven improvements. We posit that transitioning to open-source frameworks is not only ethically prudent but also essential for equitable and sustainable educational innovation globally. In the subsequent section, we articulate our detailed position advocating for a fully open-source and trustworthy GenAI approach to education, outlining practical pathways and addressing anticipated challenges.

3 Position

As discussed in Sec. 1, the prevailing dominance of closed-source, proprietary GenAI models significantly limits transparency, equity, and accountability in educational applications. To this end, we propose a fundamental shift towards the development and adoption of fully open-source GenAI

models explicitly tailored for educational purposes.

3.1 Why Focus on Open-Source GenAI for Education?

We argue for a strategic emphasis on open-source GenAI based on four critical dimensions:

Reproducibility (Data, Model, and Training). Reproducibility is essential for advancing scientific knowledge and validating educational methodologies. Open-source GenAI enables researchers and educators worldwide to reproduce experiments, verify results, and build upon prior work effectively. With unrestricted access to data, training scripts, and model checkpoints, the educational community can rigorously assess model robustness and performance across diverse learning contexts, thus promoting consistency and reliability in educational AI deployments (Fedele et al., 2024).

Accessibility (Economic Equity and Equitable Resource Distribution). The economic disparities between developed and developing regions often restrict equitable access to proprietary AI technologies, exacerbating existing educational inequalities (Bulathwela et al., 2024). Fully open-source models mitigate these financial barriers by eliminating licensing fees and enabling cost-effective implementation, thereby ensuring wider, equitable participation. Particularly for rural and economically disadvantaged areas, open-source solutions can facilitate significant improvements in educational quality and accessibility, leveling the playing field for underrepresented communities.

Transparency and Interpretability. Educational settings require a high degree of transparency and interpretability, ensuring that AI-driven educational tools are ethically aligned, unbiased, and comprehensible (Peñalvo et al., 2024). Unlike closed-source solutions that operate as opaque "black boxes", open-source GenAI models provide transparent insights into their architecture and decision-making processes (Radanliev, 2025). This transparency enables educators, learners, and stakeholders to better understand and address potential biases, promoting fairness and enhancing the ethical deployment of AI in education.

Trustworthiness (Safety, Security, and Privacy). Trust is fundamental in educational contexts, requiring AI systems to be safe, secure, and respectful of privacy (Bourg et al., 2024). Open-source models offer distinct advantages in terms of security and privacy by allowing comprehensive audits and community-driven security enhancements (Bildirici et al., 2024). Stakeholders can collaboratively identify vulnerabilities, promptly rectify security issues, and enforce rigorous privacy protections. Consequently, open-source GenAI fosters a more secure and trusted educational environment, safeguarding students' data and ensuring ethical AI use.

Therefore, embracing fully open-source GenAI models for education not only addresses critical limitations posed by proprietary systems but also fosters an equitable, transparent, and trustworthy educational landscape. The subsequent sections detail practical considerations, potential challenges, and actionable recommendations to realize this strategic shift.

3.2 Preparedness for Open-Source GenAI Platforms for Education

In this subsection, we evaluate the feasibility and infrastructure readiness for developing an open-source GenAI platform powered by LLMs specifically for educational purposes.

Existing Open-Source LLMs. Prominent open-source LLMs, including the Meta LLaMA series (e.g., LLaMA 3 (Dubey et al., 2024)), InternVL (Chen et al., 2024c,b), and Mistral (Jiang et al., 2023), demonstrate maturity, adaptability, and inherent openness required for domain-specific fine-tuning which is difficult with closed-source alternatives. These open-source models offer diverse architectures (e.g., sparse Mixture-of-Experts) and various parameter sizes (e.g., 8B to 70B), and their transparency and flexibility are crucial for research, development, and customization, especially for domain-specific fine-tuning in areas like education.

Educational Dataset Availability. Recently, researchers have made efforts in significantly expanding educational datasets across various learning domains, thus providing abundant data for fine-tuning open-source LLMs to enhance their efficacy in educational applications. Taking mathematics as a specific example, relevant teaching data may include various specialized dialogue datasets, such as real-world teacher-student interactions (e.g., Multi-turn Classroom Dialogue Dataset (Chen et al., 2024a)), real-world student interactions with an LLM-powered tutoring chat-bot (e.g., StudyChat Dataset (McNichols and Lan, 2025)), and tutoring conversations annotated or simulated by LLMs (e.g., MathDial Dataset (Macina et al., 2023) and Education Dialogue Dataset (Shani et al., 2024)).

Cost-Effective Cloud GPU Platform. Training

and fine-tuning large language models necessitate substantial computational resources, particularly high-end Graphics Processing Units (GPUs) with significant memory. While traditional cloud service providers such as AWS, Google Cloud, or Azure often entail considerable expense, several cost-effective GPU cloud platforms currently offer competitive rates for rental services, making previously unattainable projects financially viable for smaller teams and academic researchers with limited resources.

Application Deployment Infrastructure. The deployment and operationalization of GenAI educational tools are now supported by a wide range of robust platforms such as Vercel, Render, and Netlify. These services facilitate seamless backend integration, rapid prototyping, and scalable deployment of interactive learning systems, enabling educators and developers to bring AI-powered tools directly to classrooms and learners.

