LLMS' ROLES AND LIMITATIONS IN EDUCATION

Richárd Ádám Vécsey [1, 2] *

1 Hungarian Academy of Sciences, Veszprém Regional Committee — Subcommittee on Economics, Law and Social Sciences — Working Group on Social Science Research into AI and Creative Industries, Veszprém HU-8200, Hungary

2 Independent Entrepreneur, Hungary

*Correspondence: vecsey.richard@gmail.com

Abstract

The widespread integration of large language models (LLMs) into the educational landscape necessitates a comprehensive analysis of their multifaceted roles and inherent limitations. This paper first explores the significant opportunities these models present for students, teachers, and researchers, including the facilitation of personalized learning pathways, the streamlining of curriculum development, and the acceleration of AI-assisted research. However, it also rigorously examines the core challenges that arise from the technology's fundamental design. These critical issues include threats to academic integrity, such as plagiarism and overreliance; the inherent risks of data bias and cultural misrepresentation; and the persistent problem of hallucination, where models generate factually incorrect or entirely fabricated information. To mitigate these risks, this paper advocates for a multi-layered approach centered on a robust verification methodology that combines both manual and algorithmic methods. It further argues that greater transparency from developers regarding model capabilities and limitations is essential for promoting the responsible use of these tools. Ultimately, the aim of this work is to equip all educational stakeholders with the knowledge required to ethically navigate this evolving AI environment, thereby ensuring that these technologies enhance academic integrity and critical thinking rather than undermine them.

Keywords: artificial intelligence, large language models, education technology, learning methodology, teaching methodology, prompting, hallucination, bias, benchmarking, plagiarism

1. Introduction

The widespread adoption of large language models (LLMs) has initiated a profound transformation across various sectors, including the field of education. These powerful models are no longer confined to specialized research but have become ubiquitous tools, capable of assisting with tasks that range from everyday problem-solving to complex business operations. This rapid integration presents a critical need for a balanced and comprehensive understanding of their roles within academic environments. This paper explores the multifaceted capabilities and inherent limitations of LLMs from the perspective of both students and teachers, who are the primary actors affected by this technological shift. The primary goal is to provide a nuanced perspective on the application of LLMs in education, addressing their potential to facilitate personalized learning and streamline curriculum development, while simultaneously scrutinizing critical issues such as plagiarism, inherent biases, and the risk of hallucination. By examining these opportunities and threats in a single framework, this work aims to equip all stakeholders with the knowledge necessary to effectively and ethically navigate the evolving landscape of AI in education.

2. Methodology

As an AI developer, I have real experience and insights about how different large language models work and what are their limitations. The first step for researching this topic is to distinguish between actors in the education system. The main use cases are agnostic for the education level, as the tasks that have to be solved are the same on different levels. The result of the best

slicing method is: teacher, student, and researcher. Teacher and student categories are self-evident, while researcher is a person who does scientific research at a university. The roles are overlapping, one part comes from the student role, one part comes from the teacher role, with a strong need to make publications. After defining the roles, it is able to investigate the threats and challenges. Some of them come from the base concept of LLMs and is not possible to avoid. The other parts of the problems come from the human mind and how we react or connect to our environment. The responsibility of developers and people in the education system are different The first group have to build better models with less limitation. The capabilities of model also has to be articulated better. The last group of people has to be more cautious and has to use the models responsibly.

3. Results

Large language models (LLMs) offer transformative support for students across several educational tasks. For homework assistance, LLMs can help clarify complex concepts, generate structured responses, and provide instant feedback, making them especially valuable for learners without access to traditional tutoring. However, this convenience may lead to overreliance, diminishing students' ability to think critically and solve problems independently. In terms of knowledge structuring, LLMs excel at organizing raw information into outlines, summaries, and concept maps, helping students visualize relationships between ideas and retain content more effectively. This capability supports deeper learning and better exam preparation. Personalized learning is another area where LLMs shine. They adapt explanations and practice materials to match a student’s pace, style, and interests, fostering engagement and self-directed study. Finally, in AI-assisted research, students use LLMs to explore academic topics, identify relevant sources, and draft initial versions of essays or reports. These models can suggest keywords, summarize articles, and highlight gaps in literature, although students must still verify the accuracy and relevance of the generated content to maintain academic integrity.

