trends in higher education

MDPI

Review

Harnessing Large Language Models for Scalable and Effective Formative Assessment in Higher Education: A Review

Charith Narreddy [1,2], Steve Joordens [2,3] and Sapolnach Prompiengchai [2,4,*]

[1] College of Computing, Georgia Institute of Technology, Atlanta, GA 30332, USA; cnarreddy3@gatech.edu
[2] Clematis Research Empowerment Hub, Toronto, ON M2J 4X1, Canada; steve.joordens@utoronto.ca
[3] Department of Psychology, University of Toronto Scarborough, Toronto, ON M1C 1A4, Canada
[4] Department of Psychiatry, University of Oxford, Oxford OX3 7JX, UK
* Correspondence: sapolnach.prompiengchai@queens.ox.ac.uk

Abstract

Formative assessment is an integral component of higher education, fostering student learning through feedback, reflection, and iterative improvement. However, despite its pedagogical importance, widespread adoption of formative assessment is often hindered by time constraints, resource limitations, and scalability challenges. The objective of this study is to examine how large language models (LLMs) offer a potential solution to support and enhance formative assessment in higher education across diverse educational contexts by enabling automated, personalized, and scalable feedback that is sustainable and accessible. In this review, we comprehensively examine cutting-edge research and applications of LLMs in various components of formative assessment, including feedback generation, student self-assessment, peer review, and instructor support within the context of higher education. We explore the opportunities LLMs present in enhancing learning outcomes associated with formative assessments and current research gaps while critically discussing the challenges in practical implementations of integrating LLM-driven formative assessments in real-world classrooms. By synthesizing current advancements, this review provides educators and researchers with insights into the transformative potential and responsible implementation of LLM-driven formative assessments in higher education.

Keywords: formative assessment; generative AI; large-language model; educational technology

Academic Editor: Hani Morgan

Received: 22 July 2025
Revised: 1 October 2025
Accepted: 3 October 2025
Published: 22 October 2025

Citation: Narreddy, C.; Joordens, S.; Prompiengchai, S. Harnessing Large Language Models for Scalable and Effective Formative Assessment in Higher Education: A Review. *Trends High. Educ.* **2025**, *4*, 65. https://doi.org/10.3390/higheredu4040065

1. Introduction

Formative Assessments (FA) are student-centered processes designed to gather evidence of student learning and provide effective feedback aimed at enhancing future learning and teaching. These assessments help students understand their current position in the learning process, clarify their goals, and identify ways to bridge the gap between the two. FA encompasses three main components: evaluating the current knowledge levels of learners (i.e., where they are in their learning), clarifying desired goals and learning objectives (i.e., where they are supposed to be), and providing feedback to bridge their knowledge gap throughout the instructional process [1–3].

Individualized learning through FA supports the development of self-regulated learning (SRL) [4], a proactive and cyclical process where learners manage their own learning by cultivating metacognition, motivation, and strategic actions, often incorporating formative feedback [5,6]. Traditional classroom practices that embed FA include feedback-focused quizzes [7], student attitude checks [8], exit tickets [9,10], and strategic teacher-led questioning [11].

Moreover, FA promotes student learning in collaborative environments involving not only teachers and individual students but also peers. Peer assessment, where students evaluate the quality of their classmates' work or learning outcomes, is a common collaborative approach. Methods such as oral presentations, group discussions, and projects facilitate this peer engagement, contributing to deeper learning and shared understanding [12–14].

With the advent of the digital era, Technology-Enhanced Formative Assessment (FA) has been developed to increase engagement and interactivity [15]. These approaches allow for integration into online settings, such as through online reflections and group discussion forums that facilitate clarifying questions and elaboration on [16]. A key advantage of Technology-Enhanced FA is that many tasks can be completed asynchronously, offering flexibility for both learners and educators [17]. However, Technology-Enhanced FA is not limited to online environments and can also be employed in traditional classrooms, for example, by using electronic clickers to collect anonymous student responses [18].

Over time, educational technology supporting FA has evolved through several phases, where early computer-assisted learning tools in the 1980s and 1990s were utilized to administer quizzes or drill-and-practice software, which provided efficiency yet limited personalization. The rise in learning management software in the 2000s enabled the systematic integration of assessments into online and blended courses, offering automated grading, analytics dashboards, and other benefits. This led to the development of algorithm-driven customization, where student interaction data was utilized to adjust difficulty and provide target recommendations. Despite these advancements which expanded access and inclusivity, they often relied on rule-based or narrowly scoped algorithms lacking the flexibility to capture the complexity of student reasoning or provide nuanced feedback, where several barriers to effective FA implementation remain.

Teachers often face significant time constraints in designing assessments and reflecting on student results, which can contribute to an unsustainable workload [14,19,20]. Large class sizes present another challenge, as they limit the teacher's ability to provide dynamic, individualized feedback [14,21]. Additionally, effective use of FA requires teachers to develop substantial pedagogical knowledge, a process that demands considerable time and effort [22]. Resource limitations at schools, both technological and physical, can further hinder the implementation of FA [23]. Even when resources are available, cultural barriers within teacher-centered educational systems often pose additional challenges to adopting effective FA practices [3,14,24,25].

Concerns previously identified in the implementation and procurement of formative assessment (FA) within educational systems can potentially be mitigated by recent advancements in Large Language Models (LLMs). These models are capable of processing and generating human-like text [26] through training on vast amounts of textual data [27]. With the rapid growth in both the number of LLMs and related research [28], these technologies have already been integrated into diverse educational contexts. Applications include solving student questions with tailored hints, automatic grading for educators, personalized feedback for learners, and support in setting educational goals [3,29,30]. These uses address key components of the FA process and highlight the potential for further integration into higher education systems.