School and Societal Preparedness. We also assess the preparedness of educational stakeholders (Wang et al., 2023a) and the broader societal context, encompassing ethical considerations, data privacy, and public perception. While initial perceptions of AI in education among teachers and the public may have been characterized by apprehension and a focus on risks such as cheating or declined creativeness (Nazaretsky et al., 2022), recent trends indicate a growing recognition of AI's positive potential, and a notable shift towards greater awareness, acceptance, and proactive integration of AI technologies within educational settings (Kim and Kwon, 2023). Many forward-thinking schools move beyond initial hesitation to actively explore and implement AI application to education and integrating tools into their teaching practice (Xing et al., 2025). Although GenAI has brought challenges, a vast majority of students report using AI technologies in their academic activities (Vieriu and Petrea, 2025), among which virtual assistants and AI-based educational platforms are particularly popular. Additionally, despite concerns about data collection and misinformation (Yu et al., 2024), nowadays parents generally exhibit cautious optimism regarding the integration of AI in schools. They acknowledge GenAI's potential benefits in education and remain optimistic about using GenAI in academic assistance and research.

Taken together, these developments affirm the technical and institutional feasibility of building open-source GenAI platforms for education. This preparedness paves the way for scalable, equitable, and trustworthy AI integration in education, reinforcing the urgency and viability of the open-source agenda advanced in this paper.

3.3 Roadmap to Implementing a Fully Open-Source GenAI Platform

Realizing a fully open-source GenAI platform for education requires coordinated interdisciplinary action. This section outlines practical steps and roles for education researchers, computer scientists, and institutional stakeholders to co-develop a trustworthy, scalable, and pedagogically sound AI system.

Interdisciplinary Collaboration and Joint Development. Effective deployment begins with building sustained partnerships between technical and educational communities. We advocate for cross-disciplinary groups comprising computer scientists, AI engineers, education researchers, and practicing educators. These teams should engage in joint model design, dataset curation, iterative testing, and pedagogical alignment. Formal structures—such as collaborative research grants, co-development labs, and recurring workshops—can institutionalize this collaboration and ensure ongoing feedback exchange (Lan and Zhou, 2025).

Responsibilities of Key Stakeholders. *Education researchers* are responsible for defining pedagogical use cases, identifying key learning objectives, and collecting representative classroom data. Their work includes annotating instructional dialogues, modeling student misconceptions, and empirically evaluating system efficacy through classroom trials (Lan and Zhou, 2025). *Computer scientists and engineers* will lead open-source LLM development, fine-tuning, and deployment (Varghese and Ramasamy, 2024). Their responsibilities include selecting appropriate architectures, implementing scalable training and serving pipelines, and ensuring system robustness, efficiency, and privacy compliance (Varghese and Ramasamy, 2024). Together, they must co-design interfaces and feedback mechanisms to support teacher-facing and learner-facing applications (Nyaaba et al., 2024).

Technical Infrastructure. The platform must be supported by robust AI infrastructure (Varghese and Ramasamy, 2024). Training should be distributed across open-weight models (e.g., LLaMA, Mistral, DeepSeek-R1), leveraging cost-effective GPU resources and academic compute clusters. Model serving must include modular deployment pipelines, enabling localized adaptation, multilin-

gual support, and continuous monitoring. Tooling for model versioning, reproducibility, and bias auditing should be integrated from the outset.

Deployment and Maintenance. Beyond development, serving the model to educators and students requires intuitive interfaces and accessible deployment solutions (Nyaaba et al., 2024; Tang et al., 2025). Cloud-based backends (e.g., Render, HuggingFace Spaces) should be leveraged to provide browser-based, low-latency access. Open APIs must be maintained to support plug-and-play integration with existing learning management systems and teacher platforms. Continuous updates, community documentation, and open governance are essential for sustainability.

Building a fully open-source GenAI platform for education is technically feasible and institutionally necessary, but requires deliberate, collaborative, and inclusive development. This roadmap calls on both the machine learning and education communities to jointly invest in shared infrastructure, open standards, and empirical best practices to ensure that the future of educational AI is transparent, equitable, and aligned with public interest.

3.4 Key Considerations for Developing a GenAI Platform

In designing and deploying a GenAI platform for education, several foundational principles must guide development. These considerations not only ensure technical and pedagogical soundness but also underscore why open-source models are better suited for building responsible, equitable, and sustainable AI systems in education.

Reproducibility. Reproducibility is a cornerstone of academic integrity and scientific progress. In education, where empirical validation is essential, researchers must be able to replicate results, scrutinize model behavior, and extend prior work with confidence. Open-source LLMs provide access to training code, datasets, and model checkpoints, enabling rigorous peer review and longitudinal evaluation. By contrast, closed-source models obstruct this process, limiting transparency and hindering the accumulation of shared knowledge across institutions and contexts (Zhao et al., 2024).

Accessibility. Widening access to advanced GenAI tools is imperative for reducing educational inequality. Open-source models typically eliminate licensing fees and reduce the financial burden on under-resourced institutions and educators. In addition, open platforms should be paired with user-friendly interfaces, multilingual capabilities, and clear documentation to minimize technical barriers, facilitating adoption in diverse settings like underserved or rural communities.

Transparency. Transparent and interpretable AI is critical for ensuring ethical, fair, and inclusive educational outcomes. Unlike proprietary models, which obscure internal mechanisms, open-source systems allow inspection of training and decision-making processes. This is especially important in high-stakes environments like education, where algorithmic opacity can perpetuate biases or undermine trust. Open systems empower multidisciplinary oversight—engaging ethicists, educators, and policy makers in assessing fairness, interpretability, and accountability across student populations (Ratner et al., 2025; Deric et al., 2025; Quince et al., 2024; Zlotnikova et al., 2025).

