LARGE LANGUAGE MODELS

IN EDUCATION

Foundations, Customization, Applications, and Responsible Use

Edited by Burcu Bektaş Güneş

BİDGE

BIDGE Publications

LARGE LANGUAGE MODELS IN EDUCATION: FOUNDATIONS, CUSTOMIZATION, APPLICATIONS, AND RESPONSIBLE USE

Editor: Burcu Bektaş Güneş

ISBN: 978-625-372-845-8

1st Edition

Page Layout By: Gozde YUCEL

Publication Date: 20.11.2025

BIDGE Publications

www.bidgeyayinlari.com.tr - bidgeyayinlari@gmail.com
Krc Bilişim Ticaret ve Organizasyon Ltd. Şti.
Güzeltepe Mahallesi Abidin Daver Sokak Sefer Apartmanı No: 7/9
Çankaya / Ankara

Contents

FOUNDATIONS OF LARGE LANGUAGE MODELS AND THEIR EDUCATIONAL POTENTIAL

ÜMİT MURAT AKKAYA[1]

Introduction

We are in the midst of a technological shift in education, one unfolding at a pace that rivals the introduction of the personal computer or the internet. Large Language Models (LLMs) have evolved in just a few short years from academic curiosities into powerful, general-purpose tools. They are capable of generating nuanced, human-like text, engaging in complex Socratic dialogue, translating languages with remarkable fluency, and even writing functional computer code (Wang & ark., 2024). This rapid ascent from research labs to public-facing tools has profound implications for the educational landscape.

These models offer the potential to finally deliver on the long-held promise of scalable, personalized, and universally accessible learning experiences. They can act as tireless tutors, available 24/7 to answer questions, explain concepts in different ways, and adapt to a student's individual pace (Kumar & ark., 2025).

[1] Research Assistant, Gebze Teknik Üniversitesi, Bilgisayar Mühendisliği, Orcid: 0000-0002-5247-4860

They can serve as adaptive content generators, creating customized problem sets or reading materials, and as powerful creative partners for both students and teachers, brainstorming ideas or drafting lesson plans (Baidoo-Anu & Ansah, 2023). Simultaneously, their capabilities present a new and complex set of pedagogical, logistical, and ethical challenges. These range from the immediate concerns of academic integrity and the "outsourcing" of critical thinking to the deeper systemic issues of data privacy, algorithmic bias reinforcing existing inequities, and the significant environmental and financial costs of their creation and operation (Farooqi & ark., 2024; Holmes & Tuomi, 2022).

This chapter lays the foundation for understanding these remarkable yet challenging systems. We will try to solve the mistery of what LLMs are, tracing their lineage from the rule-based systems of traditional natural language processing (NLP) to the revolutionary architecture that powers them today. We will explore the specific "emergent" capabilities that make them uniquely suited for educational applications, compare the ecosystem of models available, and critically examine their inherent and often misunderstood limitations. To build effective, safe, and equitable educational tools, we must first understand the base on which they stand.

What is a Large Language Model?

At its core, a Large Language Model is a sophisticated statistical tool. It is a massive neural network, often containing hundreds of billions or even trillions of "parameters"—the values in the network that are adjusted during training. These parameters function as the model's repository of learned knowledge. These models are trained on vast, petabyte-scale quantities of text and code scraped from the internet, books, and other sources, a dataset that represents a significant portion of all human-generated text.

Its primary function is deceptively simple: to predict the next word (or, more accurately, a sub-word unit called a "token") in a sequence, given the words that came before it. Given the prompt "The capital of France is," the model calculates a probability distribution over all the words it knows, and "Paris" will have a very high probability. The model learns the statistical patterns of language (grammar, syntax, facts, reasoning styles, and even biases) by repeatedly performing this "next-token-prediction" task. This simple objective, when applied at an unprecedented scale, gives rise to the "emergent abilities" we observe, such as translation, summarization, and question-answering, none of which were explicitly programmed into the system (Wei & ark., 2022).

From Traditional NLP to the Transformer

For decades, NLP systems were brittle and complex. They often relied on hand-crafted linguistic rules and grammars, which required immense domain-specific expertise and would fail when encountering slang, metaphors, or grammatical structures they hadn't seen. Later statistical methods, while an improvement, still struggled with long-range context. The entire field was revolutionized in 2017 (Ladu & ark., 2025).

Tablo 1 A Comparison of Traditional NLP and Modern LLM Approaches

Approach	Primary Mechanism	How it Understands Context	Key Limitations
Rule-Based NLP	Hand-crafted grammars and rules (e.g., "If word X follows word Y, it is a verb")	Defined explicitly by human linguists	Extremely brittle; fails on novel language; cannot scale.
Statistical NLP	Methods like "bag-of-words" (TF-IDF) that count word frequencies.	Poor. Ignores word order (e.g., "dog bites man" vs. "man bites dog").	Lacks understanding of syntax, semantics, or context.
Recurrent/LSTM	Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks (Hochreiter & Schmidhuber, 1997).	Processes text sequentially, maintaining a "memory" of previous words.	Struggles with long-range dependencies; computationally slow (cannot be parallelized).
Transformer (LLM)	The Transformer architecture (Vaswani & ark., 2017) using self-attention mechanisms.	Processes all words at once, weighing the importance of every word to every other word.	Highly scalable; captures complex, long-range context. (Limitations are cost, bias, etc.)

Source: Created by the author

The true breakthrough came with the Google Brain paper "Attention Is All You Need" (Vaswani & ark., 2017). This paper introduced the Transformer architecture, which completely replaced the sequential processing of RNNs with a mechanism called self-attention.

The Transformer's key innovation is its ability to process all tokens in a sequence simultaneously and use self-attention to weigh the importance of all other words when processing any given word. In simple terms, for every word, the model creates a "query," "key," and "value" vector. The "query" represents the current word's request for information. It compares this query to the "key" of every other word in the sequence to determine relevance (the "attention score"). It then uses these scores to create a weighted sum of the "values," resulting in a new representation of the word that is deeply context-aware.

This attention mechanism learns, for example, that in the sentence "The student opened the book," the word "book" is highly relevant to "opened," but in "The student will book a flight," it is highly relevant to "flight." This ability to dynamically model context over long distances, and the fact that its computations could be massively parallelized on modern GPUs, made the Transformer the engine of the new LLM era.

The GPT Family, Scale, and Emergence

The Transformer architecture set the stage, but scale provided the breakthrough. Researchers discovered predictable "scaling laws," which showed that as you predictably increase model size (parameters), data volume, and computational budget, the model's performance on a wide range of tasks predictably improves (Kaplan & ark., 2020). More parameters mean the model has a greater capacity to memorize information and, more importantly, to learn complex, abstract patterns and relationships within the data.

Models in the GPT (Generative Pre-trained Transformer) family, from GPT-2 to GPT-3 (Brown & ark., 2020) and GPT-4, were a direct test of these scaling laws. As these models grew, they began to exhibit emergent capabilities, abilities that were not present in

smaller models and were not the direct object of training, appearing almost spontaneously once a certain threshold of scale was passed.

The most important of these was "in-context learning". For instance, GPT-3 was found to be a "few-shot learner" (Brown & ark., 2020). Instead of needing to be explicitly re-trained (fine-tuned) for a new task, you could simply show it a few examples of the task in its prompt, and it would understand and perform that task. This shift from a paradigm of task-specific training to one of general-purpose prompting is the defining feature of modern foundation models.

The LLM Training Pipeline: Pre-training, Instruction Tuning, and Alignment

Modern LLMs are not created in a single step. Their capabilities are the result of a sophisticated, multi-stage training pipeline.

- **Pre-training:** This is the foundational and most computationally expensive phase. The model is trained on a massive, unlabeled dataset (e.g., a large portion of the internet and all of Wikipedia) to do next-token prediction. The result of this stage is a "base model". This base model is a powerful knowledge repository and text completer, but it is not "helpful"; it is not trained to follow instructions or hold a conversation. If you gave it the prompt "Write a summary of Hamlet," it might just complete the sentence with "and Macbeth are two of Shakespeare's most famous tragedies".
- **Instruction Fine-Tuning (IFT):** To make the base model useful, it undergoes "instruction tuning", also called Supervised Fine-Tuning (SFT). In this phase, the model is fine-tuned on a smaller, high-quality dataset of "instructions" and "high-quality answers" (e.g., Prompt: "Summarize Hamlet," Answer: "Hamlet is a tragedy about..."). This dataset is often curated by human labelers. This teaches the

model to follow user commands and behave as a helpful assistant.

- **Alignment:** This final phase aims to make the model more helpful, truthful, and harmless. The most common method is Reinforcement Learning from Human Feedback (RLHF) (Ouyang & ark., 2022). In this process, several of the model's answers to a prompt are shown to human raters, who rank them from best to worst. A separate "reward model" is then trained to predict these human preferences. Finally, the LLM itself is fine-tuned using reinforcement learning to maximize the "reward" score, effectively aligning its behavior with what human raters preferred. Newer methods, like Constitutional AI (Bai & ark., 2022), attempt to achieve the same goal using AI-driven feedback based on a written constitution of principles, reducing the human labor bottleneck.

Key Capabilities for Education

The shift from task-specific models to general-purpose foundation models, accessed via prompting, is what makes LLMs so disruptive for education. A single, powerful model can be prompted to perform hundreds of different educational tasks, from tutoring to content creation.

In-Context Learning: Zero-Shot, Few-Shot, and Instruction Tuning

- **Zero-Shot Learning:** This is the model's ability to perform a task without any prior examples, thanks to its instruction-tuning. It can follow commands it has never seen before. Example: "Explain the water cycle to a 5th grader, using an analogy. Make sure to define the terms 'evaporation' and 'condensation'."

- **Few-Shot Learning:** For more complex or nuanced tasks, you can provide a few examples directly in the prompt. This "primes" the model, demonstrating the pattern or format you want it to follow. Example (Few-Shot):

Prompt: Convert the student's question into a Socratic inquiry.

Student: What's the answer to question 5? Tutor: What do you think the first step to finding the answer might be?

Student: Why is the sky blue? Tutor: That's a great question. What have you observed about the sky at different times of the day?

Student: I don't understand why the mitochondrion is called the powerhouse of the cell. Tutor:

The model recognizes the pattern (Student question -> Socratic tutor response) and will reliably generate a Socratic response, such as: "What part of the "powerhouse" metaphor is confusing you?"

Chain-of-Thought Prompting

One of the most significant recent discoveries is Chain-of-Thought (CoT) prompting (Wei & ark., 2022). Researchers found that by simply instructing a model to "think step-by-step" before giving a final answer, its accuracy on complex reasoning tasks (like math word problems or logic puzzles) increased dramatically. This simple prompt tweak coaxes the model to generate a "chain of thought," breaking down the problem and modeling the intermediate reasoning steps.

- **Standard Prompt:** "Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. Then

he gives 2 to a friend. How many tennis balls does he have left? A: 9." (Correct answer, but no reasoning)

- **CoT Prompt (Example):** "Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. Then he gives 2 to a friend. How many tennis balls does he have left? A: Let's think step-by-step. Roger starts with 5 balls. He buys 2 cans of 3 balls each, so 2 * 3 = 6 balls. He now has 5 + 6 = 11 balls. He then gives 2 balls to a friend. 11 - 2 = 9 balls. The answer is 9." (Correct answer, with reasoning)

For education, this is revolutionary. It allows the model to not only provide an answer but also to model the reasoning process itself. It suggests that LLMs may be capable of more than simple pattern-matching, perhaps engaging in a form of zero-shot reasoning (Kojima & ark., 2022). Advanced techniques, such as self-consistency, take this further by having the model generate multiple reasoning paths and then picking the most consistent answer, which further boosts accuracy (Wang & ark., 2022). For a student, this means they can see how an answer was derived, making the LLM a potential tool for metacognitive instruction.

Role-Playing and Simulation for Immersive Learning

A powerful extension of in-context learning is the model's ability to adopt a persona or simulate a scenario. By "priming" the model with a role, it can create immersive learning experiences that were previously impossible to scale.

- **Socratic Tutor:** A prompt can instruct the model to act as a Socratic tutor, never giving the answer but always responding to a student's question with another guiding question.

- **Historical Simulation:** A student can "interview" a historical figure. (e.g., "You are Julius Caesar. I am a reporter. Why did you decide to cross the Rubicon?").
- **Practice Scenarios:** A medical student can practice diagnosing a patient, a law student can practice cross-examining a witness, or a business student can practice a difficult negotiation.
- **Language Learning:** A student learning French can have a full, immersive conversation. (e.g., "You are a French baker. I want to buy a croissant. Please respond only in French and correct my grammar if I make a mistake.").

These simulations provide a safe "practice field" for students to apply knowledge and develop soft skills.

Content Generation and Augmentation

Perhaps the most immediate use case for educators is the model's ability to generate and augment content. This can significantly reduce teacher workload and provide differentiated materials.

- **Assessment Creation:** "Generate 10 multiple-choice questions about the causes of the American Civil War, suitable for a 10th-grade history class. Include an answer key."
- **Rubric Design:** "Create a detailed 6-point rubric for a persuasive essay on renewable energy, with criteria for 'Thesis,' 'Evidence,' and 'Clarity'."
- **Lesson Planning:** "Generate a 50-minute lesson plan for introducing Python 'for loops' to high school students, including a warm-up activity and a short homework assignment."

- **Differentiated Texts:** "Take this news article about a new scientific discovery and rewrite it at a 5th-grade reading level."

This capability positions the LLM as a "co-pilot" for teachers, freeing them from administrative tasks to focus on high-touch student interaction.

Model Types: Open-Source vs. Proprietary

As LLMs become critical infrastructure, it's important to understand the landscape. The models available today fall broadly into two categories, each with critical trade-offs for educational institutions.

Tablo 2 Strategic Comparison of Proprietary vs. Open-Source Models

Feature	Proprietary (Closed) Models	Open-Source Models
Examples	GPT-4 (OpenAI), Claude 3 (Anthropic), Gemini (Google)	LLaMA 3 (Meta), Mistral (Mistral AI), BLOOM (BigScience) (Kojima & ark., 2022)
Access	Paid API (e.g., $ per 1,000 tokens)	Downloadable weights (e.g., from Hugging Face)
Performance	Pro: Generally state-of-the-art; highest performance.	Con: Often lag 6-12 months behind the best proprietary models.
Ease of Use	Pro: Very easy; no infrastructure to manage.	Con: Requires significant MLOps, hardware, and technical expertise to host.
Data Privacy	Con: Major risk. Data is sent to a third party.	Pro: Fully customizable; can be deeply fine-tuned on custom data.
Customization	Con: Limited to provider's API (e.g., basic fine-tuning).	Pro: Fully customizable; can be deeply fine-tuned on custom data.
Cost	Con: Ongoing, variable "per-use" cost. Can become very expensive at scale.	Pro: No "per-use" cost. Con: High upfront and maintenance (hardware, talent).
Transparency	Con: "Black box." Training data and architecture are secret.	Pro: Auditable. Researchers can inspect weights and (often) data.

Source: Created by the author

Proprietary (Closed) Models

These are the most powerful, state-of-the-art models, developed and controlled by large corporations. They are accessed via an API, which integrates into the "API economy." Their primary benefits are their unmatched performance and ease of use. Their primary drawbacks are the lack of control, vendor lock-in, and significant data privacy risks (La Malfa & ark., 2024).

Open-Source Models

These models are released publicly (or "open-weighted"), allowing anyone to download, inspect, modify, and run them. Their primary benefits are data sovereignty (critical for student data), customizability, and lack of ongoing per-use costs. Their drawbacks are the high technical and hardware overhead required to run them effectively (Lin & ark., 2024; Touvron & ark., 2023).

The Strategic Choice for Education

This "open vs. proprietary" split presents a fundamental strategic choice for education. An individual teacher might leverage a proprietary API for its convenience and power in creating classroom materials. A school district or university, however, must heavily weigh the data privacy risks of sending sensitive student data to a third party.

Regulations like the Family Educational Rights and Privacy Act (FERPA) in the United States and the General Data Protection Regulation (GDPR) in Europe impose strict rules on how student data (personally identifiable information) is collected, stored, and used (Emon & Chowdhury, 2025). Sending prompts that contain student names, writing, or learning analytics to a third-party API provider may violate these regulations unless a very specific, and often expensive, data-processing agreement is in place. For many public institutions, investing in self-hosted, open-source models may

be the only long-term, legally-compliant path to providing custom AI tools to their students.

