A Call for Fully Open and Trustworthy Generative AI for Education

Anonymous Author(s)
Affiliation
Address
email

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 like LLaMA and DeepSeek-R1 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 [1], Parti [2], DALL-E [3, 4], and Stable Diffusion [5]), large language models (LLMs) such as GPT-series [6], Gemini [7], LLaMA [8], and DeepSeek [9], and multi-modal large language models (MLLMs) including LLaVA [10], BLIP-2 [11], and GPT-4o [12]. 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 [13, 14].

The rapid development of these powerful LLM models has led to widespread adoption across numerous domains, from healthcare [15, 16, 17], finance [18, 19], scientific discoveries [20, 21, 22], transportation [23, 24, 25], up to education [26, 27], etc. In particular, education has significantly benefited from the integration of GenAI technologies, driven by their potential to fundamentally transform teaching methodologies, curriculum design, student learning and educational research [28, 29, 30]. 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. (1) **As a tutor/assistant**: In designing an intelligent tutoring system, tips from the system or virtual tutors are crucial in helping students solve difficult questions in disciplines like mathematics. Traditionally, the manual design of general tips by teachers is time-consuming and error-prone [31], while generative AI like ChatGPT [6] presents a new channel for designing general tips [32, 33].

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

(2) **For curriculum development**: The results of research of determining whether students' math achievements improved using ChatGPT to prepare lesson plans in primary school math courses is effective in academic achievement [34]. Teachers should consider ChatGPT and their plans, combining them and benefiting from both in the implementation process [35, 36]. (3) **For enhancing student achievement**: One investigation examines the influence of ChatGPT on mathematics achievement, with a specific focus on the moderating role of students' interest in mathematics [37]. Recent studies identified a statistically significant positive moderation effect of students' interest on the association between ChatGPT usage and mathematics achievement [38, 39]. (4) **For model development**: A lightweight expression-enhanced large-scale pre-trained language model, called EBERT [40], for mathematics education. Through downstream tasks in mathematical education, models like this can outperform several former state-of-the-art baselines (such as MathBERT [41] and GPT-3 [42]) in terms of accuracy and F1-score [43, 44]. (5) **Evaluation of GenAI ability**: Recent research explores methodologies aimed at fostering a synergistic relationship between generative AI and human intelligence, encouraging the students' development of robust problem-solving strategies while harnessing the computational capabilities of AI [45, 46, 47].

Motivation. Despite these advancements, critical limitations remain prevalent in the current GenAI-for-education (GenAI4Edu) 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 [48, 49]. This opacity introduces significant challenges regarding model *interpretability*, *data privacy*, *equitable accessibility*, intellectual property rights, and accountability in educational contexts [50]. In contrast, open-source GenAI models—such as the LLaMA series and the recently emerged DeepSeek-R1 [9]—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 [9]. 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 [51].

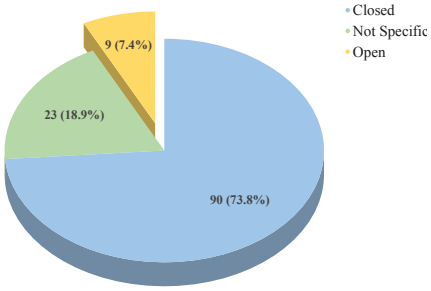


Figure 1: Percentage of Open vs. Closed-Source LLMs used in Education.

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, then present our position that open-source GenAI platform for education is promising in Sec. 3. We have presented our initial explorations of such an open source platform in Sec. 4, while Sec. 5 discuss possible counterarguments and address potential limitations of our proposed approach. Finally, we conclude with actionable recommendations and policy proposals for educators, researchers, AI developers, and policymakers in Sec. 6, aiming to catalyze a shift toward open and trustworthy GenAI in education.

2 A Review of Prior Works

Generative AI 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. [52]. 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 [31, 40, 44, 53, 54, 55, 56, 57, 58] (see Figure 1). The predominance of closed-source, commercial GenAI models underscores a significant reliance on proprietary technologies such as ChatGPT, Gemini, and DeepSeek [9], 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.

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.