Teachers also benefit from LLMs in several key areas. For curriculum development, these models can generate lesson plans, suggest learning activities, and align content with educational standards, saving educators time and effort. In grading and assessment, LLMs assist by evaluating student responses, identifying common errors, and providing constructive feedback, which helps streamline the evaluation process. They can also support tutoring by offering explanations tailored to individual student needs, enabling teachers to address diverse learning styles more effectively. Additionally, LLMs help monitor student progress by analyzing performance data and identifying areas for improvement, allowing for timely interventions. Despite these advantages, educators must remain vigilant about the limitations of LLMs, including potential biases and inaccuracies, to ensure that AI tools enhance rather than hinder the learning experience.
Researchers increasingly rely on LLMs to streamline various aspects of the academic process. These models assist in summarizing studies, synthesizing findings across multiple papers, and identifying key trends in literature. They support research design by generating hypotheses, suggesting methodologies, and helping refine research questions. LLMs also facilitate reference collection by recommending relevant sources, formatting citations, and organizing bibliographies. Additionally, researchers use LLMs to draft abstracts and introductions, and to translate complex findings into accessible language for broader dissemination. As AI tools continue to evolve, they offer promising avenues for accelerating scholarly work and enhancing interdisciplinary collaboration.

From a broader perspective, LLMs play a crucial role in content dissemination and accessibility. They can translate material into multiple languages and adjust the complexity of explanations to suit different audiences. For example, simplifying the concept of quantum computing for a fifth grader or tailoring it for a college student. This adaptability makes LLMs valuable for public education, science communication, and global outreach. Beyond translation and simplification, LLMs contribute to three additional areas: enhancing accessibility for individuals with disabilities through alternative formats, supporting lifelong learning

by curating personalized educational pathways, and fostering cross-cultural understanding by contextualizing content for diverse cultural backgrounds.

As LLMs become more integrated into educational and research workflows, users may develop an overreliance on these tools. This dependency can reduce initiative and problem-solving skills, as individuals defer to AI for tasks they might otherwise tackle independently. Over time, this reliance may hinder the development of critical thinking and creativity. It also risks creating a generation of learners and researchers who lack confidence in their own abilities. Institutions must encourage balanced use and promote human-led inquiry alongside AI support.

LLMs often generate content based on patterns found in existing data, which can lead to repetitive or derivative outputs. Students and researchers may submit AI-generated work that lacks originality or personal insight. This undermines the value of authentic intellectual contributions and may diminish the diversity of thought in academic discourse. Educators must emphasize the importance of unique perspectives and critical engagement with material. Encouraging reflection and revision can help mitigate this issue.

The ease of generating text with LLMs raises concerns about plagiarism, both intentional and unintentional. Users may copy AI-generated content without proper attribution, believing it to be original. This blurs the lines of academic integrity and complicates enforcement of plagiarism policies. Institutions must update guidelines to address AI-generated content and educate users on ethical usage. Tools for detecting AI-assisted writing may also become necessary.

LLMs usually require access to user data to provide personalized responses, raising questions about data security and privacy. Sensitive information shared with AI systems may be stored or processed in ways that users do not fully understand. This creates risks of data breaches or misuse, especially in educational settings where student data is involved. Transparency in data handling and strict privacy protocols are essential. Users should be informed about what data is collected and how it is used.

When students and researchers rely heavily on LLMs for analysis or synthesis, they may bypass the cognitive processes that foster deep understanding. This shortcut can lead to superficial learning and reduced ability to evaluate information critically. Over time, the habit of deferring to AI may erode essential reasoning skills. Educators must design tasks that require independent thought and challenge learners to engage deeply with content. AI should be positioned as a support tool, not a substitute for thinking.

The use of LLMs in tutoring, feedback, and collaboration may reduce opportunities for meaningful human engagement. Students might prefer AI assistance over seeking help from teachers or peers, leading to isolation and diminished social learning. In research, reliance on AI tools could limit interdisciplinary dialogue and peer review. Human interaction is vital for emotional support, mentorship, and the exchange of diverse ideas. Institutions should foster environments that prioritize collaboration and community.