Critical Limitations and Common Misconceptions

Despite their power, LLMs are not magical. They are tools with fundamental, deep-seated limitations. Understanding these limitations is the non-negotiable first step toward responsible implementation in education.

The "Stochastic Parrot": Understanding vs. Prediction

The most common misconception is that LLMs "understand" the world. This is a topic of intense debate. The "Stochastic Parrot" critique argues that LLMs are masters of statistical pattern-matching (Bender & ark., 2021). They have learned the relationships between words and concepts from their training data, but they lack true human-like comprehension, consciousness, grounding in reality, or intent. An LLM knows the word "apple" is statistically associated with "red," "fruit," and "pie," but it has never seen an apple, tasted one, or felt its weight.

This is a modern version of the "grounding problem" in AI (Harnad, 1990). The model's "understanding" is ungrounded from physical reality, sensory experience, or social interaction. This is why it can make errors of common sense that no human child would.

A counter-argument, highlighted by recent research, is that the "sparks" of general intelligence seen in models like GPT-4 are so advanced that the distinction between "real" understanding and "mere" complex pattern-matching becomes philosophically and practically blurred (Bubeck & ark., 2023). Regardless of this debate, it is safest to assume that models do not "know" or "believe" anything; they generate text based on learned probabilities.

Confident Plausibility: Why LLMs "Sound Right but Are Wrong"

The primary limitation of an LLM, and its greatest danger in an educational context, is the hallucination. A hallucination is when the model generates a response that is nonsensical, factually incorrect, or untethered from reality, but presents it with the same confident, plausible-sounding, and grammatically perfect language it uses for correct information (Ji & ark., 2023).

Example: A student asking for sources for a paper on the American Revolution might receive a list of five perfectly formatted APA citations, complete with authors, titles, and journal names. However, upon inspection, three of the five articles, and perhaps even one of the cited authors, do not exist.

This happens because the model's objective is not to be truthful; it is to be statistically likely. If a factual error is a "likely" sequence of words based on its training, it will generate it. It has no internal "truth-checker" or world model to cross-reference. This is why a student's reliance on an LLM as a sole source of truth is so dangerous.

Knowledge Cutoffs and Embedded Bias

LLMs suffer from two other major content flaws:

- **Static Knowledge:** Most LLMs are *"frozen in time."* Their knowledge is limited to the data they were trained on and ends at a specific "cutoff date" (e.g., *"knowledge cutoff April 2023"*). They are not aware of current events.
- **Embedded Bias:** LLMs are trained on a snapshot of the internet, which is replete with human biases. Models have been shown to reproduce and even amplify societal biases related to gender, race, and culture (Bolukbasi & ark., 2016). For example, a model might consistently generate

text associating *"doctor"* with male pronouns and *"nurse"* with female pronouns, or produce more negative text when discussing certain demographic groups, or associate *"inner-city"* with crime.

Data Privacy and Security Risks

When using a proprietary API, all data including potentially sensitive student questions, essays, and personal reflections is sent to a third-party server. This raises critical privacy concerns:

- Data Use for Training: Does the API provider have the right to use this data to train its future models?
- Compliance: Does this data transfer comply with stringent educational privacy laws like FERPA or GDPR? (Emon & Chowdhury, 2025)
- Data Breaches: The API provider becomes a high-value target for data breaches, potentially exposing student information.

These risks are a primary driver for institutions to consider open-source, self-hosted alternatives.

Environmental and Financial Costs

The "large" in Large Language Model has significant real-world costs.

- **Environmental Cost:** Training a single large foundation model requires an immense amount of computational power, consuming vast quantities of electricity and generating a significant carbon footprint, comparable to the annual emissions of hundreds of cars (Strubell, Ganesh & McCallum, 2020).

- **Financial Cost:** The hardware required for this training (thousands of specialized GPUs) costs tens or even hundreds of millions of dollars. This cost creates a high barrier to entry, concentrating power in the hands of a few large tech companies and making it difficult for academic or public-sector institutions to build their own models from scratch.

These costs are a crucial, if often invisible, limitation that shapes the entire field, driving the "open vs. closed" debate and the economics of API access.

The Context Window Limitation

A key technical limitation with major practical consequences is the "context window." This is the fixed amount of text (measured in tokens) that a model can "see" or process at one time. For older models, this was very small (e.g., 2,000 tokens, or ~1,500 words). Newer models have much larger windows (100,000 tokens or more), but they are still finite.

This has direct implications for education:

- A model with a small context window cannot read an entire textbook chapter, a long research paper, or even a full student essay before commenting on it.
- In a long tutoring conversation, the model will eventually *"forget"* what was said at the beginning.

This limitation is the primary motivation for techniques like Retrieval-Augmented Generation (RAG). RAG is a method for "feeding" the model relevant information just-in-time from an external knowledge base (like a textbook) so it can answer questions without having to fit the entire book into its context window.

Conclusion:

Large Language Models represent a new type of foundational infrastructure. They are not just another ed-tech "app" but a new utility, much like a search engine or a word processor, upon which a new generation of tools will be built. We have seen that they are built on the Transformer architecture, powered by scaling laws, and made usable through a complex pipeline of pre-training and alignment. Their capabilities, from zero-shot learning to chain-of-thought, simulation, and content generation, open up new pedagogical frontiers.

We have also seen that they are a double-edged sword. The "open vs. proprietary" models that define the market represent a crucial choice between power and privacy. Their fundamental limitations their nature as pattern predictors, not truth-tellers; their capacity for confident hallucination; their embedded biases; their finite context windows; and their significant privacy and environmental costs are the central challenges we must address.

The potential for these models to "sound right but be wrong" and to perpetuate hidden biases requires us to move forward with a mindset of critical, informed, and cautious optimism. With this foundation established, we are ready to move from the what to the how.

References

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073.

Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. Journal of AI, 7(1), 52–62.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623).

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Advances in Neural Information Processing Systems, 29.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., et al. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712.

Emon, M. M. H., & Chowdhury, M. S. A. (2025). Safeguarding student data: Privacy and security challenges in AI-powered education tools. In Enhancing Student Support and Learning Through Conversational AI (pp. 191–228). doi:10.4018/979-8-3373-3316-8.CH008

Farooqi, M. T. K., Amanat, I., & Awan, S. M. (2024). Ethical considerations and challenges in the integration of artificial intelligence in education: A systematic review. Journal of Excellence in Management Sciences, 3(4), 35–50.

Harnad, S. (1990). The symbol grounding problem. Physica D: Nonlinear Phenomena, 42(1–3), 335–346.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

Holmes, W., & Tuomi, I. (2022). State of the art and practice in AI in education. European Journal of Education, 57(4), 542–570.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., et al. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., et al. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. Advances in Neural Information Processing Systems, 35, 22199–22213.

Kumar, H., Rothschild, D. M., Goldstein, D. G., & Hofman, J. M. (2025, July). Math education with large language models: Peril or promise? In International Conference on Artificial Intelligence in Education (pp. 60–75). Cham: Springer Nature Switzerland.

La Malfa, E., Petrov, A., Frieder, S., Weinhuber, C., Burnell, R., Nazar, R., et al. (2024). Language-models-as-a-service: Overview of a new paradigm and its challenges. Journal of Artificial Intelligence Research, 80, 1497–1523.

Ladu, N. S. D., Turyasingura, B., Willbroad, B., & Atuhaire, A. (2025). The rise of transformers: Redefining the landscape of

artificial intelligence. Babylonian Journal of Artificial Intelligence, 72–76.

Lin, M. P. C., Chang, D., Hall, S., & Jhajj, G. (2024, June). Preliminary systematic review of open-source large language models in education. In International Conference on Intelligent Tutoring Systems (pp. 68–77). Cham: Springer Nature Switzerland.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.

Strubell, E., Ganesh, A., & McCallum, A. (2020). Energy and policy considerations for modern deep learning research. Proceedings of the AAAI Conference on Artificial Intelligence, 34(9), 13693–13696.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., et al. (2023). LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

Wang, S., Xu, T., Li, H., Zhang, C., Liang, J., Tang, J., et al. (2024). Large language models for education: A survey and outlook. arXiv preprint arXiv:2403.18105.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., et al. (2022). Self-consistency improves chain-of-thought reasoning in language models. arXiv preprint arXiv:2203.11171.

Wei, J., Tay, Y., Bommasani, R., & Le, Q. V. (2022). Emergent abilities of large language models. Transactions on Machine Learning Research.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824–24837.

CUSTOMIZATION AND DEPLOYMENT OF LLMS FOR EDUCATION

İLHAN AYTUTULDU[1]

Introduction

Large Language Models (LLMs) such as GPT-4, Claude, Gemini, and LLaMA 3 have rapidly evolved from research curiosities into indispensable tools across domains. Yet, their adoption in education introduces a distinctive set of opportunities and constraints. Unlike general-purpose deployments, educational settings require models that are not only accurate but also pedagogically aligned, transparent in reasoning, and safe for learners (Xu et al., 2024: 7; Zhao & Wan, 2025: 22). Off-the-shelf LLMs are typically trained on heterogeneous internet text, which means that their linguistic richness and reasoning power come without guarantees of curricular coherence or age-appropriate instruction. As a result, direct use of generic models can lead to inconsistent learning outcomes, factual drift, or even unintentional bias in student interactions (Lee et al., 2024; Delikoura & Hui, 2025).

[1] Res. Assist. Dr., Gebze Technical University, Department of Computer Engineering, Orcid: 0000-0003-4237-8442

Customization bridges this gap between general intelligence and domain-specific pedagogy. By shaping the model's inputs, outputs, and underlying parameters, educators and developers can tailor LLM behavior to reflect curriculum standards, disciplinary depth, and cultural context (Beale, 2025: 8; Zhang et al., 2025: 12; Wang et al., 2023: 6). Techniques such as prompt engineering, retrieval-augmented generation (RAG), and parameter-efficient fine-tuning enable this adaptation at varying levels of complexity and cost. At the same time, effective deployment demands careful attention to data governance, privacy, and sustainability—issues particularly sensitive when dealing with minors or institutional learning data (Dong & Xie, 2024: 5; Shan, 2025: 3; Zhao et al., 2024: 11).

This chapter explores the continuum between customization and deployment of LLMs for education. It begins by examining why educational domains necessitate specialized adaptation strategies and how pedagogical theories inform technical design. It then details practical customization approaches—from lightweight prompting frameworks to full fine-tuning pipelines—and demonstrates how these can be operationalized through scalable deployment architectures. Real-world case studies highlight systems for tutoring, assessment, and curriculum alignment. The discussion concludes with best practices for privacy, transparency, and sustainability, as well as a forward-looking perspective on personalized and multimodal educational AI.

Ultimately, the chapter argues that the educational value of large language models lies not in their generative power alone but in degree of alignment between computational design and human teaching principles. When properly customized and responsibly deployed, LLMs can act as adaptive, inclusive, and context-aware partners that extend—not replace—the expertise of human educators.

Understanding Customization Needs in Education

Adapting large language models to educational contexts is fundamentally a domain adaptation problem. Educational discourse differs sharply from the open-domain data on which LLMs are trained: it is goal-oriented, hierarchical, and evaluation-driven (Ke, Ming & Joty, 2025: 6; Afzal et al., 2024: 9; Sonkar et al., 2024: 4). Tasks such as question generation, automated grading, and concept explanation impose structured reasoning constraints that require the model to control both content scope and reasoning style. From a technical perspective, customization therefore involves three interrelated goals: (1) domain conditioning, to align model embeddings with curriculum-specific terminology and knowledge hierarchies; (2) didactic alignment, to model reasoning transparency and stepwise feedback patterns; and (3) controlled variability, to balance consistency and creativity in generated outputs (Liu et al., 2025: 8; Imperial et al., 2024: 5; Jiao et al., 2023: 480; Bhat et al., 2022: 3).

These objectives extend beyond surface fine-tuning. They require combining retrieval-grounded inference, parameter-efficient tuning, and prompt-level control to ensure factual accuracy, reproducibility, and cultural localization. In multilingual or region-specific deployments, additional layers such as bilingual adapters or localized tokenizers maintain linguistic fidelity and contextual relevance. Consequently, educational customization is not merely an interface problem but a multi-layered optimization process that integrates representation learning, controlled generation, and policy-based interaction—laying the technical foundation for the customization and deployment strategies detailed in the following sections.

Approaches to Customizing LLMs for Education

Once the need for adaptation is established, the next challenge is to determine *how* large language models can be effectively customized for educational applications. The available methods span a continuum of technical depth—from lightweight prompt engineering to parameter-efficient fine-tuning and retrieval-augmented generation (RAG). Each approach represents a different trade-off between controllability, cost, latency, and generalization capability (Pan et al., 2025: 805; Jain et al., 2025: 7; Ye, 2025: 192).

Prompt Engineering and Template Control

Prompt engineering is the most accessible and cost-effective approach to controlling LLM behavior without retraining. It is conceptually analogous to programming through natural language instructions—where the prompt defines both the task and the desired response structure. Within educational contexts, prompt engineering governs pedagogical tone, reasoning depth, and structural consistency in generated responses.

At its core, a prompt can be decomposed into three layers of control:

1. **Role Definition** – specifying the model's instructional persona (e.g., "You are a high-school mathematics tutor").
2. **Task Framing** – defining goals, reasoning constraints, and expected outputs (e.g., "Explain in steps before revealing the answer").
3. **Output Structuring** – ensuring responses conform to predictable formats (JSON, Markdown, or class-based schemas).

Instructional Role Control

By assigning **system-level instructions**, developers can enforce behavioral constraints that shape the model's response distribution. This mechanism, known as **role conditioning**, modifies the model's latent state before token generation—effectively biasing it toward a desired conversational style or reasoning depth. For example, specifying the system message as "Act as a Socratic question generator" guides the model to prioritize interrogative patterns and withhold final answers until reasoning steps are complete.

system_prompt *= """ You are an AI teaching assistant for undergraduate Operating Systems. Adopt a supportive tone, explain concepts step by step, and encourage reasoning before providing final answers."""*

user_prompt *= "Explain how semaphores prevent race conditions."*

Explanation:

This prompt enforces persona persistence and pedagogical consistency. By explicitly defining role and response behavior, the model internalizes a constrained discourse pattern, improving reproducibility across sessions.

Template Design and Structured Prompts

Templates extend role prompting by adding *modular control*. Each educational task—question generation, feedback generation, grading—can have a standardized template. This creates a reproducible pattern that enforces uniform style and evaluation logic.

TEMPLATE = """

Learning Objective

{objective}

Explanation

{explanation}

Example Question

{example}

Common Mistakes

{mistakes}

"""

Developers can instantiate this template dynamically with programmatic variables:

filled_prompt = TEMPLATE.format(**objective=** "Understand the concept of deadlock in operating systems.", **explanation=**"A deadlock occurs when two or more processes wait for resources held by each other.", **example=**"Describe a real-world analogy for deadlock.", **mistakes=**"Forgetting to mention mutual exclusion condition. Confusing deadlock with starvation.")

Structured Output and Typed Schema Control

Free-form outputs limit integration with downstream analytics systems. To make results machine-readable, models can be instructed to produce structured JSON or class-based outputs validated by Python's data classes or pydantic models. This design pattern—typed response control—bridges prompt engineering and software engineering.

from pydantic import BaseModel

class LessonOutput(BaseModel):

learning_objective: str

hints: list[str]

solution_steps: list[str]

evaluation_criteria: str

This schema-driven method transforms generative models into deterministic components that emit predictable objects. The same principle scales to grading pipelines and dashboards where each field corresponds to rubric items or assessment criteria.

Chaining and Multi-Stage Prompt Composition

Complex educational workflows, such as question generation followed by rubric evaluation, can be realized through **prompt chaining**. Here, outputs from one model call become inputs to another—forming a directed sequence of reasoning tasks. Libraries such as *LangChain* or *LlamaIndex* support these pipelines. This models *modular reasoning*. Each prompt has a well-defined contract and output type, mirroring software modularization principles.

def generate_question(concept):

return f"Generate a multiple-choice question about {concept}."

def grade_response(question, answer):

return f"Evaluate the following answer using Bloom's taxonomy.\nQuestion: {question}\nAnswer: {answer}"

question = generate_question("Virtual Memory")

evaluation = grade_response(question, "It allows more processes to be executed using disk as RAM.")