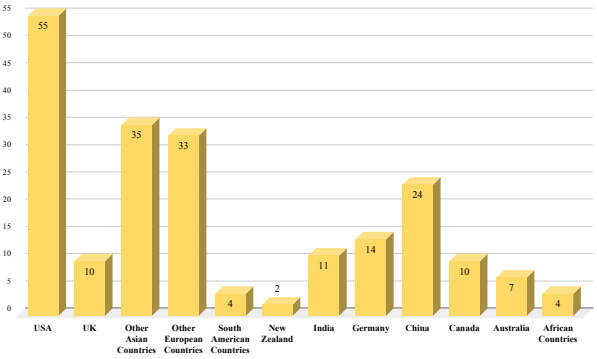


Figure 2: Geographical locations of the published papers.

Closed-source models, exemplified by GPT-4 [12] and Gemini 1.5 [59], 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.

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.

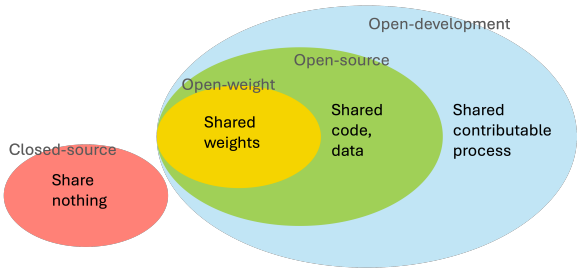


Figure 3: Definition of Open-Sourceness.

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 [60].

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 [61]. 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 [62]. Unlike closed-source solutions that operate as opaque "black boxes", open-source GenAI models provide transparent insights into their architecture and decision-making processes [63]. 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 [64]. Open-source models offer distinct advantages in terms of security and privacy by allowing comprehensive audits and community-driven security enhancements [65]. 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.

In summary, 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. The Meta LLaMA series (e.g., LLaMA 3 [8]), InternVL [66, 67], and Mistral [68] are all prominent open-source LLMs. The maturity, adaptability, and inherent openness of such open-source models enable the necessary domain-specific fine-tuning that would be difficult with closed-source alternatives. Meta's LLaMA 3 family [8] includes pre-trained and instruction-tuned generative text models, available in various parameter sizes (e.g., 8B, 70B), optimized for dialogue use cases. InternVL [66, 67], particularly InternVL3-78B [69], is one of the state-of-the-art multimodal LLMs (MLLMs) that integrate visual and linguistic capabilities during a unified pre-training stage. Mistral [68], including models like Mistral 7B and Mistral-8x22B, is another notable open-source LLM, often leveraging a sparse Mixture-of-Experts (SMoE) architecture. The transparency and flexibility inherent in these open-source models 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 following datasets. (1) **StudyChat Dataset**: a publicly available resource designed to understand real-world student interactions with an LLM-powered tutoring chat-bot, comprising conversations and student utterances [70]; (2) **Multi-turn Classroom Dialogue Dataset**: a benchmark for assessing student performance in math questions based on teacher-student dialogue data from online one-on-one tutoring classes [71]; (3) **MathDial Dataset**: a specialized dialogue tutoring dataset focused on math reasoning problems and common student confusions, richly annotated with GPT-3.5 simulated student confusions and pedagogical attributes [72]; (4) **Education Dialogue Dataset**: a novel multi-turn dataset simulating dynamic teacher-student interactions, entirely generated by prompting Gemini Ultra [73].

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 [74] 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 [75], 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 [76]. Many forward-thinking schools move beyond initial hesitation to actively explore and implement AI application to education and integrating tools into their teaching practice [77]. Although GenAI has brought challenges, a vast majority of students report using AI technologies in their academic activities [78], among which virtual assistants and AI-based educational platforms are particularly popular. Additionally, despite concerns about data collection and misinformation [79], 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. With access to state-of-the-art models, curated domain-specific datasets, affordable compute infrastructure, and a receptive user base, the necessary components are already in place. 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 across multiple domains of expertise. 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 the formation of cross-disciplinary working 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 or summits—can institutionalize this collaboration and ensure ongoing feedback exchange [80].

Responsibilities of Key Stakeholders. *Education researchers* are responsible for defining pedagogical use cases, identifying key learning objectives, and collecting representative data from real classroom contexts. Their work includes annotating instructional dialogue, modeling student misconceptions, and empirically evaluating system efficacy through classroom trials [80]. *Computer scientists and engineers* will lead the development, fine-tuning, and deployment of open-source LLMs [81]. Their responsibilities include selecting appropriate architectures, implementing scalable training and serving pipelines, and ensuring system robustness, efficiency, and privacy compliance [81]. Together, these communities must co-design interfaces and feedback mechanisms to support teacher-facing and learner-facing applications [82].