Trustworthiness. Establishing trust is foundational when integrating GenAI into education (Joubin, 2024; Henderson et al., 2025). Trustworthiness in this context encompasses a multidimensional evaluation that includes factual accuracy, safety against adversarial manipulation, robustness to unseen conditions, privacy preservation, and fairness across diverse user groups (Xing et al., 2024). These dimensions, outlined in the AutoTrust benchmark, are equally critical in classrooms, where AI should ensure safe, ethical, and equitable interactions.

Factuality and Uncertainty Awareness. Educational GenAI must avoid hallucinations or misleading content (Elsayed, 2024; Chiang, 2024). Models should provide accurate answers and signal uncertainty when relevant, to prevent overconfident misinformation—a critical requirement in high-stakes learning contexts. Open-source frameworks offer the necessary transparency to audit and refine these behaviors systematically.

Safety and Adversarial Robustness. GenAI platforms for education are susceptible to misuse via adversarial prompts inducing inappropriate or off-topic responses. Drawing from adversarial testing in vision-language models (VLMs) for autonomous systems (Xing et al., 2024), developers of educational GenAI must implement safeguards against both benign errors and malicious manipulations. Open-source development facilitates community-driven discovery and mitigation of such vulnerabilities, a process hindered in proprietary systems.

Privacy Preservation. Systems interacting with sensitive student data (e.g., learning disabilities,

behavioral feedback) must protect privacy and prevent leakage. The AutoTrust evaluation shows that even advanced models can be coaxed into revealing private details unless explicitly trained to resist such prompts. Open-source architectures allow custom privacy controls and transparency audits—essential for compliance with data protection regulations like FERPA (Alshugran and Kloub, 2025; Farmer et al., 2024).

Fairness and Bias Mitigation. AI must treat students equitably across demographic lines (Rouabhia, 2025). Educational AI must be validated to ensure consistent performance across student populations—regardless of language, socioeconomic status, or learning style. Open models permit targeted bias analysis and correction, allowing stakeholders to collaboratively enforce fairness.

In summary, trustworthiness in educational GenAI is a composite of verifiability, security, resilience, and ethics. Open-source models offer a powerful substrate for achieving these goals, as they invite scrutiny, enable adaptation, and promote accountability—attributes obscured in proprietary solutions. Adopting a trustworthiness-first design philosophy safeguards learners and enhances GenAI's reliability and social value.

3.5 Incorporating Feedback to Make GenAI-Education Platform Better

To effectively integrate a GenAI platform in education, we propose iterative improvements based on educator and practitioner feedback, ensuring the platform adapts to individual needs and continuously benefits educators (McTear, 2022).

The platform generates diverse professional development training materials for instructors and users, including detailed documentation, training videos, user guides, booklets, and an accessible support website (Pelton and Pelton, 2023). Before full implementation, education experts pilot-test it to provide structured feedback on content accuracy, pedagogical effectiveness, and user experience, guiding adjustments to training methods and interface design. Then, large-scale trials will expand to local partnership universities, collecting quantitative data and qualitative feedback via interviews from educators to optimize scalability and localization features. This input will guide model development, emphasizing local applicability and broader usability. Following regional validation, the system can be publicized to broader universities across the US and up to the global education networks. Continuous feedback collection and maintenance are crucial for enhancing practicality and addressing ethical considerations, such as diverse cultural understanding. Local experts will establish regular review cycles, analyzing user behavior and survey responses to iteratively improve instructional relevance and technological integration.

4 Our Pilot Study

Our team has made initial progress in the development of an innovative and fully open-source educational GenAI platform, training and deployment of which are both transparent. It is designed to employ an LLM to simulate responses from a student struggling with mathematics, thus enabling interaction between prospective secondary mathematics teachers (PSMTs) and simulated students. We have deployed a trial AI-powered chatbot on Render, and our basic user interface (UI) design and major features are showcased below (See Figure 4).

The LLMs, specifically LLaMA 3 model (Dubey et al., 2024) fine-tuned on multi-turn teaching conversation data from different grade level mathematics classes, are the most essential GenAI component in our platform. Our aim is to eventually develop a chat-bot that can simulate reasonable responses to math problems or teachers' instructions, and contribute to conversations just as a middle school student who is experiencing difficulties in mathematics and stuck in specific questions. Tasked with facilitating and address the queries and inquiries from students in need of mathematics learning help, users (i.e. PSMTs) can effectively practice their teaching skills, identify common student errors and misconceptions, and refine their pedagogical approaches in real-time.

While LLMs' effective mathematics instructional and tutoring ability is still questioned nowadays (Gupta et al., 2025), by casting the GenAI in the role of a student struggling with math, we actually shift the onus of correctness and instructional quality back to the human teacher, navigating its current limitations in consistently providing accurate feedback. Also, to address deficits of the current LLM and decrease the possibility of outputs inconsistent with student roles, we plan an upcoming survey with 15-20 pre-service teachers and instructors, to let them participate in the following testing process and identify cases where the GenAI model fail to simulate well.