One-Shot and Few-Shot Prompting

Another dimension of prompt engineering concerns the number of exemplars provided within the input context. Modern LLMs can perform in-context learning, adapting to a task

by observing examples embedded directly in the prompt rather than through gradient updates or fine-tuning.

When a single demonstration is included, the configuration is termed one-shot prompting; when several demonstrations are supplied, it becomes few-shot prompting. In both cases, the model implicitly constructs a task representation from these exemplars, conditioning its token-level probability distribution to mirror the structure and reasoning patterns present in the examples. This mechanism allows developers to approximate small-scale supervised learning purely through context manipulation, providing an efficient alternative to parameter training.

In educational applications, few-shot prompting is highly effective for rubric-based grading, question generation, and concept explanation tasks. Each embedded exemplar defines both the expected reasoning depth and the output format, enabling consistent behavior without modifying model weights. One-shot setups are useful for simple classification or scoring tasks; few-shot configurations provide stronger generalization and stability by exposing the model to intra-task variation.

few_shot_prompt = """

Evaluate student answers according to the rubric.

Example 1

Q: Define an operating system.

A: Software that manages hardware and resources → Grade: 10

Example 2

Q: What is a race condition?

A: Two processes accessing shared data simultaneously → Grade: 8

Now evaluate:

Q: Explain the purpose of semaphores.

A: They synchronize process access to shared memory."""

In this prompt, the model infers grading criteria from the in-context examples and applies the learned structure to the new query, performing **few-shot inference** without retraining. The approach combines the interpretability of explicit exemplars with the flexibility of generative reasoning, making it especially suitable for prototype systems where labeled data are limited.

Automatic Prompt Optimization

After designing a base prompt or few-shot template, the next step is to automatically improve it using measurable feedback. This process, called Automatic Prompt Optimization (APO), treats the prompt itself as a variable that can be optimized without changing the model's parameters. The goal is to find a prompt that produces the best results for a given task—such as grading accuracy, factual correctness, or rubric alignment.

In simple terms, APO runs a **search loop**:

1. Start with an initial prompt.
2. Generate small variations (called *mutations*) by rewording, adding examples, or changing structure.
3. Test each new version on a small evaluation dataset.
4. Keep the version that performs best and repeat.

In educational systems, automatic prompt optimization can fine-tune grading templates, tutoring responses, or feedback structures based on teacher evaluations or real student data. It provides a practical way to adapt LLM behavior to institutional goals without retraining or fine-tuning the underlying model.

Retrieval-Augmented Generation (RAG) for Educational Systems

While prompt engineering governs how a model reasons within its internal knowledge, it cannot compensate for factual gaps or curriculum-specific data that were absent from pretraining. RAG addresses this limitation by coupling the generative power of large language models with a dedicated knowledge retrieval component. In educational contexts, RAG enables models to ground their outputs in authoritative sources such as textbooks, course notes, learning management systems (LMS), and institutional repositories, thereby improving factual accuracy, curricular alignment, and explainability.

RAG Architecture and Components

A standard RAG pipeline consists of three interconnected modules:

1. **Retriever** – converts a learner query into an embedding vector and retrieves semantically related documents from a knowledge store using dense or hybrid search techniques (e.g., FAISS, ColBERT, BM25).
2. **Reader / Generator** – conditions the LLM on the retrieved passages and synthesizes a response grounded in that evidence.
3. **Indexer** – periodically encodes and updates the knowledge base to reflect syllabus revisions or new teaching materials.

This modular structure allows continual improvement without retraining the underlying model. Educators can simply update course documents in the knowledge base to refresh the system's knowledge.

Educational Use Cases

By integrating syllabus documents and textbooks into the retrieval layer, RAG systems can ensure that tutoring responses remain consistent with officially approved content. For instance, LearnRAG demonstrated that retrieval from structured course outlines improved factual grounding and reduced hallucinations in adaptive learning systems by more than 20 percent (Shan, 2025: 3).

RAG can align automated grading with institutional rubrics by retrieving exemplar answers, grading criteria, or annotated feedback examples. The generator then uses this evidence to justify grades with transparent rationale, producing explainable assessment artifacts that teachers can audit.

Frameworks such as *EduPlanner* employ multi-agent RAG pipelines where one agent retrieves learning objectives and another synthesizes lesson materials tailored to those objectives (Zhang et al., 2025: 3). This division of labor enables large-scale curriculum mapping while preserving instructional coherence.

Techniques for Enhanced Educational Grounding

Several retrieval enhancements are particularly beneficial for education:

- **Hierarchical Retrieval.** Index both *course-level* and *topic-level* materials to support multi-granular queries—from "Explain recursion" to "Give an example from Unit 3."
- **Citation Grounding.** Append references or URLs of retrieved passages to each response, improving academic traceability and deterring hallucination.
- **Query Expansion.** Reformulate learner questions using pedagogical taxonomies (e.g., Bloom's verbs: *define*, *analyze*, *evaluate*) to retrieve materials at the appropriate cognitive level (Imperial et al., 2024: 6).

- **Dynamic Context Windows.** Allocate retrieval context proportionally to question complexity, optimizing latency for real-time tutoring.
- **Multilingual Retrieval.** Deploy bilingual encoders or translation-aware retrievers to serve linguistically diverse classrooms while maintaining content fidelity.

Integration with Learning Ecosystems

Modern educational infrastructure facilitates seamless RAG deployment. Vector databases such as Milvus or Pinecone can store encoded curriculum content, while connectors allow integration with LMS platforms like Moodle or Canvas. Teachers can upload new materials directly, triggering automatic re-indexing. Through APIs, the same retrieval layer can feed both chat-based tutors and analytic dashboards, ensuring consistent pedagogical grounding across institutional tools.

Challenges and Considerations

Despite its promise, RAG introduces challenges related to **content quality, retrieval bias, and scalability**. Poorly curated or outdated materials can propagate misconceptions, and unbalanced retrieval may over-represent certain topics. Continuous evaluation pipelines that log retrieval coverage and response fidelity are therefore essential. Furthermore, as educational datasets expand, maintaining low-latency retrieval requires efficient vector compression and adaptive caching.

Parameter-Efficient Fine-Tuning (PEFT) for Educational LLMs

Prompting and retrieval ground an LLM's behavior externally, yet some educational uses—such as consistent grading, tone control, or curriculum-specific reasoning—require internal adaptation. **Parameter-Efficient Fine-Tuning (PEFT)** offers this

alignment by updating only a small fraction of model weights (usually < 5 %) while keeping the pretrained backbone frozen (Ke et al., 2025; Afzal et al., 2024). It therefore bridges full fine-tuning and zero-shot prompting in cost, stability, and pedagogical control.

Tablo 1 Summary of major PEFT methods and their educational benefits

Method	**Mechanism**	**Educational Benefit**
LoRA	Adds low-rank trainable matrices inside attention layers.	Domain-specific reasoning (e.g., physics explanations) with minimal compute.
Prefix / Prompt-Tuning	Learns "soft prompts'' prepended to embeddings.	Personalizes tutor persona or classroom language style.
Adapters / BitFit	Inserts or lightly updates small modules.	Modular subject updates (MathAdapter, HistoryAdapter).
QLoRA	Quantized LoRA for memory-limited devices.	Enables on-device fine-tuning in classrooms.

Fine-tuning data should mirror curriculum objectives, age level, and feedback style. Even small, high-quality sets of exam items or annotated answers can teach models rubric awareness. Evaluation must test *pedagogical alignment* (tone, reasoning steps) as well as factual accuracy—often through expert rubrics rather than generic text metrics (Imperial et al., 2024).

PEFT modules are lightweight and portable: institutions can host a single base model and load course-specific adapters as needed. This enables privacy-preserving, sustainable customization compliant with FERPA/GDPR (Dong & Xie, 2024). Combined with RAG, PEFT yields dual-layer alignment—stable pedagogical logic from fine-tuning plus dynamic factual grounding from retrieval.

Deployment Strategies and Infrastructure

Deploying educational language models requires infrastructures that balance **privacy, scalability, latency, and sustainability**. Once customization is complete, deployment determines how effectively models support classrooms and institutional workflows.

On-premise deployments ensure full data control and are preferred in K–12 or research settings requiring privacy and reproducibility. These setups, hosted within university or school data centers, guarantee compliance with regulations such as FERPA and GDPR while supporting integration with internal databases and analytics systems.

Cloud deployments offer elasticity and easier maintenance, providing scalable access for large user bases or multilingual courses. Such configurations enable continuous model updates but require governance over data residency, cost, and third-party API reliance.

A **hybrid architecture** combines both approaches—performing sensitive inference locally while using secure cloud services for retrieval, analytics, or load balancing (Dong & Xie, 2024). This design optimizes both privacy and accessibility, supporting continuous model updates without compromising institutional control.

To ensure reproducibility and scalability, **containerized deployments** are commonly employed. Tools such as *Docker* and *Kubernetes* enable modular orchestration of model components—fine-tuned adapters, retrieval indexes, and tutoring interfaces—within isolated microservices. This approach simplifies updates, improves fault tolerance, and allows institutions to scale individual components independently.

For **resource-limited or offline classrooms**, lightweight and quantized models can be deployed on **edge servers** or local gateways. This supports equitable access and sustainability while enabling real-time inference without internet dependency.

Integration with **Learning Management Systems (LMS)** such as *Moodle* or *Canvas* can be achieved through REST or GraphQL APIs, embedding tutoring and feedback tools directly into existing teaching platforms. Continuous **monitoring pipelines** further track model performance, bias drift, and retrieval quality over time, ensuring both pedagogical reliability and system transparency.

In essence, effective deployment transforms customized models into **operational educational ecosystems**—secure, modular, and accountable—capable of augmenting human teaching through scalable yet ethically governed AI infrastructure.

Privacy, Ethics, and Governance in Educational LLM Deployment

As LLMs enter classrooms, **privacy and governance** become central to responsible deployment. Educational data often include personally identifiable information, performance metrics, or behavioral traces that demand strict compliance with **FERPA**, **GDPR**, and regional data-protection acts. Systems must ensure that no raw student data are transmitted or stored beyond institutional control (Dong & Xie, 2024).

Key safeguards include:

- **Data Minimization:** Retain only essential inputs and anonymize stored transcripts.
- **Local Processing:** Prefer on-premise or edge inference for sensitive age groups.

- **Access Transparency:** Provide audit logs detailing how and when model outputs are generated.
- **Bias and Fairness Auditing:** Periodically test responses for cultural, gender, or linguistic bias using standardized benchmarks.
- **Explainability Tools:** Offer teachers visibility into reasoning chains or retrieval citations to support instructional oversight.

Governance frameworks should combine **technical controls** (secure APIs, encrypted storage) with **institutional policies** defining model versioning, retraining intervals, and ethical review procedures. Transparent documentation—**model cards** describing datasets, limitations, and pedagogical scope—builds trust among educators and learners (Zhao & Wan, 2025).

Case Studies and Framework Comparison

Practical implementations illustrate how customization and deployment strategies converge in real educational ecosystems.

System / Study	Customization Method	Deployment Mode	Educational Function	Reported Outcome
LearnRAG (Shan, 2025)	Retrieval-Augmented Generation	Hybrid cloud	Adaptive tutoring using syllabus documents	+20 % factual accuracy, reduced hallucination
EduPlanner (Zhang et al., 2025)	Multi-agent RAG + Prompt Templates	Cloud	Curriculum design and lesson synthesis	Coherent cross-topic planning
Adapteval (Afzal et al., 2024)	PEFT (LoRA, Adapters)	On-premise	Automated summarization and grading	Higher rubric alignment
Standardize (Imperial et al., 2024)	Prompt + RAG alignment to expert standards	Cloud	QA generation with curriculum citation	Improved explainability and compliance
COGENT (Liu et al., 2025)	Curriculum-oriented prompting framework	Edge / Local	Grade-appropriate content generation	Increased age-specific relevance

Future Directions and Research Outlook

Educational LLMs are evolving beyond text generation toward **multimodal, lifelong learning companions**. Several research frontiers define the next phase:

1. **Multimodal Integration:** Incorporating speech, handwriting, and gesture analysis for richer learner modeling.
2. **Continual and Federated Learning:** Allowing models to update from distributed classrooms without sharing raw data, supporting global yet privacy-preserving improvement.
3. **Pedagogical Explainability:** Embedding reasoning paths aligned with **Bloom's Taxonomy** to reveal cognitive levels of generated content.

4. **Cross-Institutional Interoperability:** Developing open-standard adapters so universities can share educational modules without exposing proprietary data.
5. **Sustainable AI Practices:** Using quantized inference, renewable-energy data centers, and adapter reuse to reduce carbon cost.

Ultimately, the value of educational LLMs will depend not merely on generative power but on **alignment with human learning principles**—transparency, adaptivity, and inclusiveness. When coupled with responsible deployment, these systems can transform instruction into an interactive dialogue between pedagogy and computation.

Kaynakça

Xu, H., Gan, W., Qi, Z., Wu, J., & Yu, P. S. (2024). Large language models for education: A survey. arXiv preprint arXiv:2405.13001.

Zhao, P., & Wan, X. (2025). Technical implementation of large language models in educational scenarios: A case study of DeepSeek. Advances in Management and Intelligent Technologies, 1(3).

Lee, J., Hicke, Y., Yu, R., Brooks, C., & Kizilcec, R. F. (2024). The life cycle of large language models in education: A framework for understanding sources of bias. British Journal of Educational Technology, 55(5), 1982–2002.

Delikoura, I., & Hui, P. (2025). From superficial outputs to superficial learning: Risks of large language models in education. arXiv preprint arXiv:2509.21972.

Beale, R. (2025). Dialogic pedagogy for large language models: Aligning conversational AI with proven theories of learning. arXiv preprint arXiv:2506.19484.

Zhang, X., Zhang, C., Sun, J., Xiao, J., Yang, Y., & Luo, Y. (2025). EduPlanner: LLM-based multi-agent systems for customized and intelligent instructional design. IEEE Transactions on Learning Technologies.

Wang, W., Zhao, Z., & Sun, T. (2023). Customizing large language models for business context: Framework and experiments. arXiv preprint arXiv:2312.10225.

Dong, H., & Xie, S. (2024). Large language models (LLMs): Deployment, tokenomics and sustainability. arXiv preprint arXiv:2405.17147.

Shan, R. (2025, February). LearnRAG: Implementing retrieval-augmented generation for adaptive learning systems. In 2025 International Conference on Artificial Intelligence in Information and Communication (ICAIIC) (pp. 0224–0229). IEEE.

Zhao, S., Yang, Y., Wang, Z., He, Z., Qiu, L. K., & Qiu, L. (2024). Retrieval augmented generation (RAG) and beyond: A comprehensive survey on how to make your LLMs use external data more wisely. arXiv preprint arXiv:2409.14924.

Afzal, A., Chalumattu, R., Matthes, F., & Mascarell, L. (2024). Adapteval: Evaluating large language models on domain adaptation for text summarization. arXiv preprint arXiv:2407.11591.

Ke, Z., Ming, Y., & Joty, S. (2025). NAACL2025 Tutorial: Adaptation of large language models. arXiv preprint arXiv:2504.03931.

Sonkar, S., Ni, K., Chaudhary, S., & Baraniuk, R. G. (2024). Pedagogical alignment of large language models. arXiv preprint arXiv:2402.05000.

Bhat, S., Nguyen, H. A., Moore, S., Stamper, J. C., Sakr, M., & Nyberg, E. (2022, July). Towards automated generation and evaluation of questions in educational domains. In EDM 2022 Proceedings (pp. 1–8).

Imperial, J. M., Forey, G., & Madabushi, H. T. (2024). Standardize: Aligning language models with expert-defined standards for content generation. arXiv preprint arXiv:2402.12593.

Jiao, Y., Shridhar, K., Cui, P., Zhou, W., & Sachan, M. (2023, June). Automatic educational question generation with difficulty level controls. In International Conference on Artificial Intelligence in Education (pp. 476–488). Cham: Springer Nature Switzerland.

Liu, Z., Yin, S. X., Goh, D. H. L., & Chen, N. F. (2025). COGENT: A curriculum-oriented framework for generating grade-appropriate educational content. arXiv preprint arXiv:2506.09367.