Although the current literature explores the application of large-language models to specific aspects of formative assessment such as personalized feedback [31], tutoring [32], self-regulated learning [33], and automatic MCQ generation [34], these components are typically examined in isolation. In real classroom settings, however, different FA practices do not occur independently. For example, an effective FA requires that educators not only define clear learning objectives and rubrics, but also generate appropriate assessment tasks and provide feedback that enables students to refine their learning strategies.

Recognizing this, our review goes beyond prior work by synthesizing how LLMs can support the integrated implementation of FA. Specifically, we summarize recent advances in LLM technology for higher education, examine current evaluation metrics for LLM-driven FA tasks, and provide a comprehensive overview of how LLMs can help educators in post-secondary institutions address persistent barriers to implementing FA at scale. Finally, we explore the potential of LLMs to improve meaningful pedagogical outcomes within the FA framework, while explicitly identifying gaps in the current research landscape, proposing directions for future inquiry, and examining practical challenges associated with integrating LLM-driven formative assessments in real-world classroom settings, all considered through multiple perspectives.

To identify relevant literature for this narrative review, we conducted a comprehensive keyword-based search using Boolean operators. We performed searches across multiple academic databases, including Scopus, ERIC, PubMed, IEEE Xplore, and arXiv to ensure broad coverage of the literature. Key concepts include terms related to concepts such as different aspects of formative assessments (e.g., lesson planning, question generation, and feedback), large language models, and context in education (e.g., learning outcomes, barriers to using technology, and cultural context). The terms were then combined using Boolean operator such as AND/OR.

2. What Are Large Language Models?

2.1. From Artificial Intelligence to Large Language Models

Artificial Intelligence (AI) encompasses a wide range of technologies and systems designed to mimic intelligent human behaviors [35], including complex decision making and reasoning, sensory processes, memory, and language understanding [36]. Within this domain, Machine Learning (ML) refers to computer systems that adapt and optimize themselves based on what they have "learned" [37] from specific training datasets [38]. A significant breakthrough in ML has been the development of deep learning, which employs layered neural networks composed of artificial "neurons" inspired by the human brain to automatically detect patterns in data [39,40]. These networks operate by initially assigning random values to the connection strengths between neurons, where each "strength" represents the influence one neuron's output has on the next. The network then generates an output based on the input data and current connection strengths, compares this output to the correct answer provided in the training set, and adjusts the connection strengths to minimize the difference between its guess and the true answer. This adjustment process uses a mathematical optimization method known as gradient descent [41], which iteratively improves the network's accuracy on the desired task over time [42].

These neural network foundations led to the development of recurrent neural networks, which take in sequential data where order matters and have the ability to retain memory of past inputs [43], however; training may stop or become unstable as memory storage increases because data goes through many layers which amplifies the output either making it extremely large or small, making it hard to effectively update the "strength" connections [44]. To help mitigate this, the transformer architecture was developed [45], which involved processing parts of the input simultaneously rather than sequentially by the self-attention mechanism: breaking input sequences into small instances called tokens and weighing importance to each token before processing them [46], leading to the creation of LLMs.

LLMs are trained on vast textual datasets to generate and understand human language through Natural Language Processing (NLP) [47], and they do so by predicting the most likely token given surrounding input context tokens [48,49]. During their training process,

LLMs can capture not only grammar and syntax but also logical structure and meaningful semantic relationships [50,51], demonstrating their extensive NLP capabilities.

Here, a key development in 2018 was Google's large language model (LLM), Bidirectional Encoder Representations from Transformers (BERT), which significantly improved performance on tasks requiring deep contextual understanding by predicting tokens based on both left and right context [52]. BERT served as a foundation for Google's 2019 Text-to-Text Transfer Transformer (T5), which unified various natural language processing (NLP) tasks into a single text-to-text framework to streamline LLM processing [53]. In response, OpenAI developed the second version of their generative LLM in 2019, called Generative Pre-trained Transformer 2 (GPT-2) [54]. Unlike BERT, GPT-2 generated coherent, conversational paragraphs by predicting the next token sequentially from left to right. OpenAI further advanced this approach with GPT-3 in 2020, scaling the training data to extend the model's domain of knowledge [55]. In 2022, OpenAI released ChatGPT, a web-based AI chatbot built on GPT models, recognized for its versatility and conversational abilities [56]. These foundational LLMs have since inspired newer models such as GPT-4, Meta AI's Llama family, Google Gemini, and others, which continue to push the boundaries of reasoning and natural language processing. Recent LLMs have also incorporated multimodal architectures, enabling them to process images, audio, and limited video inputs alongside text, allowing these models to tackle more nuanced and complex tasks [57].

2.2. LLM "Learning" Techniques

LLMs can "learn" through many different ML training methods. One of the most common methods is unsupervised learning, where LLMs "learn" to predict the next token in a vast, unlabeled text dataset, as most language data is unstructured and unlabeled [47], enabling them to acquire broad language understanding. On the other hand, classical supervised learning is an ML method in which training data contains explicit human-labeled input–output pairs [58]. However, since LLMs need exposure to billions if not trillions of tokens to work effectively for NLP, this approach is purely theoretical for LLMs as creating that many explicit human-labeled input–output pairs would require unfeasible compute. To mitigate these concerns, semi-supervised learning was developed and offers a powerful, feasible alternative that combines a small amount of labeled input–output pair data with a large amount of unlabeled textual data for a larger training dataset [59]. However, there still exist many scenarios where there is little to no labeled data or where creating labeled data manually would be time-consuming. To solve this, the self-supervised learning technique was developed, which only requires unlabeled and unstructured data. Using this technique, LLMs generate their own pseudo-labels rather than relying on externally provided labels [60]. This technique can be applied in formative assessment (FA), for example, by combining a small set of human-graded examples with a large volume of unlabeled learner data to train an LLM, enabling more personalized feedback generation in a cost- and time-efficient manner.