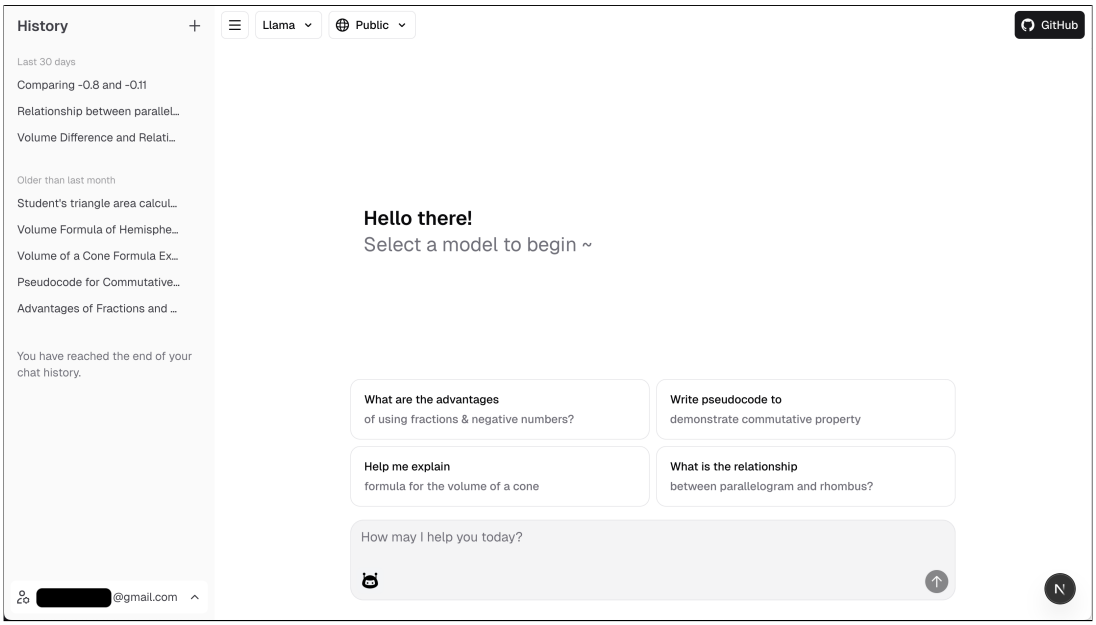
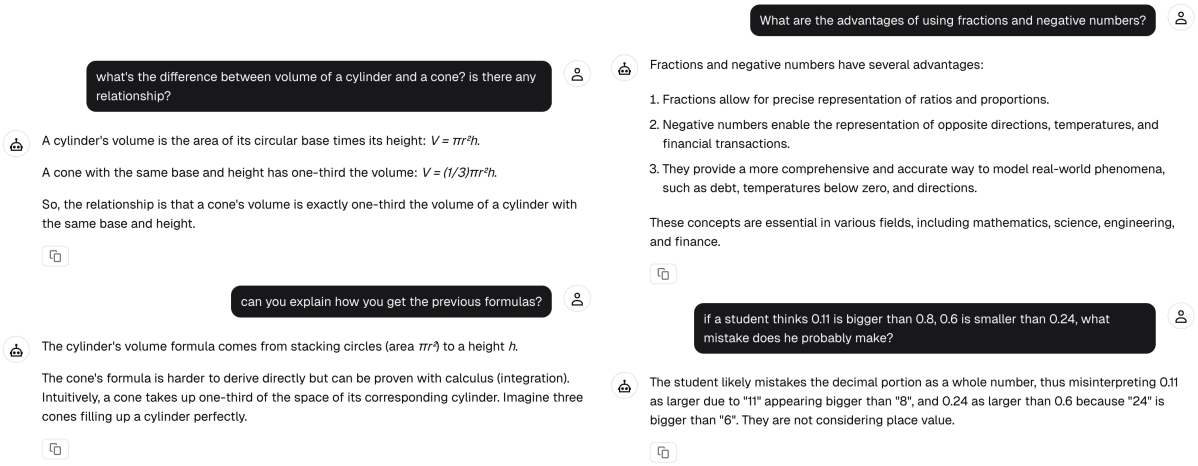


Figure 4: Our User Interface Design



(a) Problem 1

(b) Problem 2

Figure 5: Chat History between Mathematics Teachers and the Chat-Bot

5 Conclusion

We call for more open, inclusive, efficient, and trustworthy GenAI applications in education. Fully adopting open-source GenAI can enhance transparency, ensure fair access, and improve accountability—addressing many issues seen in proprietary systems. Open-source models foster collaboration, innovation, reproducibility, and rigorous peer review, all essential for credibility in education. We offer practical suggestions and policy recommendations for educators, researchers, developers, and policymakers, including joint research, co-creation, interdisciplinary events, standard data frameworks, and collaborative human–machine interaction. These strategies can help align AI advancements with effective teaching practices, supporting a more inclusive and impactful educational ecosystem.

Future Directions. Moving forward, we aim to extend our pilot system into a broader ecosystem that supports a range of instructional roles, from AI student simulators to intelligent teaching assistants. Further work is needed to design community-driven benchmarks that evaluate GenAI models for pedagogical soundness, fairness, and robustness in diverse educational settings. To support real-world adoption, we also emphasize the importance of integrating these systems into institutional infrastructures, aligning with educational standards and privacy policies. These directions will help realize an educational AI landscape that is not only technically advanced, but also open, inclusive, and grounded in public interest.

6 Limitations

While this paper advocates for the development of open-source GenAI platforms in education, we still acknowledge several limitations inherent to current open-source models and their adoption, in contrast with the compelling advantages offered by current closed-source commercial models. It is not controversial that, at least at present, existing open-source models may not always match or compete with the performance and convenience of their commercial counterparts (Chen et al., 2024b). Besides the fact that the key performance indicators of open source models are close to but have not yet exceeded closed-source commercial models, the closed-source (proprietary) models offer several other benefits that contribute to their widespread adoption. (1) **Ease of Use**: Commercial models, such as those from OpenAI (Hurst et al., 2024) and Google (Reid et al., 2024), are typically provided as managed services accessible through user-friendly web interfaces or well-documented Application Programming Interfaces (APIs), lowering technical barriers for educators and enabling rapid integration without technical expertise in model deployment or infrastructure management; (2) **Service Support**: Users of closed-source models avoid the complex and resource-intensive process of pre-training or extensive fine-tuning, as commercial providers handle continuous updates, maintenance, and even custom solutions tailored to specific industry (including education) needs, allowing quick responses to emerging demands—capabilities not yet standardized in open-source communities.