Jain, A., Cui, L., & Chen, S. (2025). Aligning LLMs for the classroom with knowledge-based retrieval—A comparative RAG study. arXiv preprint arXiv:2509.07846.

Pan, F., Zhou, Q., Guo, W., & Yang, H. (2025, April). A survey on retrieval-augmented generation in applications of education and teaching. In 2025 7th International Conference on Computer Science and Technologies in Education (CSTE) (pp. 803–807). IEEE.

Ye, Z. (2025, February). Intelligent tutoring agent with retrieval-augmented generation: A case study of quality management system course. In 2025 5th International Conference on Consumer Electronics and Computer Engineering (ICCECE) (pp. 189–194). IEEE.

LLM-SUPPORTED EDUCATIONAL APPLICATIONS: DESIGN, INTEGRATION, AND EVALUATION

SİNEM MİZANALI[1]

Introduction

Large Language Model (LLM)-based artificial intelligence tools have become a central focus for schools, teachers, and virtually every discipline related to learning. Accordingly, LLM-based artificial intelligence tools have assumed a central position in research and application areas (Giannakos et al., 2025). As the theoretical foundations of LLMs (see Chapter 1) and customization strategies (see Chapter 2) mature, the integration of these technologies into educational ecosystems has the potential to redefine teaching and learning processes.

This section addresses how LLM-based applications are designed, integrated, and evaluated in education. Within this scope, the section has three main objectives: conceptualizing examples of LLM-based applications used in education, outlining the technical and user-centric design principles of these applications, and

[1] Research Assistant, Istanbul Gedik University, Computer Engineering, Orcid: 0000-0002-3002-3057

presenting practical insights through performance and quality evaluation methods and case studies. In line with these objectives, the areas of application for LLM-based applications in education (quiz generation, essay evaluation, language learning support, coding assistants, etc.) will first be addressed, followed by blueprints for the architectural design of these applications. Subsequently, the user experience and interface dimension will be evaluated. The section will conclude with selected case studies and the results obtained from these studies.

Use Cases of LLMs in Education

The integration of artificial intelligence into educational environments has given rise to new approaches that transform the practices of both students and educators. Among the various technologies driving this transformation, Large Language Models (LLMs) are powerful tools that influence learning and teaching (Biancini, Ferrato, & Limongelli, 2025). LLMs' abilities to generate text, classify, summarize, translate, and make logical inferences make them usable in many stages of the learning process (Kasneci et al., 2023). Giannakos et al. (2025) emphasize that generative AI tools, such as LLMs, offer new opportunities for automatic content generation, formative feedback, and assessment in educational contexts; however, the pedagogical robustness of these tools has not yet been sufficiently examined. Studies on the use of LLM in education show that these systems can function not only as tools for transferring knowledge but also as tools that provide cognitive support and generate feedback. (Kasneci et al., 2023).

LLMs offer advantages in automating repetitive tasks and providing feedback when human resources are insufficient. In this context, LLM-based educational applications can be categorized into two main groups:

- **Student-Focused Applications:** These applications can be considered tools that directly support learning processes.
- **Teacher and institution-focused applications:** These can be considered tools that assist in planning, evaluating, and managing learning processes.

Student-Focused Applications

From the student perspective, LLMs can be used in various functions to support the learning process. Student-centered applications related to the use of LLMs in education are summarized below by reviewing the literature under the following subheadings.

Automatic Quiz Generation

In the field of education, Large Language Models are increasingly being used in automated question generation to simplify time-consuming assessment and evaluation processes for teachers (Biancini et al., 2025). Quiz Generation applications can generate questions tailored to the subject matter, difficulty level, and target learning outcomes based on a specific text. However, despite its potential to simplify teachers' assessment processes, automated question generation still requires human oversight for content accuracy, semantic coherence, and question variety (Azzi, Erdős, Németh, Varadarajan, & Afrifa, 2025). Therefore, LLM-based quiz tools are typically designed with the human-in-the-loop principle. In this context, Biancini et al. (2025) compared multiple-choice question generation using GPT-3.5, LlaMA 2, and Mistral and demonstrated that GPT 3.5 was significantly superior in terms of clarity and alignment with the source text under the knowledge injection strategy. The findings support the practical value of human-in-the-loop designs in measurement and assessment. As a more specific example, in a study by Dijkstra, Genc, Kayal, and Kamps (2022) GPT-3 was used to generate multiple-choice questions and

answers for reading comprehension tasks. The researchers argued that automated exam creation not only reduces the burden of manual exam design for teachers but, above all, provides a useful tool that makes it easier for students to practice and test their knowledge while learning from textbooks and preparing for exams. Such applications encourage students to practice regularly while also enabling teachers to produce more systematic, measurable, and learning-goal-aligned assessment materials.

Automated Essay Scoring and Written Feedback

Traditional forms of measurement and assessment processes in education are quite time-consuming and labor-intensive. Although providing feedback is critically important in education, evaluating open-ended responses and creating personalized feedback for each student are quite difficult and time-consuming processes (Xavier et al., 2025). Large Language Models have become the center of academic discussions in recent years with their potential to transform these processes. Recent research on this topic shows that LLMs can not only automate grading but also generate meaningful and personalized feedback that supports students' cognitive development (Fagbohun, Iduwe, Abdullahi, Ifaturoti, & Nwanna, 2024; Maity & Deroy, 2024).

LLM-based automatic essay scoring (AES) systems are one of the earliest and most widespread applications in education. These systems can holistically analyze features such as text length, consistency, vocabulary diversity, and linguistic accuracy. Fagbohun et al. (2024) emphasized in their study that models can evaluate student responses across a wide range, from short answers to long essays, in terms of content, structure, grammar, and conceptual accuracy, while also providing detailed, explanatory, and constructive feedback. Thanks to this approach, the assessment process moves beyond the concept of grading and becomes a

feedback loop that guides learning. Although studies show that LLM models perform assessments at a level of reliability similar to that of real human evaluators, the importance of human oversight should not be overlooked. Even if LLMs score and/or provide feedback with high accuracy, issues such as model bias, data privacy, and ethical accountability necessitate human oversight. In this regard, Fagbohun et al. (2024) suggest that LLMs should be designed as tools that support teachers' decision-making processes rather than replace them.

In addition to generating questions, LLMs show significant potential in response evaluation (Fagbohun et al., 2024). The ability to accurately evaluate student responses and provide feedback is a critical component of the educational process. Traditionally, this task has been performed by educators who must carefully evaluate the content and context of each response (Balfour, 2013). Consequently, LLMs have the potential to revolutionize education through response evaluation and written feedback. With their ability to understand, generate, and evaluate text, these models can help guide students on their educational journey through their capabilities in automatic assessment and constructive feedback (Maity & Deroy, 2024).

Language Learning and Personal Learning Assistants

LLM-based chat assistants offer revolutionary innovations in helping students learn a second language. A study by Kasneci et al. (2023) shows that ChatGPT and similar models interact with students in natural language, instantly correcting their mistakes, providing example sentences, and offering comments aimed at increasing learning motivation. In this context, students can practice using LLM models as tutors when learning a new language. In addition, LLMs can provide content according to the student's level and create personalized learning paths. Furthermore, they act as

partners in continuous interaction, complementing the traditional student-teacher interaction in language learning.

LLMs' ability to provide real-time feedback in areas such as pronunciation and speaking practice creates an accessible support mechanism for students with different profiles (El Shazly, 2021). Findings in the literature indicate that these approaches have a positive effect on language development (Kasneci et al., 2023).

However, as mentioned in the previous subheadings, issues such as pedagogical accuracy and cultural bias must be carefully monitored at this point. Therefore, teacher guidance and the human-in-the-loop principle are essential in the use of LLMs in the context of language learning.

Coding, creating/generating code columns

Large Language Models offer promising solutions for automatic code generation by leveraging extensive training on various code bases. Unlike traditional methods, LLMs can generate code in a wide variety of programming languages with minimal user effort (Eagal, Stolee, & Ore, 2025). Coding assistants can review the code written by students, explain error messages, offer solutions, and suggest alternative coding approaches. This allows students to receive immediate and contextual support in both debugging and design/refactoring processes.

Studies in the literature indicate that LLM-powered coding assistants reduce students' problem-solving time but may sometimes weaken the student's critical thinking process by creating overconfidence (Akçapınar & Sidan, 2024; Groothuijsen, Beemt, Remmers, & Meeuwen, 2024). Therefore, it is important that these tools are used under teacher guidance. Furthermore, while it is technically possible to design a website or mobile application using only LLMs, principles such as verification, security, and

explainability must be observed throughout the process in such applications.

Another study in computer education was contributed to the literature by MacNeil et al. (2022). In this study, GPT-3 was used to generate code explanations. This study successfully demonstrated GPT-3's potential to support learning by explaining the aspects of a specific piece of code.

In this context, when combined with appropriate pedagogical design and supervision, LLM-based coding assistants have the potential to enhance students' debugging skills, conceptual understanding, and productivity. However, they should be supported by instructional safeguards to mitigate the risks of dependency and overconfidence.

Teacher and institution-focused applications

Large language models have become an indispensable opportunity to enhance learning and teaching experiences for individuals at all educational levels, including elementary school, middle school, high school, and university, due to the diverse applications they offer. They possess features that can be beneficial to individuals at every level of education. In addition, they offer opportunities for students with special needs. In line with the principle of equal opportunity in education, every individual has the right to education regardless of religion, language, or race. In this regard, LLMs provide benefits to individuals when used for the right purposes.

Curriculum and Course Design Support

LLMs can be used to create lesson plans, learning outcomes, and activity suggestions for teachers. For example, a teacher can quickly obtain a customized lesson flow tailored to their goals by

using an English prompt such as “generate a weekly lesson plan on Newton's Laws for 10th grade students.”

Furthermore, these systems can assess learning outcomes based on students' previous performance data and adapt learning goals. In addition, they can generate improvement plans related to learning outcomes. Such tools provide information support, especially for teachers new to the field. However, the suggestions obtained should be reviewed by the teacher to ensure they are used safely in the classroom context.

Feedback, Measurement, and Analytical Reporting

LLMs can automatically analyze student performance as powerful text classifiers and provide teachers with summary and actionable reports. For example, within a writing course, students' texts can be summarized by the LLM and classified according to error types. Another alternative is to classify them according to subject matter. This allows teachers to quickly analyze which topics need improvement at the individual or class level and take action accordingly. Such systems serve as a decision support mechanism that accelerates and enriches decision-making processes rather than replacing teachers. Having teachers review reports for pedagogical appropriateness and contextual accuracy enhances the quality of feedback and strengthens the consistency of classroom practices.

Hybrid and system-level applications

The most effective scenarios in education frequently manifest in hybrid systems where student- and teacher-focused functions operate in an integrated manner. For instance, in a language learning platform, a large language model (LLM) assesses student text via the student module and offers immediate feedback, while the analytics module supplies the teacher with class-wide error statistics. Such systems can be designed utilizing a dual-agent

architecture approach: the primary agent engages with the student, whereas the secondary agent informs the teacher or generates systemic reports.

In pilot systems implemented at select universities (Kasneci et al., 2023) large language models (LLMs) have been integrated into various educational contexts, including student advising, homework oversight, and in-course assistance. The findings suggest an enhancement in student motivation; however, educators remain apprehensive regarding the reliability of the systems and the transparency of assessments.

Applications based on LLMs present considerable opportunities within the educational landscape, particularly concerning personalized learning, equitable access, and scalability. Nonetheless, challenges such as hallucination (the generation of incorrect information), bias, data privacy, and pedagogical appropriateness represent substantial limitations. The management of these risks will be addressed comprehensively in subsequent chapters of the book (see Chapter 4).

The effective implementation of these use cases depends on robust technical architecture. In this context, the next section will examine in detail the architectural design, technical integration, and performance evaluation strategies for these applications.

Architectural Blueprints

This section details how the applications described in the use cases section are built, which technical components they combine, and how LLM architectures can be integrated into education systems.

The most critical step in developing applications based on large language models in education is to correctly design the technical architecture of these systems. In this context, the success

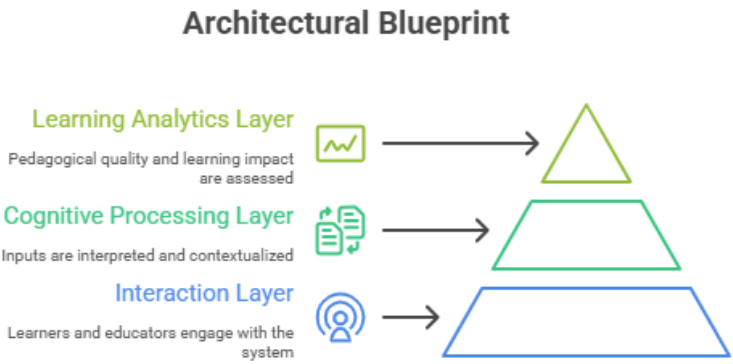
of the application depends not only on the accuracy of the model but also on the system's design, integration, reliability, and measurability qualities. An LLM-based educational application should be built on an architectural structure where the data flow is defined and user interaction is modeled to evaluate outputs.

The design of LLM-supported education systems requires an architectural framework based on solid pedagogical foundations. In this regard, the architectural plan must be aligned not only with technological capabilities but also with educational objectives. Thus, automation supports meaningful learning experiences rather than replacing them. Accordingly, the architectural plan can be structured in three layers.

- Interaction Layer (User Interface); the layer where students and educators interact with the system.
- Cognitive Processing Layer (LLM Core + Middleware); where inputs are interpreted, transformed, and contextualized.
- Learning Analytics and Assessment Layer; the layer where the pedagogical quality, reliability, and learning impact of outputs are evaluated.

Figure 1 shows the Architectural Blueprint.

Figure 1. Architectural Blueprint Structure



The *Interaction Layer* represents the user-centric interface that facilitates educational dialogue between humans and LLM. The core design principles can be listed as multi-modal interaction, adaptive interfaces, and transparency. Multi-modal interaction refers to text, voice, or visual inputs to support different learners. React or other modern user interface frameworks can be used to create accessible and inclusive user experiences.

At the heart of the architecture lies the *Cognitive Processing Layer*, which functions as the bridge between the user interface and the model's reasoning capability. It typically includes:

- Prompt Orchestration Engine: dynamically constructs prompts using contextual and pedagogical metadata (e.g., learning goals, student history).
- LLM Interface: a standardized API gateway (e.g., OpenAI, Anthropic, or open-source models) that handles requests, responses, and model selection.

- Knowledge Integration Subsystem: optional retrieval-augmented generation (RAG) modules or educational databases to provide grounded, domain-specific responses.
- Session Memory: to maintain pedagogical coherence over multi-turn interactions.

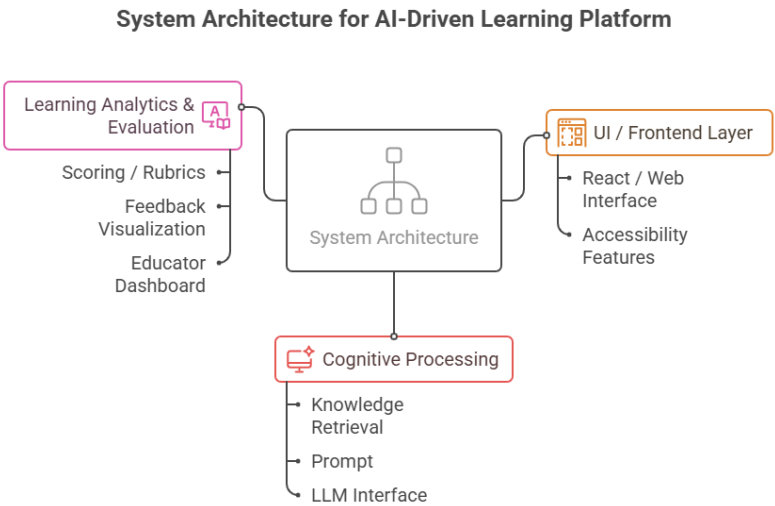
This layer embodies the system's cognitive logic, aligning model reasoning with instructional design principles.

The *Learning Analytics Layer* monitors the model's outputs, user engagement, and learning effectiveness. It may include:

- Automated feedback engines that map model outputs to rubric-based assessment frameworks.
- Human-in-the-loop evaluation workflows that enable educators to verify or adjust model-generated feedback.
- Learning analytics dashboards for visualizing progress, misconceptions, or engagement patterns. The goal is to ensure that AI outputs are pedagogically valid, not just technically fluent.