Technical Infrastructure. The platform must be supported by robust AI infrastructure [81]. 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, multilingual 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 [82, 83]. 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 it will not succeed without 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. Only through such collective effort can we 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 [51].

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 to entry. This lowers the threshold for adoption in diverse educational settings, particularly in underserved or rural communities that are most in need of scalable and affordable AI support.

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 educators and researchers to inspect how models are trained and how decisions are made. 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 [84, 85, 86, 87].

Trustworthiness. Establishing trust is foundational when integrating GenAI into educational environments [88, 89]. Trustworthiness in this context encompasses a multi-dimensional evaluation

that includes factual accuracy, safety against adversarial manipulation, robustness to unseen conditions, preservation of privacy, and fairness in the treatment of diverse user groups [90]. These dimensions—thoroughly outlined in the AutoTrust benchmark —are equally critical when deploying GenAI in classrooms, where students and educators rely on AI not only for accurate content but also for safe, ethical, and equitable interactions.

Factuality and Uncertainty Awareness. Just as in autonomous driving systems, educational GenAI must avoid hallucinations or misleading content [91, 92]. Models should not only provide accurate answers but also signal uncertainty when relevant. This helps 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, such as adversarial prompts that induce inappropriate or off-topic responses. Drawing lessons from adversarial testing in vision-language models (VLMs) for autonomous systems [90], 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. As educational GenAI systems interact with sensitive student data (e.g., learning disabilities, behavioral feedback), they must actively protect privacy and avoid unintended 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 enable institutions to impose custom privacy controls and transparency audits—essential for compliance with student data protection regulations like FERPA [93, 94].

Fairness and Bias Mitigation. AI must treat students equitably across demographic lines [95]. AutoTrust's fairness testing identifies disparities in model behavior depending on user identity or environmental features. Similarly, 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 not a monolithic criterion but a composite of verifiability, security, resilience, and ethics. Open-source models offer a uniquely powerful substrate for achieving these goals, as they invite scrutiny, enable adaptation, and promote accountability—attributes that proprietary solutions systematically obscure. By adopting a trustworthiness-first design philosophy, educational institutions can safeguard learners while enhancing the reliability and social value of GenAI systems.

3.5 Incorporating Feedback to Make GenAI-Education Platform Better

To effectively integrate a GenAI platform in education, a practical approach involves iterative improvement based on educator and practitioner feedback. This ensures the platform continuously benefits educators and adapts to their individual needs [96].

The GenAI platform generate diverse professional development training materials with various contexts for instructors and users such as detailed documentation, training videos, user guides, booklets, and an accessible support website [97]. Before full implementation, education experts will pilot-test the platform to provide structured feedback on content accuracy, pedagogical effectiveness, and user experience. The feedback will provide direction about how to adjust and refine the training methods and interface design for the platform. Then, the large-scale trials will expand to partnership universities around the local area and will be conducted by collecting quantitative useful data and qualitative feedback via interviews from various educators to optimize platform scalability and localization features. This input will guide model development, emphasizing local applicability and broader usability. Following regional validation, the system is promoted in progress and 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 the model's practicality and ensuring ethical considerations, including diverse cultural understanding.

Local experts will establish regular review cycles, analyzing user behavior and survey responses to ensure instructional relevance and technological integration get iteratively improved.

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. Our basic user interface (UI) design and major features are showcased below (See Figure 4).