LLMs can unintentionally reinforce biases present in their training data or generate plausible-sounding but inaccurate information. This poses risks in educational and research contexts where accuracy and fairness are paramount. Users may unknowingly adopt biased perspectives or rely on flawed data. Addressing this issue requires diverse training datasets and ongoing evaluation of model outputs. Critical literacy skills are also essential to identify and challenge misinformation.

The authoritative tone of LLM-generated content may lead users to trust it without question. This overconfidence can result in the acceptance of incorrect or misleading information. In academic settings, such blind trust undermines the rigor of inquiry and peer validation. Users must be trained to verify AI outputs and cross-reference with reliable sources. Encouraging skepticism and review helps maintain academic standards.

The boundaries around ownership of AI-generated content remain unclear, raising ethical and legal concerns. Users may assume that content produced by LLMs is free to use, leading to potential violations of intellectual property rights. This confusion complicates citation practices and authorship claims. Institutions should develop clear policies on the use and attribution of AI-generated work. Legal frameworks must evolve to address these emerging challenges.

LLMs operate as black boxes, making it difficult for users to understand how outputs are generated. This lack of transparency can hinder accountability and trust in AI systems. In education and research, it is important to know the rationale behind conclusions or recommendations. Promoting explainable AI and user education can help demystify these tools. Greater transparency supports informed decision-making and ethical use.
Frequent use of LLMs for writing, summarizing, or problem-solving may lead to the erosion of foundational skills. Students and researchers might lose proficiency in composition, analysis, and synthesis. This decline affects long-term academic and professional development. To counteract skill atrophy, users should be encouraged to practice and refine their abilities regularly. AI should complement, and not replace, human effort.

LLMs trained predominantly on Western-centric data may misrepresent or oversimplify non-Western cultures, languages, and perspectives. This can lead to cultural bias and exclusion in educational content and research findings. Ensuring diverse representation in training data is crucial. Users should be aware of these limitations and seek to include multiple viewpoints. Promoting cultural sensitivity enhances the inclusivity and relevance of AI-supported work.
One of the most pressing challenges in the use of LLMs is the phenomenon of hallucination, where the model generates content that is factually incorrect, misleading, or entirely fabricated. This issue stems from the probabilistic nature of LLMs, which predict the next word based on patterns in training data rather than verifying factual accuracy. Hallucinations can manifest in various forms, including actual contradictions to known facts, invented citations, or responses that sound plausible but lack grounding in reality. Technologically, hallucinations arise due to limitations in model architecture, insufficient training data coverage, and the absence of real-time fact-checking mechanisms. The lack of grounding in external knowledge bases exacerbates this problem, especially when models are prompted to answer questions outside their domain expertise. Instruction-following models may also hallucinate when given vague or overly complex prompts, leading to responses that do not align with user intent. Another type of hallucination includes non-instructional content, where the model fails to follow directives and instead produces generic or irrelevant text. These issues pose significant risks in educational and research contexts, where accuracy and reliability are paramount. Hallucinations can mislead students, distort research findings, and undermine trust in AI systems. Addressing this challenge requires improved model training, integration with verified knowledge sources, and the development of mechanisms for output validation. Users must be educated on the limitations of LLMs and encouraged to critically evaluate AI-generated content.
Bias in LLMs is a multifaceted challenge that affects the fairness and inclusivity of AI-generated content. These models learn from vast datasets that often reflect societal prejudices, stereotypes, and imbalances. As a result, LLMs may produce outputs that reinforce gender, racial, cultural, or ideological biases. In educational settings, biased content can marginalize certain groups and perpetuate inequities. Technological factors contributing to bias include skewed training data, lack of diverse representation, and insufficient filtering mechanisms. Bias can also emerge in subtle ways, such as tone, framing, or omission of perspectives. Addressing bias requires deliberate efforts to curate inclusive datasets, implement fairness-aware training techniques, and continuously audit model outputs. Transparency in model development and user awareness are critical to mitigating bias. Educators and researchers must be vigilant in identifying biased content and promoting equitable use of AI tools. Ultimately, reducing bias in LLMs is essential to ensure that these technologies support rather than hinder diversity and inclusion in education and research.