Figure 2 shows the system architecture for the AI-supported learning platform.

Figure 2. Systems Architecture for AI Driven Learning Platform



User Interface and User Experience Considerations

Providing clear and reliable feedback in systems used by students and developing data-driven dashboards for teachers are of critical importance. Educational applications built on Large Language Models require not only powerful artificial intelligence models but also interaction designs that put the user at the center. An effective user experience directly determines the success of the model. In a pedagogical context, a good LLM application is a system that not only produces accurate information but also facilitates student understanding and gives teachers a sense of control (Holmes, Bialik, & Fadel, 2019).

The study by Giannakos et al. (2025) reveals that LLM-supported educational tools require the adoption of a human-centered design in which teachers and students have initiative, control, and interpretive control over outputs generated by artificial

intelligence. This section will examine user experience in LLM-based education systems from both a student-centered and teacher-centered perspective.

Student-Centered Interaction Design

In education, the interaction between LLMs and students goes beyond the traditional "question-answer" format. The goal here is to transform the student from merely a recipient of information into an active co-learner. Therefore, the interface and interaction design should reduce the student's cognitive load, facilitate guidance, and increase motivation.

The most common form of student interaction is dialogue-based systems. An LLM-based educational chatbot not only corrects the student's mistakes but also guides the thinking process. For example:

Student: "I goed to the park yesterday."

Assistant: "You almost got it! The correct past tense of 'go' is 'went'. So the sentence should be "I went to the park yesterday."

This type of feedback provides positive reinforcement while correcting errors (scaffolded feedback). Research (Kasneci et al., 2023) shows that this type of interaction increases students' confidence and retention of information, especially in language learning.

For students with a visual learning style, graphical explanations or diagrams can be added alongside text-based outputs. Thanks to multimodal LLMs such as GPT-4V and Gemini, students can see the solution steps to a math problem through shapes rather than text. Such multiple representations support deep comprehension in learning theories. In addition, multimodal chat interfaces can support students with dyslexia or limited reading fluency. In other

words, multimodal LLMs are also highly beneficial for individuals with special needs who have temporary or permanent disabilities.

LLMs can generate responses based on the student's past performance or preferences. For example, the system learns concepts that the student struggled with from previous responses and reinforces them in subsequent tasks. LLM-based systems are evolving from neutral information providers into supportive tools that also consider the student's emotional state. Even in simple examples, systems that detect student reluctance (“Don't worry, let's try this step by step”) increase cognitive resilience.

Teacher-Centered Interaction Design

User experience in education is not limited to students alone. Control panels and analytics screens designed for teachers are elements that determine the pedagogical value of the system. Teachers need transparency, traceability, and customization features to trust LLM-based applications. LLM-based education systems are evaluated not only for technical accuracy but also for ethical interaction principles. One of the fundamental components of user experience is the concept of trust. If a student or teacher can partially understand how the system works, they can more easily trust it.

Providing the rationale behind the model's response increases user confidence. For example, if a quiz assessment tool clearly explains why an answer is incorrect, the model's explanatory nature makes the response instructive for the student. Similarly, the response will be traceable for the teacher.

Some systems increase transparency by showing which information source the answer was generated from (source attribution). However, teacher control is still important at this stage. Retrieval Augmented Generation (RAG)-based structures technically support this type of traceability (Gao et al., 2024). LLM-

based interfaces used in education must be grounded in pedagogical theories, unlike traditional software systems. In this context, the use of large language models (LLMs) in education requires a human-centered approach that goes beyond traditional user experience principles. The following principles are recommended in this context (Shneiderman, 2020);

- **Cognitive Load Management:** The interface should be free of unnecessary information and understandable in a single step. It should be designed to be instructive and reduce cognitive load. This will increase the student's attention and focus.
- **Scaffolded Interaction:** The model guides the student to the goal step by step, not all at once. This allows the student to participate in the process.
- **Dialogic Feedback:** Explanatory feedback is preferred over one-sentence responses.
- **Transparency by Design:** The model's resources, limitations, and confidence level should be made visible.
- **Inclusivity and Accessibility:** Interfaces should be accessible to different age, language, and cultural groups. In line with inclusive education, everyone should benefit from equal learning opportunities.

These principles integrate Shneiderman (2020) foundations of “human responsibility, high reliability, and enhanced creativity” into educational design. LLM-based systems in education thus become not only technically functional but also ethical, explainable, and pedagogically meaningful. Shneiderman (2020) emphasizes that human control and high automation are not opposing but complementary design goals. In educational user experience, this principle means leveraging automation for efficiency while preserving teacher expertise.

Performance and Quality Evaluation

Evaluation is central to the sustainability and reliability of LLM-based educational applications. A system must not only function correctly, but also demonstrate compliance with learning outcomes, pedagogical values, and ethical standards. Therefore, the performance of LLMs in education should be assessed not only by technical criteria, but also in terms of pedagogical effectiveness, user trust, and contribution to the learning experience (Giannakos et al., 2025; Kasneci et al., 2023).

Language generation quality, consistency, and error rate are traditional technical indicators for LLM-based systems. Metrics such as BLEU, ROUGE, and METEOR can be used in generation tasks; however, in an educational context, these metrics must be supported by pedagogical meaningfulness. For example, a quiz generation tool receiving a high BLEU score does not guarantee the conceptual accuracy of the questions or their suitability for the student's cognitive level. Therefore, hybrid evaluation frameworks have been proposed in recent years. First, the model's outputs can be scored both by automatic metrics and by expert teachers in the field. Another alternative is to analyze model errors in categories such as “factual errors,” “instruction violations,” and “pedagogical mismatch.” Finally, the same input can be tested with different models and consistency analysis can be performed (Kasneci et al., 2023).

Another element in educational applications could be considered hallucination. Misinformation carries the risk of direct learning loss and the formation of incorrect concepts in education. A three-tiered safeguard can be proposed for this risk. To prevent this, the response is based on the relevant information source, and if the reference is missing, a warning is returned. Another alternative is for the model to indicate its own confidence level. For example, it could

return a response such as “I am 70% confident about this answer.” Finally, an independent verification model can check the LLM output. These methods affect not only technical accuracy but also the student's confidence in the information. Justified explanations of responses support the student's critical thinking skills.

Performance should also be evaluated in terms of time, cost, and system stability. In educational environments, low latency, high uptime, and scalable architecture directly impact the user experience. In RAG (Retrieval-Augmented Generation)-based systems, techniques such as cache optimization and embedding-based query matching increase this efficiency (Gao et al., 2024). However, more importantly, these metrics must be balanced with pedagogical value: a meaningful explanation produced in a slightly longer time should be preferred over a fast but superficial response (Holmes et al., 2019).

Another dimension that is as important as technical accuracy is the pedagogical suitability of model outputs and their contribution to the learning process. The concept of pedagogical alignment is used to assess whether LLM-based tools are compatible with learning objectives (Giannakos et al., 2025). An application should not only provide accurate information but also support the learner's learning objectives. The most direct way to measure pedagogical quality is to track changes in student learning outcomes. This is evaluated using both quantitative (test scores, number of tasks completed) and qualitative (student self-reflection, perceived benefit) data.

The success of LLMs can be measured by their impact on student engagement and motivation. For example, Kasneci et al. (2023) showed that ChatGPT can increase children's curiosity and questioning behavior. In addition, the lack of sufficient digital and artificial intelligence literacy among teachers and students using

LLMs is considered one of the challenges in LLM usage. It is evident that trust and transparency mechanisms are necessary for user experience (Kasneci et al., 2023).

Although Large Language Models offer opportunities in learning design processes in terms of speed and creativity, automatic content, feedback, and assessment generation, and supporting students' self-regulation skills; they also bring many challenges such as model biases, ethical issues, data privacy deficiencies, the blurring of the definition of the human role, and the risk of rapid integration being pedagogically inadequate (Giannakos et al., 2025).

The sustainable integration of LLMs into education must be balanced with responsibility. While these systems can personalize learning and reduce teachers' workload, their uncritical adoption carries the risk of undermining ethical transparency. This tension between opportunity and oversight will form the basis of the ethical discussions addressed in Chapter 4.

Case Studies

Under this heading, both success indicators and limitations will be discussed based on LLM applications tested in different learning environments.

In this context, a study conducted by El Shazly (2021) examined ChatGPT's contribution to students' language learning process. In the study, students received written and verbal feedback while practicing conversation with ChatGPT. The model also instantly suggested corrections for spelling mistakes, grammar errors, and stylistic improvements. Among the results obtained in this study, it was observed that students' language anxiety decreased by 40% and that students showed more courage while practicing speaking. However, it was also seen that most students began to trust

the accuracy of the grammar corrections provided by the model, which led to a weakening of their critical thinking skills.

The study conducted by Akçapınar and Sidan (2024) examined the effect of an AI programming assistant on students' exam scores and their tendency to accept misinformation generated by AI. The authors developed a customized AI programming assistant using a GPT-based LLM. In the study, which used experimental design, students were asked to take a programming exam once with AI assistance and once without AI assistance. The results of the exam taken with AI assistance showed a significant increase. However, when examining the student-AI interaction logs for a specific question, it was found that the AI generated incorrect answers for that question for 36 students, and 33 of the 36 students who received the incorrect answer provided the wrong response to that question. Despite the obvious error in the AI-generated answer, 22 students directly copied and pasted the AI response. Only 3 students recognized the incorrect answer generated by the AI and answered the question correctly. The fact that a significant portion of the students accepted the incorrect answer provided by the AI without questioning clearly demonstrates how carefully AI tools must be used in learning environments.

Xavier et al. (2025) present a controlled experiment comparing traditional teacher feedback with LLM-supported feedback via a platform among 60 middle school students in Brazil. The results of the experiment showed no significant difference in students' perceptions of feedback quality. In other words, the vast majority of students could not distinguish feedback generated by LLM from teacher feedback. The study found that LLM assistance produced longer feedback messages without significantly increasing grading time.

The study by Biancini et al. (2025) experimentally examined the role of LLMs in assessment processes. The research comparatively evaluated the performance of GPT-3.5, LLaMA 2, and Mistral models in generating multiple-choice questions. One of the most significant contributions of the study is the knowledge injection approach, whereby test data is transferred to the model externally, independent of the model's internal knowledge. This enables teachers to have full control over the source text. This approach emphasizes the necessity of developing LLM-based systems in educational settings according to the "human in the loop" design principle. The experimental findings were obtained from 21 educators. According to the results of the study, GPT-3.5 showed statistically significant superior performance compared to the LLaMA 2 and Mistral models in all criteria. In particular, the difference was significant in the criteria of clarity and alignment with the source text. This study demonstrates that LLMs are not only linguistic production tools but can also be effective partners in designing assessment and evaluation in education. However, researchers emphasize that human expertise should not be completely eliminated. Although the questions generated by LLMs have high accuracy rates, they still require expert oversight in terms of content diversity, cognitive level (e.g., Bloom's Taxonomy), and contextual appropriateness. In this context, the fundamental approach in designing LLM-supported educational applications should be to keep the teacher at the center of the process, using the model's automation capacity as an auxiliary design element (Biancini et al., 2025). Additionally, researchers emphasize that future studies should expand LLM-based question generation not only in terms of linguistic accuracy but also in terms of cognitive level classification based on learning objectives and personalized assessment designs. Models to be developed could enable the measurement of learning outcomes at different cognitive levels

among students. This approach paves the way for LLM-supported educational applications to be integrated not only into content production but also into learning analytics and adaptive assessment processes. Thus, LLM-based assessment systems, integrated with teacher-centered design principles, lay the groundwork for an ethical, reliable, and pedagogically meaningful AI integration (UNESCO, 2023).

References

Anderson, L. W., & Krathwohl, D. R. (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. New York: Longman.

Balfour, S. P. (2013). Assessing writing in moocs: Automated essay scoring and calibrated peer review. Research & Practice in Assessment, 40-48.

Biancini, G., Ferrato, A., & Limongelli, C. (2025). Multiple-Choice Question Generation Using Large Language Models: Methodology and Educator Insights. Paper presented at the UMAP Adjunct '24: Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization, New York, NY, USA.

Fagbohun, O., Iduwe, N., Abdullahi, M., Ifaturoti, A., & Nwanna, O. (2024). Beyond traditional assessment: Exploring the impact of large language models on grading practices. Journal of Artifical Intelligence and Machine Learning & Data Science, 2(1), 1-8. doi:doi.org/10.51219/JAIMLD/oluwole-fagbohun/19

Maity, S., & Deroy, A. (2024). The Future of Learning in the Age of Generative AI: Automated Question Generation and Assessment with Large Language Models. In.

UNESCO. (2023). Guidance for generative AI in education and research. Paris, France.

Akçapınar, G., & Sidan, E. (2024). AI chatbots in programming education: guiding success or encouraging plagiarism. Discover Artificial Intelligence, 4(87). doi:https://doi.org/10.1007/s44163-024-00203-7

Azzi, A., Erdős, F., Németh, R., Varadarajan, V., & Afrifa, S. (2025). Comparative analysis of NLP-driven MCQ generators from

text sources. Computers and Education: Artificial Intelligence. doi:10.1016/j.caeai.2025.100440

Balfour, S. P. (2013). Assessing writing in moocs: Automated essay scoring and calibrated peer review. Research & Practice in Assessment, 40-48.

Biancini, G., Ferrato, A., & Limongelli, C. (2025). Multiple-Choice Question Generation Using Large Language Models: Methodology and Educator Insights. Paper presented at the UMAP Adjunct '24: Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization, New York, NY, USA.

Dijkstra, R., Genc, Z., Kayal, S., & Kamps, J. (2022). Reading Comprehension Quiz Generation using Generative Pre-trained Transformers. Paper presented at the 4th International Workshop on Intelligent Textbooks.

Eagal, A., Stolee, K. T., & Ore, J.-P. (2025). Analyzing the dependability of Large Language Models for code clone generation. Journal of Systems and Software. doi:10.1016/j.jss.2025.112548

El Shazly, R. (2021). Effects of artificial intelligence on English speaking anxiety and speaking performance: A case study. Expert Systems. doi: https://doi.org/10.1111/exsy.12667

Fagbohun, O., Iduwe, N., Abdullahi, M., Ifaturoti, A., & Nwanna, O. (2024). Beyond traditional assessment: Exploring the impact of large language models on grading practices. Journal of Artifical Intelligence and Machine Learning & Data Science, 2(1), 1-8. doi:doi.org/10.51219/JAIMLD/oluwole-fagbohun/19

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., . . . Wang, H. (2024). Retrieval-Augmented Generation for Large Language Models: A Survey. Arxiv. doi:https://doi.org/10.48550/arXiv.2312.10997

Giannakos, M., Azevedo, R., Brusilovsky, P., Cukurova, M., Dimitriadis, Y., Hernandez-Leo, D., . . . Rienties, B. (2025). The promise and challenges of generative AI in education. Behaviour & Information Technology, 44(11), 2518-2544. doi:10.1080/0144929X.2024.2394886

Groothuijsen, S., Beemt, A. v. d., Remmers, J. C., & Meeuwen, L. W. v. (2024). AI chatbots in programming education: Students' use in a scientific computing course and consequences for learning. Computers and Education: Artificial Intelligence, 7. doi:https://doi.org/10.1016/j.caeai.2024.100290

Holmes, W., Bialik, M., & Fadel, C. (2019). Artificial Intelligence in Education. Promise and Implications for Teaching and Learning: Center for Curriculum Redesign.

Kasneci, E., Sessler, K., SKüchemann, t., Bannert, M., Dementieva, D., Fischer, F., . . . Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103.

MacNeil, S., Tran, A., Mogil, D., Bernstein, S., Ross, E., & Huang, Z. (2022). Generating Diverse Code Explanations using the GPT-3 Large Language Model. Paper presented at the ICER '22: Proceedings of the 2022 ACM Conference on International Computing Education Research Switzerland.

Maity, S., & Deroy, A. (2024). The Future of Learning in the Age of Generative AI: Automated Question Generation and Assessment with Large Language Models. In.

Shneiderman, B. (2020). Human-Centered Artificial Intelligence: Three Fresh Ideas. AIS Transactions on Human-Computer Interaction, 12(3), 109-124. doi:10.17705/1thci.00131

UNESCO. (2023). Guidance for generative AI in education and research. Paris, France.