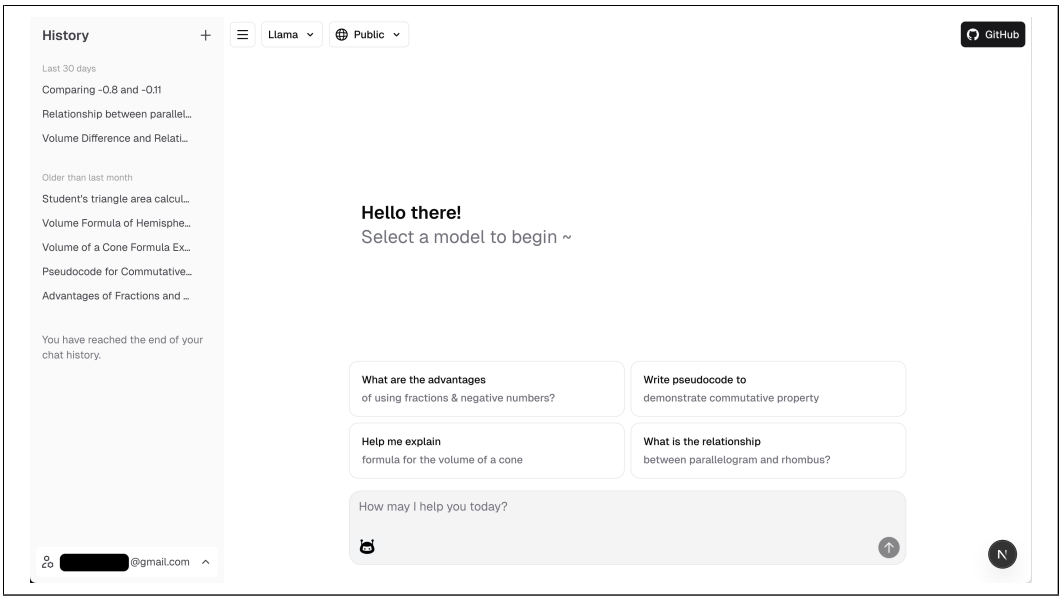
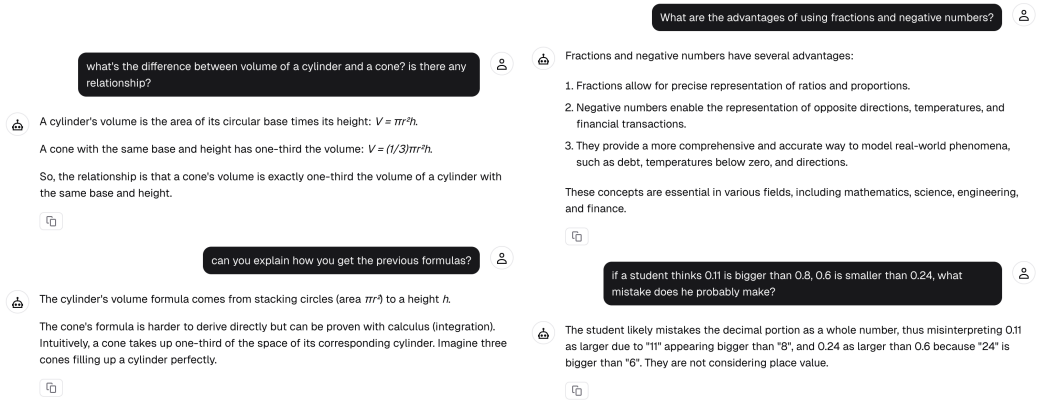


Figure 4: Our User Interface Design



(a) Problem 1 (b) Problem 2

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

The LLMs, specifically LLaMA 3 model [8] fine-tuned on 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 [98], by casting the GenAI in the role of a student struggling with mathematics, 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 to let prospective mathematics teachers and instructors participate in the following testing process and identify cases and situations where the GenAI model fail to generate satisfying responses.

5 Alternative Opinion

While this paper advocates for the development of open-source GenAI platforms in education, we still acknowledge the compelling advantages offered by current closed-source commercial models. It is 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 [67]. 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 [12] and Google [59], are typically offered as managed services accessible through user-friendly web interfaces or well-documented Application Programming Interfaces (APIs). This significantly lowers the technical barrier to entry for educators and developers, allowing for rapid integration and experimentation without deep technical expertise in model deployment or infrastructure management. The convenience is a major draw for many users. (2) **Service Support**: Users of closed-source models generally do not need to undertake the complex and resource-intensive process of model pre-training or extensive fine-tuning. Commercial providers manage the entire life-cycle of their models, including continuous improvement, updates, and maintenance. Professional technical support teams can even provide custom solutions tailored to specific industry (including education) needs, and help schools respond to sudden needs quickly, while the open source community has not yet formed a similar standardized tool chain.

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.

6 Conclusion

We call for more open, inclusive, efficient, and trustworthy generative AI 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. In parallel, we advocate for participatory governance frameworks that involve educators, researchers, and policy stakeholders in model iteration and oversight. These directions will help realize an educational AI landscape that is not only technically advanced, but also open, inclusive, and grounded in public interest.