Measuring the performance and knowledge of large language models (LLMs) is essential for ensuring their reliability, fairness, and usefulness in real-world applications especially in education and research. Without standardized benchmarks, it becomes nearly impossible to compare models, track progress, or identify areas for improvement. Reference databases serve as controlled environments where models are tested against curated tasks that reflect human reasoning, factual accuracy, and domain-specific expertise. These benchmarks help developers understand how well a model performs across different subjects, from basic arithmetic to complex medical reasoning. They also reveal weaknesses such as hallucination, bias, or poor instruction-following

behavior. For educators and researchers, these metrics offer transparency and confidence when integrating LLMs into workflows. A well-designed benchmark can simulate classroom tasks, scientific inquiry, or ethical decision-making, making it easier to evaluate whether a model is suitable for a given context. Moreover, metrics like accuracy, relevance, and reliability are not just technical, they directly impact how students learn and how researchers validate findings. As LLMs evolve, so must the benchmarks, ensuring that evaluation keeps pace with innovation. Ultimately, reference databases are the backbone of responsible AI deployment, guiding both development and usage toward higher standards.

Several reference databases have become foundational in assessing LLM capabilities. MMLU (Massive Multitask Language Understanding) tests models across 57 subjects, including law, medicine, and mathematics. It uses multiple-choice questions to evaluate both factual recall and reasoning. MMLU is especially valuable for gauging cross-domain general knowledge. Its broad scope makes it a go-to benchmark for academic and professional readiness. (Hendrycks et al. 2020) TruthfulQA focuses on how accurately and truthfully a model responds to questions designed to elicit common misconceptions. It challenges models to resist generating plausible but false information. This benchmark is critical for evaluating reliability in high-stakes domains like healthcare or journalism. It also helps identify tendencies toward hallucination or overconfidence. ARC (AI2 Reasoning Challenge) presents science questions from standardized tests, emphasizing logical reasoning and conceptual understanding. It includes both easy and difficult subsets to test nuanced performance. ARC is particularly useful in educational settings where reasoning matters more than rote memorization. It helps determine how well models can apply knowledge to solve problems. HellaSwag evaluates commonsense reasoning by asking models to complete sentences in a plausible way. The distractors are intentionally misleading, making the task challenging. This benchmark reveals how well models understand everyday scenarios and implicit human knowledge. It's a strong indicator of naturalistic reasoning ability. BIG-bench (Beyond the Imitation Game Benchmark) is a collaborative benchmark with over 200 tasks spanning ethics, linguistics, mathematics, and more. It's designed to push models beyond traditional NLP boundaries. The tasks vary in format and complexity, offering a panoramic view of model capabilities. BIG-bench is ideal for exploring emerging strengths and weaknesses in LLMs. OpenBookQA tests a model's ability to answer science questions using a small set of core facts. It requires multi-hop reasoning, combining known facts with general knowledge. This benchmark is excellent for evaluating synthesis and integration skills. It mimics how students might use textbooks and prior learning to answer complex questions. SuperGLUE builds on the original GLUE benchmark with more difficult tasks in language understanding. It includes textual entailment, question answering, and coreference resolution. SuperGLUE is a gold standard for evaluating deep linguistic comprehension. It's widely used to benchmark state-of-the-art NLP systems.

Metrics play a crucial role in evaluating the effectiveness and trustworthiness of large language models. Accuracy is one of the most fundamental metrics, indicating how often a model produces correct responses. Relevance measures how well the output aligns with the user's intent or the context of the query. Reliability assesses the consistency of a model's performance across different tasks and domains. These metrics help developers and users understand whether a model can be trusted in critical applications. Loss is another important metric, representing the difference between predicted and actual outputs during training; lower loss typically indicates better learning. (Ebert-Uphoff et al. 2021) Probabilistic distribution provides insight into the model's confidence in its predictions, helping to identify when a model is uncertain or prone to error. Together, these metrics form a comprehensive framework for evaluating LLMs. They guide model selection, fine-tuning, and deployment strategies. By analyzing these metrics, researchers can diagnose weaknesses and improve model architecture. Ultimately, robust metrics ensure that LLMs are not only powerful but also safe, fair, and aligned with human values.