Xavier, C., Xavier C. da Costa, Torrezão, N., Valdo, A. K., Alves, G., Rodrigues, L., . . . Mello, R. F. (2025). Human Teacher vs. LLM-Generated Feedback in Secondary Education: A Comparative Study on Student Perceptions. Paper presented at the 20th European Conference on Technology Enhanced Learning, ECTEL 2025.

CHAPTER 4: ETHICAL CHALLENGES, HALLUCINATION RISKS, AND RESPONSIBLE AI IN EDUCATION

BAŞAK BULUZ KÖMEÇOĞLU[1]

Introduction

Artificial Intelligence (AI) has emerged as a constitutive process in the present educational culture, and not merely as a mere form of technology, but as an epistemic and moral actor, which increasingly steers teaching, assessment and policy-making in education (Jose et al., 2025). Large language models (LLMs) and other generative AI systems have reconfigured the way knowledge is constructed, represented, and verified in a pedagogical setting. In this process of transformation, education is confronted with two tasks - how to leverage AI's personalization and efficiency, and how to protect basic academic and ethical imperatives such as fairness, integrity, and accountability (Sharma, et al., 2025; Memarian & Doleck, 2023).

As recent scholarship points out, the inclusion of AI at the centre of education is fraught with both opportunities and structural

[1] Asst. Prof. Dr., Istanbul Gedik University, Computer Engineering, Orcid: 0000-0001-9937-1036

vulnerabilities. First, adaptive learning systems and automated feedback tools offer inclusivity, prompt assessment, and data-driven teaching. On the other hand, they introduce new mechanisms of asymmetrical human judgment versus algorithmic inference, generating moral questions about transparency, bias, and epistemic dependence (Klimova & Henriksen, 2024). The normative codes that historically shaped educational practice—autonomy of the learner, validity of the measurements used and the status of the educator as moral agent— now face opaque algorithmic systems whose workings are neither fully explainable nor contestable (Bittle et al., 2025).

Among the new anxieties, hallucination — the construction of linguistically consistent but factually inconsistent output by generative AI — is an especially critical epistemic concern. In these educational contexts of factual reliability and conceptual precision which are foundational to learning, such artificial content threatens the integrity of knowledge and evaluation procedures (Elsayed, 2024) Hallucination is an example of 'systemic epistemic instability,' which creates a gap between fluent language production and reliability based on facts (Cao, 2025). This phenomenon, emerging as a co-author of the artificial intelligence education discourse, provides a perspective for some critical discussions on the reconstruction of trust, authority, and verification. This phenomenon provides a perspective for some critical discussion of rebuilding trust, authority, and verification when AI emerges as a co-author of educational discourse.

In addition, the pedagogical implications are not just about content authority, but also about academic truth and human agency. Generative AI platforms challenge both help and authorship as well as the accepted views of originality, effort and evaluative fairness (Mouta et al, 2025). AI's ability to replicate reasoning and produce assessments threatens to displace the teacher's evaluative role,

increase dependence and reduce students' critical independence (Pratiwi et al.,2025). Ethical AI in education requires "human-in-the-loop" oversight to ensure reflective judgment remains embedded in pedagogy as an end in itself (López-Meneses et al., 2025; Fajardo-Ramos et al., 2025)

These developments have provided a stimulus for the development of Responsible AI in Education (AIEd) frameworks, which aim to harmonise technological innovation with ethical governance. Research published today highlights the need for multi-stakeholder accountability: developers to provide fairness and explainability, institutions to embed governance systems, educators to ensure that educators choose the appropriate use, and learners to acquire AI literacy skills (Fu & Weng, 2024; Zhuang et al., 2025). The aim is not just risk mitigation, but to build ethically sustainable learning environments in which automation does not replace human judgment, but rather serves human judgment.

Thus, it is the purpose of this chapter to achieve three objectives. It firstly describes the pedagogical reweavings and tensions introduced by AI in education through the use of the artificial intelligence with respect to the pedagogical transformation and conflicts. Second, it studies four core ethical dilemmas -data ethics, bias, transparency and accountability -which it builds on recent empirical data. Third, it discusses hallucination as an epistemic phenomenon and conceptualizes governance under Responsible AI principles. The chapter concludes with policy-oriented reflections on balancing innovation with responsibility and on nurturing a human-centred AI pedagogy. By locating ethical reflection within the realm of grounded educational practice, the chapter contends that responsible AI is not a technical add-on, but a moral and epistemological precondition for the sustainability of truth, justice, and trust in a technology-mediated education.

The Rise of AI in Education: Opportunities and Tensions

Evolution of AI Tools in Education

The integration of AI into educational environments has evolved from an experimental phase to a systemic one. Early machine learning applications focused on adaptive testing and automatic grading. However, advancements in natural language processing have enabled large-scale interactions with intelligent tutoring systems and generative assistants (Chandrakant, 2025). These developments represent a paradigm shift from rule-based automation to probabilistic reasoning. AI now models learning trajectories, predicts potential performance, and co-produces instructional materials.

AI-powered tools, such as writing support systems, conversational tutors, and personalized analytics dashboards, are increasingly integrated into both formal and informal learning settings (Zhai et al., 2021; Chu et al., 2025) . Applications of LLM-based platforms synthesize course content, offer tailored feedback, and facilitate peer collaboration (Zhang et al, 2025; Guo et al., 2024; Naie et al., 2024; Shahzad et al., 2025; Abu-Rasheed et al., 2024). Concurrently, these AI technologies are becoming integral to administrative decision-making, including admissions screening, early warning systems, and curriculum optimization, extending their influence beyond pedagogy to the governance of educational organizations (Zhao et al., 2025; Hu et al., 2024; Chu et al., 2025). This growth has blurred the line between human and computational authority in the learning process. While traditional educational technologies acted as mediating resources, AI systems now assume interpretative and evaluative roles previously held by educators. The transition from "supporting cognition" to "substituting cognition" necessitates a rethinking of the epistemic division of labor in learning (Grinschgl & Neubauer, 2022; Zhuang et al., 2025)

Pedagogical Transformations

The pedagogical implications of the adoption of AI systems are not just in terms of the efficiency gains. Recent academia also focuses on the way AI systems disrupt epistemological underpinnings of the teaching and learner context by mediating access to, representation and legitimacy of knowledge (Creely & Carabott, 2025; Marshall et al., 2024). Adaptive learning environments change the pace and sequencing of teaching, and generative AI tools also change the way students think about authorship, originality, and reflective thinking (Li et al.,2024; Martin et al.,2020; Baidoo-Anu & Ansah, 2023). Teachers are increasingly seen as curators of AI-generated information, helping their students process information with interpretive evaluation, not direct instruction. This shift accords with post-human pedagogy: learning emerges through interaction between human and non-human agents (Katsenou et al., 2025). And yet, it also spawns tensions around control, reliability, and the decline in professional competence. An expanding body of academic evidence indicates that AI-enhanced pedagogy has the capacity to promote engagement as well as formative feedback, but also to create a passive reliance on algorithm-mediated recommendations (Pitts et al., 2025). In other words, AI's transformative pedagogy is dialectical: while it increases human capability, it also restricts it within algorithmic boundaries.

Ethical Tensions

AI's advent as a tool to educate holds pedagogical potential, but it also breeds enduring ethical friction in education. Three issues—epistemic dependence, assessment fairness, and human agency—remain especially pertinent in present scholarly research.

1.Epistemic Dependence

Epistemic dependence is the extent to which learners (and educators) look to AI systems as authoritative sources of knowledge (Kahl, 2025). When LLM-driven systems produce fluent yet unverifiable outputs, it may be easier to embrace these outputs with little critical examination of what is generated.

This dependency undermines students' epistemic agency — their ability to assess and justify belief — and has the power to supplant critical reasoning with algorithmic trust. Empirical evidence indicates that extended exposure to AI explanations leads to lower metacognitive awareness and higher acceptance of inaccurate claims (Yeh & Siah, 2025). The educational task is, then, to build learning environments in which AI literacy — the capacity to interpret, question, and verify machine outputs rather than uncritical consumption — is engendered (Ng et al., 2024; Daher, 2025; Southworth et al., 2023).

2.Assessment Fairness

While AI-driven assessment systems promise objectivity, they can duplicate hidden biases inherent in their training data. Automated essay scoring, speech recognition, and predictive analytics have been shown to discriminate against students because of linguistic variation, accent, or socioeconomic background (Baker & Hawn, 2022; Jones-Jang et al. (2025) emphasize that perceptions of fairness are crucial to AI-driven grading as a matter of legitimacy: even statistically credible models can damage trust if learners consider decision-making opaque as well as culturally insensitive. Consequently, fairness in assessment must be understood as both an algorithmic and a relational construct—requiring scoring criteria to be transparent and opportunities for human appeal, according to Fu and Weng (2024). Institutional policies should aim to have AI systems that complement, rather than displace, educators' evaluative judgment, striking a balance between efficiency and equity.

3.Human Agency

In the context of artificial intelligence (AI), human agency refers to individuals' capacity to act and make choices independently in accordance with their own beliefs, values, and goals. It encompasses multiple dimensions, including intentionality (acting with purpose), autonomy, adaptability, and responsibility (ethical accountability) (Holmes, 2024). In education, the evolution of AI reflects a shift toward Paradigm Three (AI-empowered, learner-as-leader), in which learners are seen not as passive recipients but as active participants with autonomy to define their own academic goals (Ouyang & Jiao, 2021). AI offers potential to enhance human capabilities by providing personalized learning experiences, increasing efficiency, and supporting decision-making—an augmentation-oriented approach that envisions a hybrid model where technology amplifies human potential (Sethuraman, 2025; Roe & Perkins, 2024). However, the use of AI in education also carries risks of diminishing agency. One of the most prominent risks is cognitive offloading, where students overly rely on technology instead of developing their own mental strategies—thereby weakening critical thinking, self-reflection, and independent problem-solving skills (Shum, 2024; Holmes, 2024). As seen in early AI approaches (Paradigm One: AI-directed, learner-as-recipient), AI systems may predetermine learning paths, restricting learners' autonomy (Ouyang & Jiao, 2021). AI can also steer student preferences toward narrow algorithmic defaults, increasing the risk of modal collapse and creating an algorithmic panopticon that limits meaningful choice (Bozkurt, 2025). To preserve human agency, it is essential that educators maintain control over core pedagogical decisions (human oversight) and employ AI in ways that support distinctly human capacities such as critical thinking, creativity, and ethical judgment (Mouta, Pinto-Llorente & Torrecilla-Sánchez,

2025; Sethuraman, 2025). AI should augment, not replace, human intelligence (Shum, 2024).

Ethical Challenges of AI in Educational Contexts

The ethical use of AI in education is about how to keep alive basic values – autonomy, justice, privacy, and accountability – in environments increasingly mediated by data and algorithms. While AI systems have the capacity of enhancing learning contexts, they also serve to instantiate new asymmetries of knowledge and power, especially when learners become data subjects and educators are subsumed under algorithmic decision-making (Nguyen et al., 2023; Adams et al., 2023). This section analyses four interrelated domains that structure current ethical debates in educational AI: data ethics and privacy, algorithmic bias and fairness, transparency and explainability, and accountability and governance.

1.Data Ethics and Privacy

Data ethics and privacy poses among the most significant ethical problems to be resolved today within educational institutions (Mienye and Swart, 2025; Akgun and Greenhow, 2022). Student and teacher privacy abuses have heightened due to the growing use of AI systems. Privacy violations are situations where people disclose an excessive amount of personal information (including metadata) on the internet; such metadata may include linguistic features, racial identity, biographical information, and geolocation information (Regan and Jesse, 2019; Remian, 2019). Laws to protect sensitive data do exist, though violations of data access and security by big tech companies employing AI technologies have exacerbated privacy fears (Stockman and Nottingham, 2022). Although AI systems frequently ask users for permission to access their personal data, many users give it without understanding, and without having the chance to think first and foremost, about what they are revealing to others. Such unreflective data sharing infringes upon personal

autonomy (i.e., human agency) and control over their own privacy (Nguyen et al., 2023). Indeed, in the case that it is a school that requires such systems, students and parents become implicated in the question of ethics: “even if they explicitly consent to participate, they are being forced to take part in it, since they can no longer opt out” (Turner, Pothong & Livingstone, 2022).

Yet another major ethical question brought up with the incorporation of AI in education is surveillance. These systems utilize algorithms and machine learning models to harvest granular data collected on the actions and preferences of students and teachers. AI-based surveillance doesn’t just keep an eye on what people do; it also monitors and predicts what its users will do next (Charteris, 2022; Ryymin, 2021; Dai, Thomas and Rawolle, 2025). For example, these monitoring technologies can be embedded as predictive systems designed to forecast learners’ learning performance, strengths, weaknesses, and behavioral patterns (Alamri & Alharbi, 2021; Almalawi, Soh and Samra, 2024). Issues arise when such predictors do end up challenging the autonomy of the individual, the ability to act on what is to one’s liking or on an even-valued basis. Algorithmic predictions threaten that autonomy and freedom, which both students and educators require, for information. Once students know that their thoughts and behaviours are tracked by the AI systems, they become restricted in the extent of their learning engagement and lose confidence in the degree to which their ideas are their own (Lo Piano, 2020; Regan & Jesse, 2019; Akgun & Greenhow, 2022). We can mitigate these safety and privacy issues by educating both teachers and learners at a better level about the ethical nature of AI. For this reason, the MIT Media Lab and others have launched a series of "AI and Data Privacy" workshops on the topic, which engage students ages 7–14 years old in critical reflection (Akgun & Greenhow, 2022). They encourage learners to examine how algorithms interpret human behavior and

also how a child's choices with regard to online consent and the legal principles of the Children's Online Privacy Protection Act (COPPA) are an issue for them (Anderson, 2024).

Additionally, regions like the European Union (EU) have published guidelines addressing ethical responsibilities, such as the Assessment List for Trustworthy AI (ALTAI) (Hleg, 2020). Its guidelines emphasize privacy, preventing surveillance, and eliminating discrimination as matters of utmost importance. As the role of AI in education continues to expand, what practitioners need to know about potential dangers and ethical aspects is the biggest challenge.

2.Algorithmic Bias and Fairness

Algorithmic bias is one of the most clearly documented ethical challenges in educational AI. Systems trained on historical data reproduce existing inequalities and encode cultural, linguistic, or gendered biases (Baker & Hawn, 2022) For example, automated essay scoring models have been found to favour academic writing styles that conform to the norms of the dominant linguistic culture, which, in some cases, has led to the exclusion of minority language learners (Matta, Mercer & Keller-Margulis, 2023). This aligns predictive analytics used for at-risk identification — predictive models often link socio-economic characteristics to performance, exacerbating structural disadvantage (Almalawi, Soh & Samra, 2024). Fairness in education includes perceived justice, beyond statistical inequities, — the extent to which students and teachers see algorithmic decisions as legitimate (Lünich, Keller & Marcinkowski, 2024). Jones-Jang et al. (2025) emphasize that fairness perception is determined by transparency, feedback mechanisms, and the ability to question AI outputs. Without them, even technically correct systems undermine institutional trust. It takes multi-level intervention to mitigate algorithmic bias: inclusive

sampling of data, bias auditing, explainable decision-path visualization, and a hybrid evaluation using human & machine judgment. Fairness, cannot be completely automated, but must be "co-produced through human oversight and algorithmic accountability" (Kyriakou & Otterbacher, 2023).

3.Transparency and Explainability

Transparency and explainability are two epistemic preconditions of ethical AI for education (Contreras & Jaimes, 2024). When users cannot see what any given output from a system means—as a score, as a recommendation, or as feedback—the validity of such a decision vanishes. But most of these educational AI solutions are black boxes, especially the deep neural network ones, which are not subject to pedagogical scrutiny (Yue, Jong & Dai, 2022; Balasubramaniam et al., 2023). The difficulty of interpreting the decisions of complex AI models—such as Large Language Models (LLMs)—undermines trust among educators, students, and other stakeholders (Geethanjali & Umashankar, 2025). For this reason, transparency and explainability are emphasized as essential quality requirements in ethical guidelines for AI systems and are considered fundamental prerequisites for the successful integration of AI into educational environments (Balasubramaniam et al., 2023; Raza et al., 2024). Nearly all organizations examined highlight the importance of transparency and regard explainability as an integral component of it. The primary purpose of incorporating transparency and explainability is to build and maintain trust. In high-stakes contexts such as education, AI systems must be transparent, auditable, and aligned with human values (Balasubramaniam et al., 2023).