Tangentially, reinforcement learning has also emerged as a powerful technique to optimize task objectives with human preferences, which involves a system where the LLM interacts with an environment and receives feedback in terms of rewards to promote desired behaviors [61,62]. In one study [63], LLMs were utilized to generate FA questions combined with a specific reinforcement learning technique that decreases the number of questions that need to be asked to learners to gauge their current knowledge levels. Furthermore, research has led to the technique: reinforcement-based learning from human feedback, incorporating human input to improve desired task performance [64].

After LLM training from these aforementioned techniques, LLMs can be further trained on a small task-specific dataset in a process called fine-tuning through supervised learning

(also called Instruction fine-tuning [65]), self-supervised or reinforcement learning, often requiring specific prompting techniques to optimize the model's output [66]. In evaluating how effectively LLMs can assess undergraduate students' answers [59], a fine-tuned model trained through supervised learning on student self-explanations, where students describe and explain each step of a task, outperformed various base LLMs in terms of accuracy [67].

2.3. Prompt Engineering in LLMs

Prompt engineering, which refers to the process of carefully designing and structuring input instructions to LLMs, has emerged as a crucial step to harness the full capabilities of LLMs. With effective prompt engineering, non-expert users can enhance the understandability, accuracy, and task alignment of LLMs' output without needing to change the LLMs' internal weights [68,69].

When LLMs are prompted to perform tasks that were not previously provided in the training dataset nor with any task-specific examples embedded into the prompt itself, it is defined as zero-shot prompting [70]. However, the simplicity of this prompting method has its trade-offs as it may not be ideal for heavily complex or domain-specific tasks. This is specifically seen when the model utilizing zero-shot learning performed poorly in an FA task for short answer grading [71]. Despite these shortcomings, model reasoning can be further enhanced in zero-shot learning with the integration of role-play prompting [72], which involves prompting an LLM to simulate and mimic a specific persona.

Building on this, task-specific examples can be added inside the prompt to guide the LLM's response, thereby increasing the amount of information the model has to better tailor the output response, which is called few-shot learning [73]. This process relies on in-context learning, which is the method that allows the LLMs to adapt to specific tasks based on information from user input prompts [55,74]. One study tested multimodal LLMs trained on different prompt engineering techniques on their ability to score and provide effective formative feedback on paper-based FA assignments [75], which resulted in the LLMs that utilized few-shot learning usually outperforming those that used zero-shot learning.

Another notable prompt engineering technique is chain of thought prompting, which involves providing the model with examples on how to reason step-by-step to guide the LLM to output its full reasoning process rather than only a direct answer, thereby improving capabilities of LLMs in complex reasoning tasks [76]. One study evaluated the automated scoring capabilities of LLMs on student written assignments [77], a key task in formative assessment where teachers commonly use writing to gauge student learning. In this study, GPT models were tested using various prompt engineering techniques. The model that incorporated chain of thought prompting along with the problem context and scoring rubric within the prompt outperformed those using zero-shot or few-shot learning, with or without chain of thought prompting.

In the original chain of thought prompting method, human-crafted step-by-step reasoning examples are included in every prompt, regardless of the task. However, because different tasks may require distinct reasoning strategies to produce optimal outputs, active prompting combined with chain of thought prompting was developed. In this technique [78], the LLM is presented with a range of task-specific questions to identify those it is most uncertain about. These selected questions are then manually annotated with step-by-step reasoning and used as part of the prompts for new questions, improving the model's performance. One study applied this method to formative assessment tasks such as automated grading and generating effective feedback, finding that active prompting combined with chain of thought prompting achieved the highest overall scores across multiple evaluation metrics, outperforming zero-shot, few-shot, and few-shot with chain of thought prompting approaches [79].

2.4. LLM Performance Benchmarks

Originally, tasks for language understanding often utilize standard evaluation metrics for NLP, either with text classification or text generation through token prediction. For classification tasks, accuracy, which measures how often the model is correct overall compared to the desired answers; precision, which measures how often positive predictions of belonging to a class are correct; and recall, which measures how much relevant tasks were completed regardless of being correct or incorrect [80], which are then combined into the F-1 score [81] to balance measure specific trade-offs. Thus, if we want to test how well LLMs can help in supporting different aspects of FA, we also need various metrics to evaluate the efficacy of such LLMs (Table 1).

In the first component of FA, where learners are in their current learning process, LLMs can assist with tasks such as automated scoring of student assignments, generating knowledge checks, and detecting errors or misconceptions. These tasks require robust benchmarking methods to evaluate how accurately and appropriately LLM outputs align with desired responses. For automated assignment scoring, effectiveness is typically assessed by measuring the correlation between LLM-generated scores and those given by human evaluators. One study used equivalence testing to determine whether LLM grading was statistically comparable to that of human evaluators [82], while another study utilized the Pearson correlation coefficient [83] to assess the closeness of LLM-generated scores to human-assigned ones [84].