References

[1] 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. Photorealistic text-to-image diffusion models with deep language understanding. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, volume 35, pages 36479–36494, 2022.

[2] 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. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022, 2022.

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

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

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

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

[7] 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, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805, 2023.

[8] 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, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024.

[9] 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, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu

Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025.

[10] Xingyu Bruce Liu, Vladimir Kirilyuk, Xiuxiu Yuan, Alex Olwal, Peggy Chi, Xiang 'Anthony' Chen, and Ruofei Du. Visual captions: Augmenting verbal communication with on-the-fly visuals. In Albrecht Schmidt, Kaisa Väänänen, Tesh Goyal, Per Ola Kristensson, Anicia Peters, Stefanie Mueller, Julie R. Williamson, and Max L. Wilson, editors, *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023*, pages 108:1–108:20. ACM, 2023.

[11] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR, PMLR, 2023.

[12] 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, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll L. Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, and Dane Sherburn. Gpt-4o system card. *CoRR*, abs/2410.21276, 2024.

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

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

[15] Yuqing Wang, Yun Zhao, and Linda R. Petzold. Are large language models ready for healthcare? A comparative study on clinical language understanding. In Kaivalya Deshpande, Madalina Fiterau, Shalmali Joshi, Zachary C. Lipton, Rajesh Ranganath, Iñigo Urteaga, and Serene Yeung, editors, *Machine Learning for Healthcare Conference, MLHC 2023*, volume 219 of *Proceedings of Machine Learning Research*, pages 804–823. PMLR, 2023.

[16] Sagar Goyal, Eti Rastogi, Sree Prasanna Rajagopal, Dong Yuan, Fen Zhao, Jai Chintagunta, Gautam Naik, and Jeff Ward. Healai: A healthcare LLM for effective medical documentation. In Luz Angelica Caudillo-Mata, Silvio Lattanzi, Andrés Muñoz Medina, Leman Akoglu, Aristides Gionis, and Sergei Vassilvitskii, editors, *Proceedings of the 17th ACM International Conference on Web Search and Data Mining, WSDM 2024*, pages 1167–1168. ACM, 2024.

[17] Jefferson Hunter, Gregg Nicandri, and Kevin J Bozic. Value-based healthcare: How can large language model (llm) technology be integrated with patient-reported outcomes? *Clinical Orthopaedics and Related Research*, pages 10–1097, 2022.

[18] Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 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, 2023.

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

[20] Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. A comprehensive survey of scientific large language models and their applications in scientific discovery. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024*, pages 8783–8817. Association for Computational Linguistics, 2024.

[21] Oskar Wysocki, Magdalena Wysocka, Danilo S. Carvalho, Alex Teodor Bogatu, Danilo Miranda Gusicuma, Maxime Delmas, Harriet Unsworth, and André Freitas. An llm-based knowledge synthesis and scientific reasoning framework for biomedical discovery. *CoRR*, abs/2406.18626, 2024.

[22] Chandan K. Reddy and Parshin Shojaee. Towards scientific discovery with generative AI: progress, opportunities, and challenges. In Toby Walsh, Julie Shah, and Zico Kolter, editors, *Proceedings of the AAAI Conference on Artificial Intelligence, AAAI 2025*, pages 28601–28609. AAAI Press, 2025.

[23] 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. Generative AI in transportation planning: A survey. *CoRR*, abs/2503.07158, 2025.

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

[25] Huanqi Yang, Rucheng Wu, and Weitao Xu. Transcompressor: Llm-powered multimodal data compression for smart transportation. In Weisong Shi, Deepak Ganesan, and Nicholas D. Lane, editors, *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking, ACM MobiCom 2024*, pages 2335–2340. ACM, 2024.

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

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

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

[29] Hyoungwook Jin, Seonghee Lee, Hyungyu Shin, and Juho Kim. Teach AI how to code: Using large language models as teachable agents for programming education. In Florian 'Floyd' Mueller, Penny Kyburz, Julie R. Williamson, Corina Sas, Max L. Wilson, Phoebe O. Toups Dugas, and Irina Shklovski, editors, *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024*, pages 652:1–652:28. ACM, 2024.

[30] Yukyeong Song, Jinhee Kim, Zifeng Liu, Chenglu Li, and Wanli Xing. 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, 2024.