The current commercial model is inevitably the mainstream choice for educational institutions due to its substantial convenience, immediate utility, complete functional chain, and reliability; while it is still undeniable that open source models have great market potential because of better performance in data privacy protection and localized customization, and they will become an indispensable market trend for future educational applications.

Our current study is also limited, focusing on secondary-level mathematics without validation in other domains. The publicly available data we use may under-represent certain curricula, classroom discourse styles, and learner demographics, probably introducing distribution shifts and fairness concerns. Also, the open-source LLMs used in our prototype still lag behind top proprietary models on math reasoning and dialogue robustness, which affects the fidelity of student simulation. As for evaluation, our pilot is small-scale and user-study results with pre-service teachers are pending, so claims about training effectiveness and classroom impact still remain preliminary. And our evaluation primarily uses task- and dialogue-level metrics; measures for long-term learning outcomes are out of scope for this submission and will be addressed in follow-up work. Future research should empirically quantify these gaps through larger-scale evaluations to accelerate open-source advancements.

7 Ethical Considerations

Our proposed open-source GenAI approach prioritizes ethical deployment in education. All data used for model fine-tuning and training are sourced from publicly available educational resources or published classroom and tutoring datasets (e.g., MathEDU (Hsu et al., 2025) and MathDial (Macina et al., 2023)); no private student/teacher records or personally identifiable information are included, and we only used the data for research purposes.

Additionally, the planned surveys with pre-service math teachers to evaluate model weaknesses and student simulation efficacy have undergone and passed strict Institutional Review Board (IRB) review at our institution. Participants provide informed consent, with anonymized responses and the option to withdraw at any time, and no sensitive personal data will be collected. During pilot use, logs will be anonymized and stored under restricted access.

References

Ikpe Justice Akpan, Yawo Mamoua Kobara, Josiah Owolabi, Asuama A. Akpan, and Onyebuchi Felix Offodile. 2025. Conversational and generative artificial intelligence and human-chatbot interaction in education and research. *International Transactions in Operational Research*, 32(3):1251–1281.

Mohammad Aljanabi, Mohanad Ghazi, Ahmed Hussein Ali, and Saad Abas Abed. 2023. Chatgpt: open possibilities. *Iraqi journal for computer science and mathematics*, 4(1):7.

Tariq A Alshugran and Lina H Kloub. 2025. Preserving student privacy while leveraging generative ai in higher education. In *2025 Northeast Section Conference*.

Nguyen Tuan Anh and Phan Duy Hung. 2024. Development of vietnamese large language model for elementary math solving problem. In *International Conference on Cooperative Design, Visualization and Engineering, CDVE 2024*, volume 15158 of *Lecture Notes in Computer Science*, pages 73–82. Springer.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, and 33 others. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Bright Asare, Yarhands Dissou Arthur, and Francis Ohene Boateng. 2023. Exploring the impact of chatgpt on mathematics performance: The influential role of student interest. *Education Science and Management*, 1(3):158–168.

Oyebola Olusola Ayeni, Nancy Mohd Al Hamad, Onyebuchi Nneamaka Chisom, Blessing Osawaru, and Ololade Elizabeth Adewusi. 2024. Ai in education: A review of personalized learning and educational technology. *GSC Advanced Research and Reviews*, 18(2):261–271.

Fatih Bildirici and 1 others. 2024. Open-source ai: An approach to responsible artificial intelligence development. *REFLEKTİF Sosyal Bilimler Dergisi*, 5(1):73–81.

Chris Bourg, Sue Kriegsman, Nick Lindsay, Heather Sardis, Erin Stalberg, and Micah Altman. 2024. Generative AI for Trustworthy, Open, and Equitable Scholarship. *An MIT Exploration of Generative AI*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*.

Sahan Bulathwela, María Pérez-Ortiz, Catherine Holloway, Mutlu Cukurova, and John Shawe-Taylor. 2024. Artificial intelligence alone will not democratise education: On educational inequality, techno-solutionism and inclusive tools. *Sustainability*, 16(2):781.

Cecilia Ka Yuk Chan and Wenjie Hu. 2023. Students' voices on generative ai: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, 20(1):43.

Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J. Page, and Vivian A. Welch. 2019. Cochrane handbook for systematic reviews of interventions. *Hoboken: Wiley*, 4.

Jiahao Chen, Zitao Liu, Mingliang Hou, Xiangyu Zhao, and Weiqi Luo. 2024a. Multi-turn classroom dialogue dataset: Assessing student performance from one-on-one conversations. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024*, pages 5333–5337. ACM.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, and 16 others. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12).

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, CVPR 2024*, pages 24185–24198.

L. Chiang. 2024. Navigating hallucinations in generative ai for education: A case study in legal teaching and learning. In *Proceedings of 17th annual International Conference of Education, Research and Innovation, ICERI 2024*, pages 7170–7174. IATED.