The effectiveness of LLMs in generating knowledge checks can be evaluated based on how well the generated questions align with established cognitive frameworks. One study assessed question quality using Bloom's Taxonomy, a framework that categorizes learning objectives by specificity and complexity [85]. Another study evaluated question quality by comparing student quiz performance across two weeks—one week in which students received LLM-generated multiple-choice questions every 10 to 15 min to aid information recall, and another week without the LLM-generated questions [86]. In another FA task of error and misconception detection, one study was able to benchmark LLMs on identifying gaps in students' code explanations from expert-annotated code explanations by using four different similarity metrics [87]: chrF for character similarity [88], METEOR for word similarity [89], BERTScore for comparing tokens based on their meaning in the context of tokens beside them [90], and Universal Sentence Encoder for encoding sentences and comparing them based on semantic similarity [91]. Another study focused on improving SRL in students for grammatical accuracy used LLMs to identify incorrect reasoning or responses and was benchmarked using precision, recall, and F1-scores [92].

In addition to helping identify current learner understanding levels, LLMs can also help assist in educating learners on where they should aim to be, namely by generating and filtering learning goals. One study tested how well an LLM-integrated chatbot system could identify students' learning goals and plans based on previous interactions. Here, two human evaluators viewed the LLM response to see if it correctly detected the students' goals and plans, where Cohen's Kappa was utilized, which is the statistical measure for agreement between the evaluators [93,94]. Meanwhile, another study focused on using LLMs to assist in generating effective learning objectives, and here, the LLM output quality was evaluated by seeing how well the output aligned with Bloom's Taxonomy [95], which was also utilized in the first FA component. For roadmapping, which involves generating step-by-step mini goals, one study guided LLMs to generate personalized and pedagogically effective learning paths, which was evaluated by accuracy to learners' needs and goals, user satisfaction, and quality of learning paths in terms of how logical the roadmap is based on specific concepts [96].

Furthermore, the main component of FA is to help bridge this gap between where learners are currently in their learning process and where they are going. A Likert scale based on six defined effective feedback criteria has been used for human experts to evaluate LLM-driven feedback [97]. Likert scale has been used for experts to benchmark LLMs for FA in other scenarios, including the use of LLM in lesson planning [98,99]. Another study integrates deep learning and an LLM to provide personalized feedback in a course where one group receives weekly LLM generated feedback and a control group does not, and the performance difference on assessments between the groups was analyzed [100]. Furthermore, significance testing [101] and effect size [102] have been used to compare LLM-assisted gamified education system aimed at improving SRL against the traditional, non-AI learning system.

Table 1. List of LLM benchmarks for FA tasks.

FA Component	FA Task	Benchmarking Techniques
Where Learners Are in Their Current Learning Process	Automated Student Assignment Scoring	Equivalence Testing: [82] Pearson Coefficient: [83,84]
	Generating Knowledge Checks	Blooms Taxonomy Alignment: [85] With/Without Intervention Student Learning Outcomes Comparison: [86]
	Error/Misconception Detection	chrF, METEOR, BERTScore, Universal Sentence Encoder: [87–91] Precision, Recall, F1-Scores: [80,81,92]
Learners' Goals/Where they Should Be	Generating Goals/Learning Objectives	Cohen's Kappa: [93,94] Blooms Taxonomy Alignment: [95]
	Roadmapping	Accuracy, User Satisfaction, Quality of Learning Paths: [96]
How to Bridge Gap Between Current Knowledge and Goals	Providing personalized feedback	Expert Scoring through Likert Scales: [97] With/Without Intervention Student Learning Outcomes Comparison: [100]
	Improving SRL	Significance Testing and Effect Size: [101–103]
	Improving Teaching Instruction	Human Answered Likert Scale: [98,99]

Despite current LLM benchmarks' breadth covering a wide range of scenarios, they still have well-known gaps. Many benchmarks often rely on fixed "gold" answers or labels, but these can be unreliable as even one imperfect "gold" answer or label could cause a major concern [80]. This poses a particular challenge for effective FA implementation that relies on complex, individualized learning. In FA scenarios, student responses likely have a mix of partially correct and incorrect parts, where rigid "gold" answers or labels may provide incorrect or unfair feedback rather than the goal of effective FA: to guide learning.

Furthermore, many current LLM and natural language generation benchmarks rely heavily on quantitative correlation-based metrics like Pearson coefficients [83] or expert agreement scores, which mainly assess alignment with desired human outputs but may not truly reflect pedagogical effectiveness or meaningfulness [3,104]. Finally, the absence of standardized benchmark datasets makes it extremely difficult to compare results across

studies or platforms and thus decreases reproducibility, making it important for more holistic benchmarking frameworks to be proposed and implemented in the future.

3. Formative Assessment Applications with Large Language Models

Although barriers to implementing effective FA exist, LLMs offer promising solutions to mitigate them. Through automating some redundant parts of the FA process [105,106], LLMs can increase scalability and effectiveness, thereby transforming traditional FA practices by overcoming these challenges.

3.1. BERT

The Bidirectional Encoder Representations from Transformers (BERT) model, developed by Google in 2018, was one of the first large language models (LLMs) to effectively capture relationships between words across different contexts [107]. Its ability to model bidirectional word dependencies made it well-suited for tasks such as short-answer grading. When combined with transfer learning, the practice of adapting pre-trained models to specific tasks improved efficiency and reduced the need for large, annotated datasets [108], allowing BERT to achieve performance that surpassed previous LLM benchmarks in short-answer grading [109]. Notably, the model outperformed human evaluators when trained on more than 45–50% of the total dataset, marking a significant advancement in the possibility of integrating LLMs into formative assessment. By automatically analyzing trends in student responses, such models can help educators identify areas where learners struggle, reducing the time burden traditionally associated with manual grading.