To address this transparency challenge, **Explainable Artificial Intelligence (XAI)** has emerged as a critical paradigm aimed at reducing opacity in educational AI applications (Ali &

Husain, 2015; Geethanjali & Umashankar, 2025). XAI seeks to make AI models understandable to humans by providing clear and justified explanations for their decisions—for example, in automated grading, personalized learning pathways, or virtual tutor recommendations (Rachha & Seyam, 2023). Such explanations may be produced through inherently interpretable models, such as decision trees or rule-based systems, or through post-hoc techniques like **SHAP** and **LIME (**Raza et al., 2024)**.**

XAI fosters deeper learning by enabling students not only to see *what* they should improve, but also to understand *why* these improvements are necessary (Singh, 2025). Furthermore, transparency and explainability are essential for ensuring fairness, as they help identify and mitigate algorithmic bias and ethical concerns during system development and deployment (Akinrinola et al., 2024; Johnson & Zhang, 2024). In this way, XAI supports the adoption of AI systems as trustworthy and responsible partners that reinforce educational goals (Rachha & Seyam, 2023).

4.Accountability and Governance

The widespread adoption of artificial intelligence (AI) systems in education has created an ethical dilemma centered on the question of **who should be held responsible** in high-stakes scenarios involving algorithmically mediated decisions—for example, exam scoring or vocational guidance systems (Herrera-Poyatos et al., 2025; Ramnani, 2024). **Accountability** refers to being answerable for the consequences of one's actions or decisions, and it is regarded as a foundational component of a democratic, tolerant, and inclusive society (Porayska-Pomsta & Rajendran, 2019). In education, accountability is closely tied to internal performance monitoring, which ensures that institutional decisions align with intended outcomes (Algazo & Ibrahim, 2024). However, the tendency of digital governance systems to rely on standardized and

quantifiable data may conflict with educators' professional autonomy by constraining their discretion and their ability to adapt decisions to local contexts. This risks limiting *phronesis*—the capacity for context-informed practical judgment (Larsen, 2025).. Because the complexity and opacity of AI systems may obscure decision-making processes, the absence of clear accountability mechanisms risks turning ethical failures into technological inevitabilities. In Responsible AI (RAI) systems, **auditability** is an *ex ante* requirement ensuring that decisions and processes are traceable and verifiable, while **accountability** pertains to *post hoc* evaluation of whether the system has performed as intended (Herrera-Poyatos et al., 2025).

The responsible integration of AI applications in education requires a holistic governance framework that addresses ethical, legal, and social dimensions. **Governance** defines how power is distributed, how resources are managed, and how complex systems are directed. In higher education, AI integration has the potential to enhance decision-making and operational efficiency through data-driven insights and automation (Herrera-Poyatos et al., 2025; Algazo & Ibrahim, 2024; Mariam, Adil & Zakaria, 2024). Yet this also introduces challenges related to data privacy, ethical concerns, and institutional power dynamics. Proposed frameworks such as the AI Ecological Education Policy Framework aim to address these challenges across three main dimensions: governance, pedagogy, and operations. Within this framework, the **Governance Dimension**, initiated by institutional leadership, bears primary responsibility for addressing ethical concerns such as academic integrity, data privacy, transparency, accountability, and security (Chan, 2023).. Institutions must be transparent about the algorithms they use, their functions, and their potential biases or limitations—an essential step in building trust among students and staff regarding the use of AI technologies. Ultimately, governance for responsible

AI systems should focus on ethical and lawful use, encouraging policymakers and administrators to develop frameworks that ensure processes remain transparent, inclusive, and aligned with educational values (Mariam, Adil & Zakaria, 2024).

Hallucination and the Epistemic Integrity of AI Systems

The rapid integration of Artificial Intelligence (AI) systems—particularly Large Language Models (LLMs)—into critical domains such as education and science introduces significant challenges to epistemic integrity (Chen, 2025). Hallucinations occur when an AI system produces information that is fabricated, nonsensical, or factually incorrect, even though it appears fluent, syntactically coherent, and persuasive (Wachter, Mittelstadt & Russell, 2024; Li, Yi & Chen, 2025). This phenomenon stems from the foundational architecture of such models: LLMs operate not through conceptual understanding or causal reasoning, but by predicting the next token in a sequence based on statistical patterns within vast datasets. This probabilistic nature is the core reason why hallucination is an inherent feature of LLM outputs (Yingzhe, 2025; Li, 2023). Hallucinations may take the form of factual inconsistencies, invented references, or subtler distortions such as consensus illusion or oversimplification (Li, Yi & Chen, 2025; Yingzhe, 2025). The polished and authoritative style in which such inaccuracies are presented can diminish users'—especially students'—capacity for critical evaluation. Hallucinatory content can disrupt students' conceptual scaffolding, increasing the risk of developing ingrained misconceptions that are difficult to correct in the long term (Elsayed, 2024; Yingzhe, 2025; Ayeni et al.,2024) . This threat to epistemic integrity is not merely a technical flaw but points toward a deeper philosophical issue known as the "Accuracy Paradox". The paradox suggests that excessive optimization for accuracy, intended to reduce hallucinations, may cause greater harm by generating an illusion of epistemic certainty and fostering

uncritical user trust. While accuracy typically denotes statistical consistency with existing "ground truth" datasets, epistemic truth is more complex and requires contextualization, justification, and robustness against error (Li, Yi & Chen, 2025; Laux, Wachter & Mittelstadt). Since LLMs are often optimized for fluency and rhetorical persuasiveness, they may convince users of their reliability even when their outputs lack epistemic validity. The portrayal of AI systems as "knowing agents" and the attribution of human-like cognitive status to them (anthropomorphism) blur the boundaries between human and machine epistemic processes, leading users to overestimate AI capabilities. This dynamic can constrain human epistemic agency, the essential ethical capacity of individuals to maintain control over their own processes of knowledge formation—crucial in educational contexts (Chen, 2025).

To safeguard epistemic integrity and overcome the Accuracy Paradox, regulatory frameworks and system designs must shift away from narrow accuracy metrics toward epistemic reliability (Li, Yi & Chen, 2025). In scientific domains, ensuring that LLMs make reliable contributions requires embedding them into rigorous workflows. Systems such as AlphaFold and GenCast employ strategies like theory-guided training (encoding physical and chemical laws to guide learning) and confidence-based error screening (e.g., the ensemble of probabilistic predictions in GenCast or the pLDDT scores in AlphaFold) to flag potential errors (Rathkopf, 2025).. In education, however, the primary challenge is resisting the development of passive dependency habits that may arise from the convenience and speed of AI tools (Chen, 2025). Therefore, it is essential to equip students to become critical evaluators of AI-generated content (Yingzhe, 2025). Pedagogical strategies should explicitly teach verification techniques such as lateral reading and aim to cultivate epistemic sensitivity—the ability

to recognize when critical scrutiny is necessary (Elsayed, 2024). Ultimately, overcoming these challenges requires an approach to AI integration that leverages the power of such systems while upholding core educational values and aligning with human expertise and epistemic agency (Chen, 2025).

Responsible AI Frameworks for Educational Practice

The rapid integration of Artificial Intelligence (AI) systems—especially generative AI models such as Large Language Models (LLMs)—into educational environments offers unprecedented opportunities for personalizing learning experiences, optimizing administrative processes, and enhancing instructional quality (Nguyen & Nguyen, 2025; Chan et al., 2025; Bearman, Ryan & Ajjawi, 2023). However, this technological expansion also brings complex ethical challenges and risks, including issues related to academic integrity (plagiarism, overreliance), data privacy and security, and algorithmic bias (Chan et al., 2025; Tirado, Mulholland & Fernandez, 2024; Zhu, Sun & Yang, 2025). To address these challenges and ensure the equitable distribution of AI's benefits, Responsible and Trustworthy AI (RAI) frameworks have become essential. These frameworks aim to ensure that AI systems are designed and deployed in ways that minimize potential harm and maximize societal benefit (Tirado, Mulholland & Fernandez, 2024). Core principles of Responsible AI include fairness and bias mitigation, transparency, accountability, safety, and explainability. International organizations (such as UNESCO) and industry initiatives converge around these foundational values (Nguyen & Nguyen, 2025; Tirado, Mulholland & Fernandez, 2024).

Applying Responsible AI frameworks in educational practice requires translating abstract ethical principles into concrete, actionable strategies. To this end, models tailored to specific contexts—such as Learning Analytics (LA)—have been developed

(Tirado, Mulholland & Fernandez, 2024). For example, an integrated responsible and trustworthy AI framework has been created to analyze and support students' learning engagement, implemented across five stages and encompassing three modules: Explainable AI (XAI), Safeguard and Auditing, and Adversarial Training. XAI models provide interpretable information—such as decision rules or variable importance rankings—to evaluate learning performance. Safeguard and Auditing modules provide complementary predictions to prevent students with poor learning performance from being misidentified as normal learners and to issue early warnings for at-risk students. Another structural model, the Integrity AI Model, guides ethical AI integration by focusing on three levels: guiding principles, educational activities, and impact/empowerment. Additionally, the GAIDL Framework, designed to offer practical guidance for Higher Education Institutions (HEIs), aligns ethical AI considerations with the stages of the software development life cycle (requirements and data collection, design, development, testing, deployment, and monitoring) (Chou, 2023).

The success of Responsible AI frameworks depends on how much stakeholders—especially teachers and students—value ethical priorities. Studies involving K-12 teachers show that fairness and safety consistently emerge as the highest priority values across different scenarios. Fairness requires that AI systems do not perpetuate existing inequities and ensure equitable outcomes for all learners. Safety encompasses the accuracy, reliability, and robustness of AI systems, while also aiming to minimize psychological, emotional, or academic harm to staff and students. Transparency, on the other hand, is critical for building trust by enabling users to understand the reasons behind algorithmic decisions and for strengthening students' capacity for critical thinking (Yin, Karumbaiah & Acquaye, 2025). However, the lack of interpretability mechanisms in most models (88.1%) creates a

significant transparency gap—one that conflicts with regulatory frameworks such as the EU AI Act, which mandates transparency in high-risk AI applications (Floridi et al., 2018). Effective RAI integration requires continuous AI literacy and ethics education for both teachers and students (Nofirman et al., 2025; Smith et al., 2025; Chan et al., 2025). Such training should prepare students to critically evaluate AI's limitations, bias potentials, and the trustworthiness of its outputs (Watson, 2025). This multidimensional approach is essential to ensure that AI functions as a responsible and ethical ally in education (Nguyen & Nguyen, 2025).

Conclusion and Policy Recommendations

Artificial Intelligence (AI) has shifted from its technical utility to an epistemic and moral actor that profoundly changes the ways we teach, learn and assess. With educational systems becoming weighted toward computational judgment, the historically normative foundations of education—including learner autonomy (which is central to many types of pedagogic practice), evidential rigor (and thus to pedagogical legitimacy), epistemic trust, and the moral agency exercised by the educator themselves—are at odds with inscrutable algorithmic architectures that are neither fully questioned nor contested. This shift compels education to face a difficult trade-off between two imperatives: to seize personalization, scalability, and efficiency from AI, while at the same time safeguarding fairness, academic integrity, transparency, and accountability as non-negotiable ethical commitments.

Among the most significant tensions unveiled in this chapter are those of epistemic dependence and hallucination. Epistemic dependence arises when fluent algorithmic outputs replace the critical reasoning of students and educators, thereby progressively making deference to machine-generated knowledge the new normal. Hallucination is, on the other hand, a case of deeper epistemic

instability—the generation of linguistically consistent but factually incorrect content, that undermines the very scaffolding of conceptual understanding. These risks taken together highlight a wider accountability gap: the opacity of AI decision-making can turn ethical failures into apparently unavoidable technological results.

In this regard, Responsible AI (RAI) should not be considered an extension of technology, but must be understood as the epistemological and moral prerequisite to maintaining truth, justice, and trust in digitally mediated education. To ensure successful implementation of RAI, they need a governance approach that combines innovation, as well as ethical stewardship. This new model must ensure multi-stakeholder accountability (developer vs. institution vs. educator vs. learner) while also ensuring that AI does not substitute for, but rather augments, human judgment.

Human-in-the-loop oversight is important in pedagogical decision-making, allowing reflective judgment to remain an educational value, as opposed to an operational constraint. Explainable AI (XAI) paradigms are the key to this, they offer the understanding when reasoning becomes interpretable algorithm, letting students understand not just what to do better but why. At the same time, AI literacy – the ability to discern, verify, and question AI outputs without simply consuming them — needs to be a core part of modern education. Educators, for their part, need to be prepared to use AI in ways that reinforce human capabilities such as critical thinking, creativity, ethical reasoning and contextual judgment.

If educational systems are to cultivate a future shaped by AI in ethically and epistemically robust ways, the primary challenge resides not in resisting technological transformation, but in designing and governing AI systems that expand rather than diminish human agency. Embedding ethical, legal, pedagogical, and

epistemological principles across all stages of institutional AI development and deployment will enable the emergence of learning environments that are technologically sophisticated, equitable, transparent, and resilient in their epistemic foundations.

References

Abu-Rasheed, H., Weber, C., & Fathi, M. (2024, May). Knowledge graphs as context sources for llm-based explanations of learning recommendations. In *2024 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1-5). IEEE.

Adams, C., Pente, P., Lemermeyer, G., & Rockwell, G. (2023). Ethical principles for artificial intelligence in K-12 education. *Comput. Educ. Artif. Intell.*, *4*, 100131.

Akgun, S., & Greenhow, C. (2022). Artificial intelligence in education: Addressing ethical challenges in K-12 settings. *AI and Ethics*, *2*(3), 431-440.

Akinrinola, O., Okoye, C. C., Ofodile, O. C., & Ugochukwu, C. E. (2024). Navigating and reviewing ethical dilemmas in AI development: Strategies for transparency, fairness, and accountability. *GSC Advanced Research and Reviews*, *18*(3), 050-058.

Alamri, R., & Alharbi, B. (2021). Explainable student performance prediction models: a systematic review. *IEEE Access*, *9*, 33132-33143.

Ali, S., & Husain, A. (2015). Explainable AI for Next-Generation Learning Systems: A Path Towards Trust and Transparency. *e-IIEMS J*, 59.

Algazo, F. A., & Ibrahim, S. (2024). University governance and accountability. *Asian Journal of Research in Education and Social Sciences*, *6*(2), 528-535.

Almalawi, A., Soh, B., Li, A., & Samra, H. (2024). Predictive models for educational purposes: A systematic review. *Big Data and Cognitive Computing*, *8*(12), 187.

Anderson, H. (2024). The guardian of the digital era: Assessing the impact and challenges of the children's online privacy protection act. *Law and Economy*, *3*(2), 6-10.

Ayeni, O. O., Al Hamad, N. M., Chisom, O. N., Osawaru, B., & Adewusi, O. E. (2024). AI in education: A review of personalized learning and educational technology. *GSC Advanced Research and Reviews*, *18*(2), 261-271.

Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, *7*(1), 52-62.

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 32 (4), 713–742. https://doi.org/10.1007/s40593-022-00292-9

Balasubramaniam, N., Kauppinen, M., Rannisto, A., Hiekkanen, K., & Kujala, S. (2023). Transparency and explainability of AI systems: From ethical guidelines to requirements. *Information and Software Technology*, *159*, 107197.

Bearman, M., Ryan, J., & Ajjawi, R. (2023). Discourses of artificial intelligence in higher education: A critical literature review. *Higher Education*, *86*(2), 369-385.

Bozkurt, A. (Ed.). (2025). Algorithmically manufactured minds: Generative and agentic AI in a time of post-truth, reconfiguration of student agency and death of critical pedagogy. *Open Praxis*, *17*(2), 206-210.

Cao, M. (2025). Factual Consistency in Neural Text Generation: Detecting, Correcting, and Understanding Hallucinations (Doctoral dissertation, McGill University (Canada)).

Chan, C. K. Y. (2023). A comprehensive AI policy education framework for university teaching and learning. *International journal of educational technology in higher education*, *20*(1), 38.

Chan, M. M., Rosales, M., Hernandez-Rizzardini, R., & Amado-Salvatierra, H. R. (2025, April). Ethical AI in Education: A Proposed Model for Responsible Integration. In *2025 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1-7). IEEE.