[31] Jiyou Jia, Tianrui Wang, Yuyue Zhang, and Guangdi Wang. 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, 2024.

[32] Ethan Prihar, Morgan P. Lee, Mia Hopman, Adam Tauman Kalai, Sofia Vempala, Allison Wang, Gabriel Wickline, Aly Murray, and Neil T. Heffernan. Comparing different approaches to generating mathematics explanations using large language models. In Ning Wang, Genaro Rebolledo-Mendez, Vania Dimitrova, Noboru Matsuda, and Olga C. Santos, editors, *International Conference on Artificial Intelligence in Education, AIED 2023*, volume 1831 of *Communications in Computer and Information Science*, pages 290–295. Springer, 2023.

[33] Felipe Urrutia and Roberto Araya. 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, 2024.

[34] Muhammet Remzi Karaman and İdris Göksu. Are lesson plans created by chatgpt more effective? an experimental study. *International Journal of Technology in Education*, 7(1):107–127, 2024.

[35] Rizwaan Malik, Dorna Abdi, Rose E. Wang, and Dorottya Demszky. Scaling high-leverage curriculum scaffolding in middle-school mathematics. In David Joyner, Min Kyu Kim, Xu Wang, and Meng Xia, editors, *Proceedings of the Eleventh ACM Conference on Learning @ Scale, L@S 2024*, pages 476–480. ACM, 2024.

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

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

[38] Wahid Yunianto, Selen Galic, and Zsolt Lavicza. 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, 2024.

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

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

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

[42] 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, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020.

[43] Shihao Xu, Yiyang Luo, and Wei Shi. 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, 2024.

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

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

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

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

[48] Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. Opening up chatgpt: Tracking openness, transparency, and accountability in instruction-tuned text generators. In Minha Lee, Cosmin Munteanu, Martin Porcheron, Johanne Trippas, and Sarah Theres Völkel, editors, *Proceedings of the 5th International Conference on Conversational User Interfaces, CUI 2023*, pages 47:1–47:6. ACM, 2023.

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

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

[51] 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. Fully open source moxin-7b technical report. *CoRR*, abs/2412.06845, 2024.

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

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

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

[55] Nasrin Dehbozorgi, Mourya Teja Kunuku, and Seyedamin Pouriyeh. Personalized pedagogy through an llm-based recommender system. In Andrew M. Olney, Irene-Angelica Chounta, Zitao Liu, Olga C. Santos, and Ig Ibert Bittencourt, editors, *International Conference on Artificial Intelligence in Education, AIED 2024*, volume 2151 of *Communications in Computer and Information Science*, pages 63–70. Springer, 2024.

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

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

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

[59] 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, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, and et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530, 2024.

[60] Andrea Fedele, Clara Punzi, Stefano Tramacere, et al. 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, 2024.

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

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

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

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

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

[66] 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. 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, 2024.

[67] 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, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12), 2024.

[68] 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. Mistral 7b. *CoRR*, abs/2310.06825, 2023.

[69] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *CoRR*, abs/2504.10479, 2025.

[70] Hunter McNichols and Andrew S. Lan. The studychat dataset: Student dialogues with chatgpt in an artificial intelligence course. *CoRR*, abs/2503.07928, 2025.

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

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

[73] 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. Multi-turn reinforcement learning from preference human feedback. *CoRR*, abs/2405.14655, 2024.

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

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

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

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

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

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

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

[81] Nisha Varghese and Gobi Ramasamy. 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, 2024.

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

[83] Qianwen Tang, Wenbo Deng, Yidan Huang, Shuaijie Wang, and Hao Zhang. 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, 2025.

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

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

[86] Zachery Quince, Kathy Petkoff, Ruby N Michael, Scott Daniel, and Sasha Nikolic. 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, 2024.

[87] Irina Zlotnikova, Hlomani Hlomani, Tshepiso Mokgetse, and Kelebonye Bagai. 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, 2025.

[88] Alexa Alice Joubin. 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, 2024.

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

[90] 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. Autotrust: Benchmarking trustworthiness in large vision language models for autonomous driving. *CoRR*, abs/2412.15206, 2024.

[91] Hassan Elsayed. 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, 2024.

[92] L. Chiang. 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, 2024.

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

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

[95] Ridha Rouabhia. 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, 2025.

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

[97] Timothy Pelton and Leslee Francis Pelton. 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), 2023.

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