Validating the outputs of large language models (LLMs) is a fundamental requirement for their responsible and effective use in education, research, and professional domains. These models generate text based on statistical patterns learned from vast datasets, which means they can produce content that sounds convincing but is factually incorrect, misleading, or ethically problematic. Without proper validation, users may unknowingly rely on hallucinated information, leading to flawed decisions, academic dishonesty, or the spread of misinformation. In educational settings, unverified outputs can compromise learning

outcomes, misguide students, and erode critical thinking skills. In scientific research, they can distort findings, misrepresent data, or propagate pseudoscientific claims. Verification acts as a safeguard, helping users distinguish between reliable insights and speculative noise. It also reinforces ethical standards, especially in contexts where accuracy, truthfulness, and fairness are paramount. As LLMs become more integrated into daily workflows, the need for robust validation mechanisms grows. Whether through automated systems or human oversight, verification ensures that AI-generated content aligns with real-world knowledge, respects contextual nuances, and supports informed decision-making. Ultimately, validation is not just a technical necessity, it is a moral and intellectual imperative that upholds the integrity of AI-assisted work.

Algorithmic verification refers to the use of computational techniques to assess the validity, consistency, and reliability of LLM outputs. These methods are designed to scale efficiently and provide rapid feedback, making them suitable for high-volume or real-time applications. One common approach is fact-checking against structured databases such as Wikidata, PubMed, or encyclopedic repositories. These sources offer verified information that can be used to cross-reference model-generated claims. Another method involves consistency testing, where the same prompt is rephrased and submitted multiple times to check if the model produces stable and coherent responses. Probabilistic scoring is also widely used, evaluating the confidence levels of predictions and flagging low-certainty outputs for further review. Semantic similarity analysis compares generated text to known correct answers using vector embeddings, helping identify deviations from expected content. Adversarial prompting is another strategy, where misleading or ambiguous inputs are used to expose weaknesses in reasoning and instruction-following. (Tekin et al. 2024) Some systems incorporate competing networks or reasoning models to validate outputs through logical inference. (Xu et al. 2025) These algorithmic tools can also detect stylistic anomalies, such as overly generic or repetitive phrasing, which may indicate low-quality generation. In educational platforms, automated grading systems use algorithmic checks to evaluate student submissions for correctness and originality. Despite their efficiency, algorithmic methods have limitations. They may struggle with context-sensitive tasks, cultural nuances, or ethical judgments. Moreover, reliance on incomplete or biased datasets can lead to flawed verification outcomes. To address these challenges, hybrid systems combine multiple algorithmic strategies, enhancing robustness and adaptability. Continuous learning mechanisms allow these systems to evolve based on user feedback and emerging data. Transparency is also crucial. Users should understand how verification algorithms work and what their limitations are. Algorithmic verification must also be domain-aware, adapting its logic to the specific requirements of fields like medicine, law, or education. Ultimately, algorithmic verification provides a scalable foundation for validating LLM outputs, but it must be complemented by human oversight to ensure comprehensive and context-aware evaluation.

Manual verification involves human reviewers assessing the accuracy, relevance, and ethical integrity of LLM-generated content. This process is indispensable for capturing subtleties that automated systems may overlook. Experts in specific domains, such as medicine, law, or education, bring contextual knowledge that enables deeper scrutiny of model outputs. Manual verification often includes cross-referencing with authoritative sources, evaluating logical coherence, and identifying potential biases or harmful language. Reviewers assess whether the tone and style of the response are appropriate for the intended audience, especially in sensitive or high-stakes contexts. In academic research, peer review serves as a form of manual validation, where scholars examine AI-generated hypotheses, summaries, or literature reviews. In customer service, human agents may verify chatbot responses before they are delivered to users, ensuring clarity and empathy. Manual verification also plays a critical role in multilingual applications, where cultural and linguistic nuances affect interpretation. It allows for ethical judgment, such as determining whether a response respects diversity and avoids discriminatory implications. In legal contexts, manual validation ensures compliance with regulatory standards and case law. In healthcare, clinicians review diagnostic suggestions from LLMs to confirm their validity and safety. Teachers use manual verification to assess the pedagogical value of AI-generated lesson plans, quizzes, or feedback. Despite its resource intensity, manual verification offers depth and nuance that algorithmic methods cannot match. It supports iterative improvement by providing feedback that helps fine-tune models and training data. Human reviewers can also identify emerging patterns of error or bias, contributing to long-term model refinement. Collaboration between reviewers and