BERT has also been applied to classify aspects of student teamwork in online team chats, identifying elements such as coordination, mutual performance monitoring, constructive conflict, and team emotional support [110]. Compared to a previously effective model, BERT demonstrated fewer classification errors, identified more relevant teamwork messages, and showed greater generalizability to unseen data. This application is particularly valuable in the context of online group-based formative assessments, where student interaction tends to be more limited than in face-to-face settings. By analyzing the quality of student interactions and collaborative contributions, LLMs can help educators evaluate the effectiveness of group-based versus individual assessment approaches. Additionally, LLMs can be used to automate feedback on assignment-specific criteria and recommend targeted teamwork skills for improvement, further reducing the time required for analyzing formative assessment outcomes.

3.2. Longformer

Evolving from BERT, the Longformer Transformer model was introduced in 2020 to address the need for processing longer sequences of text more efficiently [111]. This model has been applied in automatic summary evaluation, where it demonstrated a low average deviation, only 13%, from expert-assigned scores when assessing summaries of a passage's main idea [112]. Unlike short-answer grading, summary evaluation involves analyzing entire texts, yet LLMs such as Longformer still show promise in providing detailed and personalized feedback. This supports students in achieving learning goals and enhances formative assessment practices.

Beyond summary evaluation, the Longformer has also been used for automated assessment of instructional classroom discussions based on raw text transcripts. Quality-assessment models that employed data labeling techniques to segment instructional dialog into specific pedagogical strategies, or "talk moves," outperformed Longformer in this context [113]. These talk moves, such as "Strong Text-based Evidence," "Press," and "Recap or Synthesize Ideas", are used by instructors to promote student engagement and deeper

understanding. Integrating LLMs with such classification schemes can provide educators with valuable feedback on the effectiveness of their instructional strategies. Additionally, LLMs can help identify inaccuracies or missed opportunities within teacher discourse and suggest formative assessment techniques aligned with the subject matter, streamlining instructional planning and feedback delivery.

3.3. ChatGPT

Integrating elements from its predecessors, OpenAI created Chat-Generative Pre-Trained Transformer (ChatGPT), a user-friendly chatbot hosted online that was fine-tuned to effectively handle dialogs. ChatGPT has been applied to the education sphere, where GPT-3.5 was utilized to improve automated evaluation of student text responses with limited datasets through data augmentation, artificially increasing training diversity size [114]. Specifically, prompt-based text data augmentation was used with GPT-3.5 to fine-tune an instance of a BERT model, where in a majority of test cases, the BERT model trained on artificially generated responses outperformed a LLM augmenting data by itself. This demonstrates that to an extent, a small dataset can be valuable when augmented to optimize performance. Automated evaluation of student text responses is also a component of FA, where the notion that only a small dataset is needed to optimize performance means that for simpler FA, only a small number of rubrics or examples need to be utilized in prompting, limiting the amount of time spent on creating FA.

GPT-4 builds upon the foundation of GPT-3.5 with the addition of multimodal capabilities, enabling it to process and integrate information across text and images. Compared to GPT-3.5, GPT-4 showed a stronger correlation with human expert scores when automating quantitative feedback in a dataset of 243 economics exams [115]. It has also been applied to automate qualitative evaluation feedback in open-ended science formative assessments [79]. In this context, a few-shot learning approach was used, where the model was trained on a small dataset and employed Chain-of-Thought reasoning to generate intermediate explanations while solving problems. A key feature of this application was the integration of a human-in-the-loop strategy to support Active Learning. This process involved identifying cases where the model made incorrect predictions, after which humans reintroduced these examples using Chain-of-Thought prompting to help fine-tune the model. A disagreement detection method was used to identify mismatches between model and human outputs, guiding further refinement. The study compared several approaches, including zero-shot, few-shot, few-shot with Chain-of-Thought, and the combined Chain-of-Thought plus Active Learning method. The final approach proved the most effective in automating qualitative evaluations. The ability of GPT-4 to generate both quantitative and qualitative feedback has significant implications for formative assessment. It enables scalable, personalized, and actionable insights for learners while reducing the workload required from educators.

3.4. Google Gemini

In response to the rapid proliferation of large language models, Google developed Gemini as an evolution of the BERT architecture. This model, alongside ChatGPT, was used to generate 18 lesson plans for teachers, which were then evaluated using subject-specific learning theories [116]. Analysis of the LLM-generated outputs showed that the learning objectives identified by the models closely aligned with those created by human educators, indicating that AI-generated lesson plans are suitable for classroom integration. Although both models produced effective content, differences were observed in their approaches. Google Gemini tended to include assessment types such as quizzes, while ChatGPT emphasized group discussions. Gemini also provided differentiated instruction by including

content sections tailored for students who were ahead or behind in their learning, whereas ChatGPT focused more on outlining teacher responsibilities and strategies to motivate student engagement. These complementary strengths suggest that LLMs can greatly reduce the time required for curriculum planning in formative assessment. By leveraging the unique capabilities of different models, educators can develop more comprehensive and efficient lesson plans that better address diverse learning needs and clarify complex concepts for students.