Chandrakant, N. S. M. (2025). AI-powered teaching assistants: Enhancing educator efficiency with NLP-based automated feedback systems. *International Journal of Science and Research Archive*, 14(3), 009-018.

Charteris, J. (2022). Post-panoptic accountability: making data visible through 'data walls' for schooling improvement. *British Journal of sociology of Education*, *43*(3), 333-348.

Chou, T.N. (2023). Apply an Integrated Responsible AI Framework to Sustain the Assessment of Learning Effectiveness. In *Proceedings of the 15th International Conference on Computer Supported Education (CSEDU 2023)* - Volume 2, pages 142-149 ISBN: 978-989-758-641-5; ISSN: 2184-5026

Chen, B. (2025). Beyond tools: Generative AI as epistemic infrastructure in education. *arXiv preprint arXiv:2504.06928*.

Chu, Z., Wang, S., Xie, J., Zhu, T., Yan, Y., Ye, J., ... & Wen, Q. (2025). Llm agents for education: Advances and applications. *arXiv preprint arXiv:2503.11733*.

Creely, E., & Carabott, K. (2025). Teaching and learning with AI: an Integrated AI-Oriented Pedagogical Model. *The Australian Educational Researcher*, 1-22.

Contreras, M. R., & Jaimes, J. O. P. (2024, October). AI Ethics in the Fields of Education and Research: A Systematic

Literature Review. In *2024 International Symposium on Accreditation of Engineering and Computing Education (ICACIT)* (pp. 1-6). IEEE.

Daher, R. (2025). Integrating AI literacy into teacher education: a critical perspective paper. *Discover Artificial Intelligence*, *5*(1), 217.

Dai, R., Thomas, M. K. E., & Rawolle, S. (2025). Revisiting Foucault's panopticon: how does AI surveillance transform educational norms?. *British Journal of Sociology of Education*, *46*(5), 650-668.

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

Fajardo-Ramos, D. C., Andrés, C., & Mella-Norembuena, J. (2025). *Human-in-the-Loop Assessment with AI: Implications for Teacher Education in Ibero-American Universities*. In Frontiers in Education (Vol. 10, p. 1710992). Frontiers.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and machines*, *28*(4), 689-707.

Fu, Y., & Weng, Z. (2024). Navigating the ethical terrain of AI in education: A systematic review on framing responsible human-centered AI practices. *Computers & Education: Artificial Intelligence,* 7, 100306. https://doi.org/10.1016/j.caeai.2024.100306

Geethanjali, K. S., & Umashankar, N. (2025, February). Enhancing Educational Outcomes with Explainable AI: Bridging

Transparency and Trust in Learning Systems. In *2025 International Conference on Emerging Systems and Intelligent Computing (ESIC)* (pp. 325-328). IEEE.

Grinschgl, S., & Neubauer, A. C. (2022). Supporting cognition with modern technology: Distributed cognition today and in an AI-enhanced future. *Frontiers in Artificial Intelligence*, *5*, 908261.

Guo, S., Latif, E., Zhou, Y., Huang, X., & Zhai, X. (2024). Using generative AI and multi-agents to provide automatic feedback. *arXiv preprint arXiv:2411.07407*.

Herrera-Poyatos, A., Del Ser, J., de Prado, M. L., Wang, F. Y., Herrera-Viedma, E., & Herrera, F. (2025). Responsible Artificial Intelligence Systems: A Roadmap to Society's Trust through Trustworthy AI, Auditability, Accountability, and Governance. *arXiv preprint arXiv:2503.04739*.

Hleg, A. (2020). *The assessment list for trustworthy AI (ALTAI). High-level expert group on artificial intelligence. European Commission, Brussels*.

Hu, B., Zheng, L., Zhu, J., Ding, L., Wang, Y., & Gu, X. (2024). Teaching plan generation and evaluation with GPT-4: Unleashing the potential of LLM in instructional design. *IEEE Transactions on Learning Technologies*, *17*, 1445-1459.

Holmes, W. (2024). AI, AIED and human agency. AI For Teachers (AI4T).

Jones-Jang, S. M., Park, J., & Lee, M. (2025). Fairness perceptions of AI in grading systems: Examining how discontent with the status quo and outcome favorability reduce AI reluctance. *Computers & Education: Artificial Intelligence*, 8, 100385. https://doi.org/10.1016/j.caeai.2025.100419

Johnson, M., & Zhang, M. (2024). Examining the responsible use of zero-shot AI approaches to scoring essays. *Scientific Reports*, *14*(1), 30064.

Jose, B., Cleetus, A., Joseph, B., Joseph, L., Jose, B., & John, A. K. (2025). Epistemic authority and generative AI in learning spaces: rethinking knowledge in the algorithmic age. *Frontiers in Education* 10 (1647687). Frontiers.

Kahl, P. (2025). Epistemic clientelism theory: Power dynamics and the delegation of epistemic agency in academia.

Katsenou, R., Kotsidis, K., Papadopoulou, A., Anastasiadis, P., & Deliyannis, I. (2025). Beyond Assistance: Embracing AI as a Collaborative Co-Agent in Education. *Education Sciences*, *15*(8), 1006.

Kyriakou, K., & Otterbacher, J. (2023). In humans, we trust: Multidisciplinary perspectives on the requirements for human oversight in algorithmic processes. *Discover Artificial Intelligence*, *3*(1), 44.

Larsen, E. (2025, June). Democratic Accountability in the Digital Governance of Education: A Review of Tensions and Challenges. In *Conference Proceedings. The Future of Education 2025*.

Laux, J., Wachter, S., & Mittelstadt, B. (2024). Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*, *18*(1), 3-32.

Li, H., Xu, T., Zhang, C., Chen, E., Liang, J., Fan, X., ... & Wen, Q. (2024). Bringing generative AI to adaptive learning in education. *arXiv preprint arXiv:2402.14601*.

Li, Z., Yi, W., & Chen, J. (2025). Beyond Accuracy: Rethinking Hallucination and Regulatory Response in Generative AI. *arXiv preprint arXiv:2509.13345*.

Li, Z. (2023). Why the European AI Act transparency obligation is insufficient. *Nature machine intelligence*, *5*(6), 559-560.

Lo Piano, S. (2020). Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. *Humanities and Social Sciences Communications*, *7*(1), 1-7.

López-Meneses, E., López-Catalán, L., Pelícano-Piris, N., & Mellado-Moreno, P. C. (2025). Artificial Intelligence in Educational Data Mining and Human-in-the-Loop Machine Learning and Machine Teaching: Analysis of Scientific Knowledge. *Applied Sciences*, 15(2), 772.

Lünich, M., Keller, B., & Marcinkowski, F. (2024). Fairness of academic performance prediction for the distribution of support measures for students: Differences in perceived fairness of distributive justice norms. *Technology, Knowledge and Learning*, *29*(2), 1079-1107.

Mariam, G., Adil, L., & Zakaria, B. (2024). The integration of artificial intelligence (ai) into education systems and its impact on the governance of higher education institutions. *International Journal of Professional Business Review: Int. J. Prof. Bus. Rev.*, *9*(12), 13.

Marshall, S., Blaj-Ward, L., Dreamson, N., Nyanjom, J., & Bertuol, M. T. (2024). The reshaping of higher education: technological impacts, pedagogical change, and future projections. *Higher Education Research & Development*, *43*(3), 521-541.

Martin, F., Chen, Y., Moore, R. L., & Westine, C. D. (2020). Systematic review of adaptive learning research designs, context, strategies, and technologies from 2009 to 2018. *Educational Technology Research and Development*, *68*(4), 1903-1929.

Matta, M., Mercer, S. H., & Keller-Margulis, M. A. (2023). Implications of bias in automated writing quality scores for fair and equitable assessment decisions. *School Psychology*, *38*(3), 173.

Memarian, B., & Doleck, T. (2023). Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and higher education: A systematic review. *Computers and Education: Artificial Intelligence*, 5, 100152.

Mienye, I. D., & Swart, T. G. (2025). ChatGPT in education: A review of ethical challenges and approaches to enhancing transparency and privacy. *Procedia Computer Science*, *254*, 181-190.

Mouta, A., Pinto-Llorente, A. M., & Torrecilla-Sánchez, E. M. (2025). "Where is Agency Moving to?": Exploring the Interplay between AI Technologies in Education and Human Agency. *Digital Society*, *4*(2), 49.

Nair, I. J., Tan, J., Su, X., Gere, A., Wang, X., & Wang, L. (2024, November). Closing the Loop: Learning to Generate Writing Feedback via Language Model Simulated Student Revisions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 16636-16657).

Ng, D. T. K., Su, J., Leung, J. K. L., & Chu, S. K. W. (2024). Artificial intelligence (AI) literacy education in secondary schools: a review. *Interactive Learning Environments*, *32*(10), 6204-6224.

Nguyen, A., Ngo, H. N., Hong, Y., Dang, B., & Nguyen, B. P. T. (2023). Ethical principles for artificial intelligence in

education. *Education and information technologies*, *28*(4), 4221-4241.

Nguyen, P. B. T., & Nguyen, T. N. P. (2025). Ethıcal And Responsıble Aı In Educatıonal Use Of Chatgpt. *Tpm–Testing, Psychometrics, Methodology İn Applied Psychology*, *32*(S7 (2025): Posted 10 October), 1322-1335.

Nofirman, N., Vann, R., Sefrizal, L., & Dara, R. (2025). The Ethics of AI in Education: An Empirical Study on Students' and Teachers' Attitudes toward Responsible AI Use. *Al-Hijr: Journal of Adulearn World*, *4*(2), 101-114.

Ouyang, F., & Jiao, P. (2021). Artificial intelligence in education: The three paradigms. *Computers and Education: Artificial Intelligence*, *2*, 100020.

Pitts, G., Rani, N., Mildort, W., & Cook, E. M. (2025). Students' Reliance on AI in Higher Education: Identifying Contributing Factors. *arXiv preprint arXiv:2506.13845*.

Porayska-Pomsta, K., & Rajendran, G. (2019). Accountability in human and artificial intelligence decision-making as the basis for diversity and educational inclusion. In *Artificial intelligence and inclusive education: Speculative futures and emerging practices* (pp. 39-59). Singapore: Springer Singapore.

Pratiwi, H., Riwanda, A., Hasruddin, H., Sujarwo, S., & Syamsudin, A. (2025). Transforming Learning or Creating Dependency? Teachers' Perspectives and Barriers to AI Integration in Education. *Journal of Pedagogical Research*, 9(2), 127-142.

Rachha, A., & Seyam, M. (2023). Explainable AI in education: Current trends, challenges, and opportunities. *SoutheastCon 2023*, 232-239.

Ramnani, S. (2024). Exploring Ethical Considerations of Artificial Intelligence in Educational Settings: An Examination of Bias, Privacy, and Accountability. *International Journal of Novel Research and Development (IJNRD)*, *9*(2), 2456-4184.

Rathkopf, C. (2025). Hallucination, reliability, and the role of generative AI in science. *arXiv preprint arXiv:2504.08526*.

Raza, S., Fatima, I., Arif, S., Sharif, M., Jalal, M. S., & Muhammad, Z. (2024). The future of learning: building trust and transparency in AI education. *Journal of management practices, humanities and social sciences.*, *8*(3), 62-74.

Regan, P. M., & Jesse, J. (2019). Ethical challenges of edtech, big data and personalized learning: Twenty-first century student sorting and tracking. *Ethics and Information Technology*, *21*(3), 167-179.

Remian, D. (2019). Augmenting education: ethical considerations for incorporating artificial intelligence in education.

Roe, J., & Perkins, M. (2024). Generative AI and agency in Education: A critical scoping review and thematic analysis. *arXiv preprint arXiv:2411.00631*.

Ryymin, E. (2021). Perspectives from higher education: applied sciences university teachers on the digitalization of the bioeconomy. *Technology Innovation Management Review*, *11*(2).

Saarela, M., Gunasekara, S., & Karimov, A. (2025, May). The EU AI Act: Implications for Ethical AI in Education. In *International Conference on Design Science Research in Information Systems and Technology* (pp. 36-50). Cham: Springer Nature Switzerland.

Sethuraman, K. R. (2025). Artificial Intelligence and Education: Preserving Human Agency in a World of

Automation. *SBV Journal of Basic, Clinical and Applied Health Science*, *8*(2), 41-42.

Shahzad, T., Mazhar, T., Tariq, M. U., Ahmad, W., Ouahada, K., & Hamam, H. (2025). A comprehensive review of large language models: issues and solutions in learning environments. *Discover Sustainability*, *6*(1), 27.

Sharma, S., Mittal, P., Kumar, M., & Bhardwaj, V. (2025). The role of large language models in personalized learning: a systematic review of educational impact. *Discover Sustainability*, 6(1), 1-24.

Shum, S. B. (2024, January). Generative AI for Critical Analysis Practical Tools Cognitive Offloading and Human Agency. In *CEUR Workshop Proceedings*. CEUR-WS.

Singh, A. (2025). Evaluating the transparency and explainability of llm-based educational systems. *Available at SSRN 5198565*.

Smith, S. M., Tate, M., Freeman, K., Walsh, A., Ballsun-Stanton, B., & Lane, M. (2025). A university framework for the responsible use of generative AI in research. *Journal of Higher Education Policy and Management*, 1-20.

Stockman, C., & Nottingham, E. (2022). Surveillance capitalism in schools: What's the problem. *Digital Culture & Education*, *14*(1), 1-15.

Southworth, J., Migliaccio, K., Glover, J., Glover, J. N., Reed, D., McCarty, C., ... & Thomas, A. (2023). Developing a model for AI Across the curriculum: Transforming the higher education landscape via innovation in AI literacy. *Computers and Education: Artificial Intelligence*, *4*, 100127.

Tirado, A. M., Mulholland, P., & Fernandez, M. (2024). Towards an operational responsible AI framework for learning analytics in higher education. *arXiv preprint arXiv:2410.05827*.

Turner, S., Pothong, K., & Livingstone, S. (2022). Education Data Reality: The challenges for schools in managing children's education data.

Varshini, K. N., & Katreddi, S. S. (2025). Education in The Age of AI: Challenges in Ethics, Technology, and Pedagogy. *Systemic Analytics*, *3*(4), 242-247.

Wachter, S., Mittelstadt, B., & Russell, C. (2024). Do large language models have a legal duty to tell the truth?. *Royal Society Open Science*, *11*(8), 240197.

Yeh, S. N., & Siah, C. J. R. (2025). Students' Attitudes Toward LLMs and Its Association with Metacognitive Abilities: A Cross-sectional Study. *Nurse Education in Practice*, 104567.

Yin, Y., Karumbaiah, S., & Acquaye, S. (2025, June). Responsible AI in Education: Understanding Teachers' Priorities and Contextual Challenges. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2705-2727).

Yingzhe, L. I. (2025). Addressing "Hallucinations" in AI-Generated Content: Strategies for Developing Student Fact-Checking and Information Evaluation Skills. *Artificial Intelligence Education Studies*, *1*(2), 48-62.

Yue, M., Jong, M. S. Y., & Dai, Y. (2022). Pedagogical design of K-12 artificial intelligence education: A systematic review. *Sustainability*, *14*(23), 15620.

Watson, J. (2025). Ethical Considerations for Responsible AI Use in the Classroom. *NACTA Journal*, *69*(TT).

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., ... & Li, Y. (2021). A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. *Complexity*, *2021*(1), 8812542.

Zhang, Z., Dai, Q., Bo, X., Ma, C., Li, R., Chen, X., ... & Wen, J. R. (2025). A survey on the memory mechanism of large language model-based agents. *ACM Transactions on Information Systems*, *43*(6), 1-47.

Zhao, R., Bobrov, A., Li, J., Aloisi, C., & He, Y. (2025, November). Learnlens: Llm-enabled personalised, curriculum-grounded feedback with educators in the loop. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 625-633).

Zhu, H., Sun, Y., & Yang, J. (2025). Towards responsible artificial intelligence in education: a systematic review on identifying and mitigating ethical risks. *Humanities and Social Sciences Communications*, *12*(1), 1-14.

Zhuang, M., Long, S., Martin, F., & Castellanos-Reyes, D. (2025). The affordances of Artificial Intelligence (AI) and ethical considerations across the instruction cycle: A systematic review of AI in online higher education. *The Internet and Higher Education*, 101039.


BİDGE