developers fosters transparency and accountability in AI systems. Manual verification is especially important when dealing with controversial topics, where ethical sensitivity and contextual awareness are paramount. It ensures that AI outputs align with human values and societal norms. From the scope of the environmental impact, manual verification usually overperforms the automated solution, since constant iterations may require more and more deeper prompts and uses more memory and energy to produce the output (Cheung et al. 2025). Combining manual and algorithmic verification creates a balanced approach, leveraging the strengths of both to achieve reliable and trustworthy results. As LLMs become more sophisticated, the role of human judgment in verification will remain essential, guiding the responsible evolution of AI technologies.

One of the most fundamental challenges in training large language models (LLMs) is the weighting of the training data. If a model is exposed to a disproportionate number of publications that promote misleading or harmful ideas, such as 100 articles claiming smoking is not bad for health versus only 5 scientifically rigorous papers detailing its risks, the model may learn and reproduce the dominant but incorrect narrative. This imbalance can skew the model's understanding and lead to outputs that reflect popular misconceptions rather than scientific consensus. The problem is exacerbated when the model lacks mechanisms to prioritize high-quality, peer-reviewed sources over unverified or biased content. Without proper weighting, the model's outputs may reinforce misinformation, especially in domains like health, law, and education. This issue highlights the need for more sophisticated data curation strategies that account for source credibility and factual accuracy. Ultimately, the quality of a model's knowledge is only as good as the data it learns from, and improper weighting can undermine its reliability and trustworthiness.

Another major problem is the temporal context of the training data. For example, if a model is trained on materials from the Middle Ages, when geocentrism was the dominant belief, it may conclude that heliocentrism is a fringe or incorrect idea. Similarly, training a model on 19th-century texts could lead it to treat general relativity or the Higgs boson as speculative fiction rather than established science. This temporal bias can distort the model's understanding of scientific progress and lead to outputs that misrepresent current knowledge. The challenge becomes even more complex when considering concepts that are currently theoretical, such as the Dyson sphere. While this idea is science fiction today, it could become reality in the distant future. LLMs must navigate the fine line between speculative and verified knowledge, recognizing that today's fiction may be tomorrow's fact. This requires models to be context-aware and capable of signaling the epistemic status of the information they present. Moreover, it underscores the importance of aligning model training with the actual state of human knowledge, rather than attempting to exceed it. A model should reflect the best available understanding, not fabricate insights beyond the scope of its data.

A further issue is the blending of scientific facts with non-scientific claims in model outputs. Foundational concepts like heliocentrism, the existence of bacteria, general relativity, and the Higgs boson are supported by rigorous empirical evidence and form the backbone of modern science. Yet LLMs often present these alongside pseudoscientific ideas, science fiction tropes, or theories with contradictory evidence, without clearly signaling the difference. This can lead to confusion, especially for users unfamiliar with the nuances of scientific consensus. The model's training data may include both peer-reviewed research and unverified online content, which it treats with similar linguistic confidence. As a result, users may encounter outputs that blend truth with fiction, undermining trust in the model's reliability. This issue is compounded when the model is asked to explain complex scientific phenomena, where simplification can distort meaning or introduce inaccuracies. Without built-in mechanisms to prioritize verified knowledge, LLMs risk amplifying misinformation and misrepresenting the boundaries between science and speculation.