Longchao Da, Tiejin Chen, Zhuoheng Li, Shreyas Bachiraju, Huaiyuan Yao, Li Li, Yushun Dong, Xiyang Hu, Zhengzhong Tu, Dongjie Wang, Yue Zhao, Xuanyu Zhou, Ram Pendyala, Benjamin Stabler, Yezhou Yang, Xuesong Zhou, and Hua Wei. 2025. Generative AI in transportation planning: A survey. *CoRR*, abs/2503.07158.

I de Zarzà, J de Curtò, Gemma Roig, and Carlos T Calafate. 2023. Optimized financial planning: Integrating individual and cooperative budgeting models with llm recommendations. *AI*, 5(1):91–114.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 81 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Nasrin Dehbozorgi, Mourya Teja Kunuku, and Seyedamin Pouriyeh. 2024. Personalized pedagogy through an llm-based recommender system. In *International Conference on Artificial Intelligence in Education, AIED 2024*, volume 2151 of *Communications in Computer and Information Science*, pages 63–70. Springer.

Elena Deric, Domagoj Frank, and Dijana Vukovic. 2025. Exploring the ethical implications of using generative AI tools in higher education. *Informatics*, 12(2):36.

Zhiyi Duan, Hengnian Gu, Yuan Ke, and Dongdai Zhou. 2024. EBERT: A lightweight expression-enhanced large-scale pre-trained language model for mathematics education. *Knowledge-Based Systems*, 300:112118.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Hassan Elsayed. 2024. The impact of hallucinated information in large language models on student learning outcomes: A critical examination of misinformation risks in ai-assisted education. *Northern Reviews on Algorithmic Research, Theoretical Computation, and Complexity*, 9(8):11–23.

Ryan L Farmer, Adam B Lockwood, Anisa Goforth, and Christopher Thomas. 2024. Artificial intelligence in practice: Opportunities, challenges, and ethical considerations. *Professional Psychology: Research and Practice*.

Andrea Fedele, Clara Punzi, Stefano Tramacere, and 1 others. 2024. The altai checklist as a tool to assess ethical and legal implications for a trustworthy ai development in education. *Computer Law & Security Review*, 53:105986.

Seyum Getenet. 2024. Pre-service teachers and chatgpt in multistrategy problem-solving: Implications for mathematics teaching in primary schools. *International Electronic Journal of Mathematics Education*, 19(1).

Sagar Goyal, Eti Rastogi, Sree Prasanna Rajagopal, Dong Yuan, Fen Zhao, Jai Chintagunta, Gautam Naik, and Jeff Ward. 2024. Healai: A healthcare LLM for effective medical documentation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining, WSDM 2024*, pages 1167–1168. ACM.

Adit Gupta, Jennifer M. Reddig, Tommaso Calò, Daniel Weitekamp, and Christopher J. MacLellan. 2025. Beyond final answers: Evaluating large language models for math tutoring. *CoRR*, abs/2503.16460.

Michael Henderson, Margaret Bearman, Jennifer Chung, Tim Fawns, Simon Buckingham Shum, Kelly E Matthews, and Jimena de Mello Heredia. 2025. Comparing generative ai and teacher feedback: student perceptions of usefulness and trustworthiness. *Assessment & Evaluation in Higher Education*, pages 1–16.

Wei-Ling Hsu, Yu-Chien Tang, and An-Zi Yen. 2025. Mathedu: Towards adaptive feedback for student mathematical problem-solving. *CoRR*, abs/2505.18056.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, and 79 others. 2024. Gpt-4o system card. *CoRR*, abs/2410.21276.

Jaeho Jeon and Seongyong Lee. 2023. Large language models in education: A focus on the complementary relationship between human teachers and chatgpt. *Education and Information Technologies*, 28(12):15873–15892.

Jiyou Jia, Tianrui Wang, Yuyue Zhang, and Guangdi Wang. 2024. The comparison of general tips for mathematical problem solving generated by generative ai with those generated by human teachers. *Asia Pacific Journal of Education*, 44(1):8–28.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Hyoungwook Jin, Seonghee Lee, Hyungyu Shin, and Juho Kim. 2024. Teach AI how to code: Using large language models as teachable agents for programming education. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024*, pages 652:1–652:28. ACM.

Alexa Alice Joubin. 2024. Enhancing the trustworthiness of generative artificial intelligence in responsive pedagogy in the context of humanities higher education. In *General Aspects of Applying Generative AI in Higher Education: Opportunities and Challenges*, pages 207–220. Springer.

Muhammet Remzi Karaman and İdris Göksu. 2024. Are lesson plans created by chatgpt more effective? an

experimental study. *International Journal of Technology in Education*, 7(1):107–127.

Keunjae Kim and Kyungbin Kwon. 2023. Exploring the AI competencies of elementary school teachers in south korea. *Computers and Education: Artificial Intelligence*, 4:100137.

Shalece Kohnke and Tiffanie Zaugg. 2025. Artificial intelligence: An untapped opportunity for equity and access in stem education. *Education Sciences*, 15(1):68.

Min Lan and Xiaofeng Zhou. 2025. A qualitative systematic review on ai empowered self-regulated learning in higher education. *npj Science of Learning*, 10(1):21.

Ehsan Latif, Gyeong-Geon Lee, Knut Neumann, Tamara Kastorff, and Xiaoming Zhai. 2024. G-sciedbert: A contextualized LLM for science assessment tasks in german. *CoRR*, abs/2402.06584.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR, PMLR.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023b. Large language models in finance: A survey. In *Proceedings of the 4th ACM International Conference on AI in Finance, ICAIF 2023*, pages 374–382. ACM.

Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. 2023. Opening up chatgpt: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of the 5th International Conference on Conversational User Interfaces, CUI 2023*, pages 47:1–47:6. ACM.

Xingyu Bruce Liu, Vladimir Kirilyuk, Xiuxiu Yuan, Alex Olwal, Peggy Chi, Xiang 'Anthony' Chen, and Ruofei Du. 2023. Visual captions: Augmenting verbal communication with on-the-fly visuals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023*, pages 108:1–108:20. ACM.

Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621. Association for Computational Linguistics.

Rizwaan Malik, Dorna Abdi, Rose E. Wang, and Dorottya Demszky. 2024. Scaling high-leverage curriculum scaffolding in middle-school mathematics. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale, L@S 2024*, pages 476–480. ACM.

Hunter McNichols and Andrew S. Lan. 2025. The study-chat dataset: Student dialogues with chatgpt in an artificial intelligence course. *CoRR*, abs/2503.07928.

Michael F. McTear. 2022. *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots*. Synthesis Lectures on Human Language Technologies. Springer Nature.

Wesley Morris, Langdon Holmes, Joon Suh Choi, and Scott Crossley. 2024. Automated scoring of constructed response items in math assessment using large language models. *International journal of artificial intelligence in education*, pages 1–28.

Ryosuke Nakamoto, Brendan Flanagan, Taisei Yamauchi, Yiling Dai, Kyosuke Takami, and Hiroaki Ogata. 2023. Enhancing automated scoring of math self-explanation quality using llm-generated datasets: A semi-supervised approach. *Computers*, 12(11):217.

Tanya Nazaretsky, Moriah Ariely, Mutlu Cukurova, and Giora Alexandron. 2022. Teachers' trust in ai-powered educational technology and a professional development program to improve it. *British journal of educational technology*, 53(4):914–931.

Matthew Nyaaba, Lehong Shi, Macharious Nabang, Xiaoming Zhai, Patrick Kyeremeh, Samuel Arthur Ayoberd, and Bismark Nyaaba Akanzire. 2024. Generative AI as a learning buddy and teaching assistant: Pre-service teachers' uses and attitudes. *CoRR*, abs/2407.11983.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Timothy Pelton and Leslee Francis Pelton. 2023. Adapting chatgpt to support teacher education in mathematics. In *Society for Information Technology & Teacher Education International Conference*, pages 1662–1670. Association for the Advancement of Computing in Education (AACE).

Francisco José García Peñalvo, Marc Alier, Juanan Pereira, Maria Jose Casany, and 1 others. 2024. Safe, transparent, and ethical artificial intelligence: Keys to quality sustainable education (sdg4). *IJERI: International Journal of Educational Research and Innovation*, pages 1–21.

Ethan Prihar, Morgan P. Lee, Mia Hopman, Adam Tauman Kalai, Sofia Vempala, Allison Wang, Gabriel Wickline, Aly Murray, and Neil T. Heffernan. 2023. Comparing different approaches to generating mathematics explanations using large language models. In *International Conference on Artificial Intelligence in Education, AIED 2023*, volume 1831 of *Communications in Computer and Information Science*, pages 290–295. Springer.

Zachery Quince, Kathy Petkoff, Ruby N Michael, Scott Daniel, and Sasha Nikolic. 2024. The current ethical considerations of using genai in engineering education and practice: A systematic literature review.

In *Proceedings of the 35th Annual Conference of the Australasian Association for Engineering Education, AAEE 2024*, pages 509–517. Engineers Australia Christchurch.

Petar Radanliev. 2025. Ai ethics: Integrating transparency, fairness, and privacy in ai development. *Applied Artificial Intelligence*, 39(1):2463722.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR.

Nitin Rane. 2023. Enhancing mathematical capabilities through chatgpt and similar generative artificial intelligence: Roles and challenges in solving mathematical problems. *SSRN 4603237*.

Sara Ratner, Rebecca Williams, and Elizabeth Wonnacott. 2025. Exploring ai in education through interdisciplinary collaboration. In *INTED2025 Proceedings*, pages 5650–5655. IATED.

Chandan K. Reddy and Parshin Shojaee. 2025. Towards scientific discovery with generative AI: progress, opportunities, and challenges. In *Proceedings of the AAAI Conference on Artificial Intelligence, AAAI 2025*, pages 28601–28609. AAAI Press.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, and 34 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530.

Ioannis Rizos, Evaggelos Foykas, and Spiros V Georgakopoulos. 2024. Enhancing mathematics education for students with special educational needs through generative ai: A case study in greece. *Contemporary Educational Technology*, 16(4).

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, pages 10674–10685. IEEE.

Ridha Rouabhia. 2025. Ethical implications of ai: Examining bias and fairness in ai-powered education. In *Using AI Tools in Text Analysis, Simplification, Classification, and Synthesis*, pages 411–438. IGI Global Scientific Publishing.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, volume 35, pages 36479–36494.

Lior Shani, Aviv Rosenberg, Asaf B. Cassel, Oran Lang, Daniele Calandriello, Avital Zipori, Hila Noga, Orgad Keller, Bilal Piot, Idan Szpektor, Avinatan Hassidim, Yossi Matias, and Rémi Munos. 2024. Multi-turn reinforcement learning from preference human feedback. *CoRR*, abs/2405.14655.

Jia Tracy Shen, Michiharu Yamashita, Ethan Prihar, Neil T. Heffernan, Xintao Wu, and Dongwon Lee. 2021. Mathbert: A pre-trained language model for general NLP tasks in mathematics education. *CoRR*, abs/2106.07340.