3.5. LLaMA

Extended writing assistant capabilities have also been demonstrated by Meta's Large Language Model Meta AI (LLaMA). In a comparison between the open-source LLaMA 2 model and GPT-3.5, both were evaluated on their ability to provide feedback aimed at improving students' argumentation and writing skills [117]. Out of 63 participants, 50 were able to understand the recommendations provided by the models and how to implement them. The perceived quality of feedback, rated on a scale from 1 to 10, averaged 7.51 for LLaMA 2 and 7.65 for GPT-3.5, indicating both models were well-received by learners. Further evaluation of LLaMA's capabilities was conducted using LLaMA-34b-Instruct-hf, a fine-tuned version of LLaMA 2, which was compared to GPT-4 in providing individualized formative feedback for coding-related tasks, such as identifying and fixing bugs [118]. Results showed that GPT-4 successfully resolved 94.33 percent of bugs across 106 programs, whereas the LLaMA-based model achieved a success rate of 66.03 percent. These findings demonstrate the growing potential of LLMs in supporting formative assessment by offering personalized feedback across both writing and technical domains.

3.6. Other Fine-Tuned and Specialized LLMs

Fine-tuned hybrid LLMs have also been applied in FA, including the multimodal Macaw model, which integrates three modules for encoding multimodal data, leveraging LLMs, and unifying multimodal inputs for coherent output generation [119]. This model was fine-tuned to generate multiple-choice questions resembling those found in college-level science textbooks, aiming to evaluate whether the quality of AI-generated questions could match or surpass that of human-created ones. When compared to a zero-shot model, BootIt Next Generation (BING) chat, the results showed that in six out of seven cases, the quality of questions produced by both LLMs was not significantly different from those generated by humans [120]. These findings suggest that hybrid LLMs hold promise for efficiently generating high-quality educational materials within the formative assessment framework.

3.7. FLAN-T5

After BERT, Google developed the Text-To-Text Transfer Transformer (T5) [53], where its ability to be fine-tuned allowed Google to develop it to the Fine-tuned Language Net (FLAN)-T5 [121]. In one study, a FLAN-T5 model was compared to LLaMa 2 and GPT models in their ability to assess how correct students' explanations were in undergraduate computer science classes, where FLAN-T5 had the higher accuracy [67]. Different FLAN-T5 sized models have also been integrated in a virtual tutor for masters' students to help answer questions on fermentation [122]. This model used a zero-shot, in-context learning approach, and utilized a specific semantic searching technique to better understand the user query by prompting the LLM with external context. The results of this study showcase that for each size of the FLAN-F5 model, the output of the LLM is similar to the original answers, where the FLAN-F5 Base model has the highest similarity.

3.8. Mistral-7B with Parameter-Efficient Fine-Tuning

Another important fine-tuning approach is parameter-efficient fine-tuning, which optimizes only a small subset of a model's parameters while keeping the majority fixed, resulting in more efficient training and deployment [123]. This method has been applied to enhance LLM performance in answering textbook questions by incorporating supplementary instructional materials and tailoring the models to specific academic disciplines [124]. Combined with retrieval-augmented generation, a technique that retrieves relevant context to inform the model's responses, this approach was shown to improve output quality. When applied to True/False and Matching questions, the integration of parameter-efficient fine-tuning with retrieval-augmented generation achieved a question-answering accuracy of 86.34 percent, outperforming the 84.42 percent accuracy achieved using retrieval-augmented generation alone [124]. Perhaps, efficient fine-tuning strategies could improve the accuracy and relevance of LLM-generated responses in formative assessment contexts.

4. Opportunities for Innovative LLM Implementations in FA

4.1. Storytelling and LLM

Storytelling, defined as the use of narrative to engage an audience, plays a crucial role in fostering human connection and enhancing communication [125]. With recent advancements in large language models (LLMs), these systems have demonstrated strong performance in summary evaluation tasks, often producing outputs that closely resemble those of human writers, although they may still show some limitations in tasks requiring nuanced sentence-level explanation [126]. The narrative generation capabilities of LLMs have been further enhanced by integrating them with symbolic planning systems, which rely on explicitly defined instructions to guide content generation [127]. This integration helps reduce disjointed or repetitive information in the stories produced [128]. One such implementation, TattleTale, combined automated planning with LLMs to guide narrative generation. The system ensured that the nouns and verbs present in the final output closely aligned with the elements specified in the plan, resulting in more coherent and goal-directed narratives [129].

These narrative generation techniques hold strong potential for application in formative assessment. They can be used to evaluate students' comprehension, creativity, and ability to apply learned knowledge in real-world contexts. By embedding LLMs within interactive storytelling environments, students can be presented with immersive scenarios that require the integration of previously acquired concepts. As learners engage with these narratives, LLMs can provide progressive feedback loops in the form of formative nudges such as hints or gentle corrections that guide students toward their learning goals without constraining their creative expression. This approach offers a dynamic and engaging method of reinforcing understanding while maintaining the motivational aspects of storytelling. However, the effectiveness of current storytelling LLM systems in large-scale classroom settings has largely been left unexplored, where future research should experiment on how these system interventions affect subjects' learning across a diverse sample.

4.2. LLM-Integrated Intelligent Tutoring Systems

Previously, Intelligent Tutoring Systems (ITS), which are computerized platforms designed to deliver adaptive, real-time formative feedback, have been used to reduce teacher workload and enhance instructional quality by offering programmed pedagogical interventions [130]. Studies have shown no significant difference in the effectiveness of ITS compared to human tutoring, demonstrating their reliability in supporting student

learning [123,124]. More recently, the integration of artificial intelligence frameworks into ITS has further improved student engagement, learning outcomes, and performance. These advancements have been supported by techniques such as performance prediction, learning behavior analysis, and enhancements in SRL [131]. Emerging research has also proposed methods to optimize feedback generation in LLM-powered ITS through structured prompt design and empirical evaluation of their impact on learning [126].