Another significant problem is the model's tendency to generate content that reflects cultural bias, outdated information, or fabricated details. For instance, when discussing controversial topics or emerging research, LLMs may fail to account for the latest findings or present a balanced view. They might cite studies that have been debunked or rely on sources that lack credibility. Additionally, the model's inability to express uncertainty means that speculative claims are often delivered with unwarranted confidence. This is particularly dangerous in fields like medicine, law, or history, where precision and context are critical. The presence of contradictory evidence in the training data can also lead to inconsistent outputs, where the model offers conflicting

explanations depending on how a question is phrased. Furthermore, the inclusion of science fiction and pseudoscientific narratives in the training corpus introduces stylistic and conceptual confusion. Users may receive responses that sound plausible but are rooted in fictional or fringe ideas. New problems are emerging as well, such as the model's difficulty in handling interdisciplinary questions that span multiple domains of knowledge. It may struggle to synthesize information accurately or misinterpret the relationships between concepts. These limitations highlight the need for improved data curation, clearer epistemic boundaries, and more robust verification systems to ensure that LLMs support informed and responsible use.

4. Conclusions

The integration of Large Language Models (LLMs) into the educational landscape marks a significant and irreversible transformation. As this paper has demonstrated, these tools offer powerful capabilities for students, teachers, and researchers, from personalized learning and curriculum development to accelerated research and enhanced content accessibility. LLMs can serve as invaluable assistants, streamlining repetitive tasks and providing novel ways to engage with information. However, the benefits are inextricably linked to a set of profound and complex challenges that demand careful consideration.

The analysis has revealed that core limitations, such as hallucination, bias, and a lack of real-time grounding in factual knowledge, are not mere technical glitches but inherent characteristics of the current probabilistic model architecture. These issues manifest in outputs that can be factually incorrect, culturally skewed, or misleadingly confident, posing significant threats to academic integrity and the development of critical thinking skills. We have shown that reliance on LLMs without critical oversight can lead to a decline in foundational skills and a dangerous overconfidence in unverified information.

To navigate this new environment, a dual approach is essential. First, developers must continue to innovate, creating more robust models with better data curation, transparent methodologies, and built-in verification mechanisms. The ongoing development of benchmarks and metrics is critical for objectively measuring and improving model performance. Second, and perhaps more importantly, the responsibility lies with the users in the education system. It is imperative that students, teachers, and researchers adopt a mindset of critical engagement and verification. The techniques of manual and algorithmic validation are not optional but are fundamental safeguards against the inherent flaws of LLMs.

In sum, the future of AI in education is not about replacing human intellect, but about augmenting it. By understanding both the roles and the limitations of LLMs, we can harness their potential while mitigating their risks. The path forward requires a collaborative effort to promote digital literacy, ethical use, and a balanced approach that champions human-led inquiry supported by, but not subservient to, artificial intelligence.

References

1. Hendrycks, D, Burns, C, Basart, S, Zou, A, Mazeika, M, Song, D & Steinhardt, J. 2020, 'Measuring massive multitask language understanding,' arXiv:2009.03300
2. Ebert-Uphoff, I, Lagerquist, R, Hilburn, K, Lee, Y, Haynes, K, Stock, J, Kumler, C, & Stewart, J Q. ' CIRA Guide to Custom Loss Functions for Neural Networks in Environmental Sciences' arXiv:2106.09757, 2021
3. Tekin, S F, Ilhan, F, Huang, T, Hu, S & Liu L 2024, 'LLM-TOPLA: Efficient LLM Ensemble by Maximising Diversity' arXiv:2410.03953
4. Xu, F, Hao, Q, Zong, Z, Wang, J, Zhang Y, Wang, J, Lan, X, Gong, J, Ouyang, T, Meng, F, Shao, C, Yan, Y, Yang, Q, Song, Y, Ren, S, Hu, X, Li, Y, Feng, J, Gao, C & Li Y 2025, ' Towards Large Reasoning Models: A Survey of Reinforced Reasoning with Large Language Models' arXiv:2501.09686
5. Cheung, K S, Kaul, M, Jahangirova, G, Mousavi, M R & Zie E 2025 ' Comparative Analysis of Carbon Footprint in Manual vs. LLM-Assisted Code Development' arXiv:2505.04521