Yukyeong Song, Jinhee Kim, Zifeng Liu, Chenglu Li, and Wanli Xing. 2024. Students' perceived roles, opportunities, and challenges of a generative ai-powered teachable agent: A case of middle school math class. *Journal of Research on Technology in Education*, pages 1–19.

Qianwen Tang, Wenbo Deng, Yidan Huang, Shuaijie Wang, and Hao Zhang. 2025. Can generative artificial intelligence be a good teaching assistant?—an empirical analysis based on generative ai-assisted teaching. *Journal of Computer Assisted Learning*, 41(3):e70027.

Marek Urban, Filip Dechterenko, Jirí Lukavský, Veronika Hrabalová, Filip Svacha, Cyril Brom, and Kamila Urban. 2024. Chatgpt improves creative problem-solving performance in university students: An experimental study. *Computers & Education*, 215:105031.

Selin Urhan, Oguzhan Gençaslan, and Senol Dost. 2024. An argumentation experience regarding concepts of calculus with chatgpt. *Interactive Learning Environments*, 32(10):7186–7211.

Felipe Urrutia and Roberto Araya. 2024. Who's the best detective? large language models vs. traditional machine learning in detecting incoherent fourth grade math answers. *Journal of Educational Computing Research*, 61(8):187–218.

Nisha Varghese and Gobi Ramasamy. 2024. Unveiling the potential of large language models: Redefining learning in the age of generative ai. In *Intersection of AI and Business Intelligence in Data-Driven Decision-Making*, pages 389–414. IGI Global.

Aniella Mihaela Vieriu and Gabriel Petrea. 2025. The impact of artificial intelligence (ai) on students' academic development. *Education Sciences*, 15(3):343.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S. Yu, and Qingsong Wen. 2024. Large language models for education: A survey and outlook. *CoRR*, abs/2403.18105.

Xinghua Wang, Linlin Li, Seng Chee Tan, Lu Yang, and Jun Lei. 2023a. Preparing for ai-enhanced education: Conceptualizing and empirically examining teachers' AI readiness. *Computers in Human Behavior*, 146:107798.

Yuqing Wang, Yun Zhao, and Linda R. Petzold. 2023b. Are large language models ready for healthcare? A comparative study on clinical language understanding. In *Machine Learning for Healthcare Conference, MLHC 2023*, volume 219 of *Proceedings of Machine Learning Research*, pages 804–823. PMLR.

Shuo Xing, Hongyuan Hua, Xiangbo Gao, Shenzhe Zhu, Renjie Li, Kexin Tian, Xiaopeng Li, Heng Huang, Tianbao Yang, Zhangyang Wang, Yang Zhou, Huaxiu Yao, and Zhengzhong Tu. 2024. Autotrust: Benchmarking trustworthiness in large vision language models for autonomous driving. *CoRR*, abs/2412.15206.

Wanli Xing, Yukyeong Song, Chenglu Li, Zifeng Liu, Wangda Zhu, and Hyunju Oh. 2025. Development of a generative ai-powered teachable agent for middle school mathematics learning: A design-based research study. *British Journal of Educational Technology*.

Shihao Xu, Yiyang Luo, and Wei Shi. 2024. Geo-llava: A large multi-modal model for solving geometry math problems with meta in-context learning. In *Proceedings of the 2nd Workshop on Large Generative Models Meet Multimodal Applications*, pages 11–15.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. 2022. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022.

Yaman Yu, Tanusree Sharma, Melinda Hu, Justin Wang, and Yang Wang. 2024. Exploring parent-child perceptions on safety in generative AI: concerns, mitigation strategies, and design implications. *CoRR*, abs/2406.10461.

Wahid Yunianto, Selen Galic, and Zsolt Lavicza. 2024. Exploring computational thinking in mathematics education: Integrating chatgpt with geogebra for enhanced learning experiences. *International Journal of Education in Mathematics, Science and Technology*, pages 1451–1470.

Fan Zhang, Chenglu Li, Owen Henkel, Wanli Xing, Sami Baral, Neil Heffernan, and Hai Li. 2024a. Mathllms: Ai cyberinfrastructure with pre-trained transformers for math education. *International Journal of Artificial Intelligence in Education*, pages 1–24.

Yi Zhang, Guangyou Zhou, Zhiwen Xie, and Jimmy Xiangji Huang. 2024b. Number-enhanced representation with hierarchical recursive tree decoding for math word problem solving. *Information Processing & Management*, 61(2):103585.

Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. 2024c. A comprehensive survey of scientific large language models and their applications in scientific discovery. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024*, pages 8783–8817. Association for Computational Linguistics.

Pu Zhao, Xuan Shen, Zhenglun Kong, Yixin Shen, Sung-En Chang, Timothy Rupprecht, Lei Lu, Enfu Nan, Changdi Yang, Yumei He, Xingchen Xu, Yu Huang, Wei Wang, Yue Chen, Yong He, and Yanzhi Wang. 2024. Fully open source moxin-7b technical report. *CoRR*, abs/2412.06845.

Ou Zheng, Mohamed A. Abdel-Aty, Dongdong Wang, Zijin Wang, and Shengxuan Ding. 2023. Chatgpt is on the horizon: Could a large language model be all we need for intelligent transportation? *CoRR*, abs/2303.05382.

Irina Zlotnikova, Hlomani Hlomani, Tshepiso Mokgetse, and Kelebonye Bagai. 2025. Establishing ethical standards for genai in university education: a roadmap for academic integrity and fairness. *Journal of Information, Communication and Ethics in Society*, 23(2):188–216.