Building on these developments, we propose that these feedback-generation techniques can be effectively applied to LLMs within concept-check-based formative assessment. LLMs can process large volumes of student-specific data to rapidly generate tailored feedback. This process may begin with the LLM posing diagnostic questions to identify knowledge gaps or misconceptions. Once the student responds, the system can assess understanding and clarify earlier errors through targeted follow-up questions. To increase efficiency, a pipeline can be implemented where broad diagnostic responses are automatically fed into the LLM, which then pre-generates context-specific explanations. By the time the student answers the follow-up, the system is prepared to provide immediate, personalized feedback. Leveraging the multimodal capabilities of LLMs, such as accepting both text input and uploaded student work, can allow the ITS to better understand a student's thought process and detect inefficiencies in reasoning.

Additionally, LLMs can enhance formative assessment by integrating SRL principles to support deeper learning [132]. These principles can be combined with the Chain-of-Thought and Active Learning approach, as demonstrated in GPT-4 [79], to help students explicitly map conceptual connections and reflect on their familiarity with subject content. Such strategies support mental mapping within FA, allowing students to identify areas needing further attention. Educators can also use this documentation to improve pedagogical strategies by understanding how students connect and apply key concepts. The potential of ITS-powered LLMs extends further when student feedback is contextualized using their individual interests [133]. Personalizing formative assessments in this way promotes engagement and helps the LLM better interpret the student's cognitive patterns to generate more meaningful feedback. Future research should explore how integrating contextual interests with historical learning data and current performance can support goal setting. LLMs can assist by generating templates for clear, long-term goals that are broken down into manageable, short-term objectives in a cyclical and adaptive format, minimizing the time required for students to develop actionable plans. Finally, the goal-setting approach supported by LLMs can also be applied to teachers' content planning for formative assessment. Drawing from Baytak's findings on LLM integration into lesson plans [116], we propose that LLMs be prompted with instructional content, teaching strategies, and student outcomes to offer recommendations for more effective formative assessment practices. This would allow educators to adapt their teaching based on data-driven insights, further enhancing the overall learning process. Despite these promising applications, further empirical research is needed on combining difference sources of information on the student to optimize adaptive learning as well as short-term and long-term goal attainment.

4.3. Collaborative Learning and Dynamic Learning

A recent direction in state-of-the-art LLM research has focused on the advanced language understanding and generation capabilities of LLMs to support group discussion and collaborative learning [134]. LLM-powered chatbots have shown promise in engaging students and enhancing classroom debate discussions by posing thought-provoking questions and offering relevant comments [135]. These contributions help scaffold student learning within the context of formative feedback by guiding them from their current understanding toward more advanced conceptual knowledge.

The integration of LLMs into collaborative learning has progressed further through systems that adopt a devil's advocate role, where the LLM critiques specific parts of student discussions to encourage higher-quality discussion [136]. Despite these advances, LLMs still face challenges in determining appropriate moments to contribute in group settings, as most have been fine-tuned for one-on-one interactions rather than group conversations. To address this, researchers have explored advanced prompting strategies that involve multiple LLM agents, with the most coherent and relevant discussion topics being selected for continuation [137]. This approach opens new possibilities for collaborative problem-solving and feedback in LLM-supported virtual learning environments [138].

Furthermore, within the context of collaborative learning, role-play prompting technique allows LLMs to portray and embody not only specific human characteristics [139] but also systems to help simulate complex interactions, behaviors and reasoning [72]. Being able to adapt a specific personality trait [140], such as in gamifying learning, can improve student engagement [141] and academic outcomes [142,143].

Combining recent advances in storytelling, ITS, role-play prompting techniques, we could imagine creating an LLM-driven interactive learning environment for FA. In this approach, students engage in a story-based scenario where each participant adopts a different perspective or role within a topic. The ITS constructs a branching decision tree that guides students through scenario-based decision making, presenting highly specific questions tailored to each student's conceptual understanding. Students progress through the tree by selecting options and answering questions, advancing when correct and receiving feedback when incorrect. After each interaction, roles are randomized to ensure that all students engage with multiple perspectives and demonstrate comprehensive understanding of the material. Over time, by analyzing students' learning habits, strengths, and areas for improvement, the ITS can automatically adjust the difficulty of questions and tailor feedback to optimize learning outcomes. This integration of adaptive storytelling, role-play, and real-time formative assessment has the potential to transform how students engage with and master complex concepts.

LLMs can also be applied in different peer-assessments, especially group discussions. Previous research found the integration of interactive LLM-powered devil advocates for group decision making, which challenges or questions decisions and promotes higher-quality discussion [136]. It may be possible that this devil's advocate system can also be combined with an opposite system that helps continue the current subject matter flow. Here, LLMs can serve to generate thought-provoking, engaging prompts that reflect specific topics of concern, while also moderating and guiding the conversations through suggesting transitions and gently redirecting the conversations if they are off-topic. LLMs can also transcribe and analyze the group discussions to find areas that have similar opinions or differing ones that reflect groupthink, to foster in-depth discussion. LLMs can also be utilized in the peer-grading process, where models can help summarize peer feedback from FA to make it easier for students to identify areas of improvement, thereby helping the students better understand how they can reach their learning goals. In addition, LLMs can also help students improve upon the peer-feedback they give by providing consistency and quality checks to ensure that it is constructive, toning down subject language that may be viewed harshly. LLMs can also help provide real-time translations for peer-grading comments between different languages, promoting the accessibility of quality formative feedback for non-native speakers. Current LLM-based systems face challenges in managing multi-agent interactions. Future research should explore robust evaluation techniques for LLM-facilitated collaborative learning.

4.4. Game-Based Learning Systems

Game-based learning is an approach that uses digital and in-person game play to help facilitate learning and achieve defined learning outcomes while increasing students' engagement usually through an incentive-based system [144]. Many different games have been utilized over the years in game-based learning systems, and they also provide highly personalized experiences due to students being able to learn at their own pace and learning style [145].

LLMs have recently emerged as an integral part in the evolution of game-based learning by connecting personalized, adaptive agents with a captivating and focused environment. One significant advancement is the use of adaptive LLM-based student goal and planning generation seen within the environment *Crystal Island* [146], thereby assisting in the SRL and FA framework for learners to more efficiently create effective learning goals and proactive steps to achieve them. Here, LLMs can generate an eagerness to learn through playful game-based learning, ultimately leveraging the full extent of capabilities of LLMs in education [147].

Many studies have previously connected the improvement of student performance and engagement with competition-based learning, which is a student-centered approach focused on problem-solving and other competition-centered approaches with peers [148,149]. Previous research on competition-based learning has then been extended to game-based environments [150,151] to further increase student motivation and performance, which is an integral part of FA for continuous feedback loops that emphasize the learning process. However, creating these game-based learning systems might not always be viable as it might take an extremely long amount of time to build and educators might not have the skills or experience to create systems that are pedagogically effective and engaging. In these competition-based and game-based environments, future research should integrate LLMs to generate these systems efficiently, while following the FA framework with tailored and actionable feedback for students to help them better understand their current knowledge and how to reach their learning goals by prompting the LLMs with rubrics on knowledge mastery criteria, common student misconceptions, and typical learning development. However, designing and implementing LLM-powered game-based learning environments still remains highly resource-intensive, where future research should delve into how to decrease time needed for program development.

4.5. Virtual Reality and Multi-Agent Classrooms

Advancements in computer graphics and artificial intelligence have enabled the development of virtual reality, a technology that allows users to simulate hyper-realistic scenarios in a new digital environment in which the user can interact with [152]. Recently, however, the development of state-of-the-art technologies has led to the creation of improved virtual reality systems that are now being integrated in higher education [153]. These systems offer students opportunities for learning with active engagement, especially with learning abstract concepts [154], which is integral to better help learners move forward in the FA framework.

In parallel, multi-agent classrooms refer to digital environments where multiple autonomous agents [155], including humans and AI-powered technology, are utilized to collaborate and assist in the student learning process. LLMs have recently been incorporated as agents into these classrooms [156], where they effectively simulated dynamic learning scenarios and enhanced learning outcomes through increased collaboration, thereby connecting to the FA framework by assisting the learners in bridging the gap of reaching their goals with personalized LLM agent guidance.

Moreover recently, one study has delved into LLM integration with multi-agents specifically in ITS [157], connecting to the FA framework by first identifying the gap between the learners' goals and what skills are required, then creating a personalized plan to reach the learners' goals based on learners' profile, and finally providing dynamic feedback to the student all with the benefit of being supported through multiple LLM agents where subtasks are distributed between collaborating agents to optimize efficiency. This specific distributed agent approach is efficient but can be further enhanced where future research should apply distributed computing with parallel processing on multiple machines to enable faster real-time decision making that enhances the FA process with more efficient dynamic feedback, thereby improving scalability for applications with large computational demand [158].

In addition, research has also investigated virtual reality systems used in a game-based learning platform integrated with LLMs [159], which provides a more immersive and interactive environment as LLM-integrated characters, and their dialog is dynamically changed based on fictional simulated experiences. Furthermore, research has also delved into how these systems can be combined with the multimodal capabilities of LLMs taking in facial features or movements to understand and subsequently simulate human emotions and personality to increase student engagement [160]. This connects to the FA framework as an engaging way to receive personalized feedback to help learners move forward in learning process as the dialog between characters could be individualized based on the specific goals and the current knowledge of the student input from different FA, where future research should potentially explore which LLMs are best suited for this as well as how to optimize it for this pedagogical task if fine-tuning is required. Current VR and multi-agent LLM systems have not been widely testes across diverse educational contexts and future research should focus on testing the scalability of this implementation.

5. Challenges of Integrating LLMs in FA

5.1. Standards and Educational Technology Procurement

Educational Technology (EdTech) is a vehicle to deliver effective FA and can be defined as any type of technology that can be combined with teaching to foster and enhance learning, including improving problem-solving skills [161], even if the agents in the learning process are unaware of the technologies' purpose [162].

With a rapid increase in EdTech integration and EdTech procurement, different institutions have varying strategies in the process of selecting, obtaining, and implementing EdTech. Thus, it becomes helpful to create specific standards for procuring LLM-integrated EdTech such that institutions know who and when to trust the EdTech in question, particularly when there are novel LLM-driven EdTech frequently coming out into the market [163,164]. Although even when creating standards is seen as important, formally certifying them is not seen as a top priority by educational institutions. However, there are several way to address the challenge of procuring the right technology: involving teachers and other practitioners in procurement, knowing exactly what you need from assessments and finding meaningful product effectiveness evidence, creating a national LLM-driven EdTech product information website [165].

Although LLM in FA EdTech integration has the potential to transform learning processes to optimize learning effectiveness, there still exist many current challenges and barriers to institutional adoption that must be solved first. One challenge is a type of "technical" bias from constraints of software or hardware design that cause inherent inequalities in EdTech integration based on structural digital divides [166], such as gender disparities in digital access, freedom, and literacy [167]. Furthermore, the lack of clarity on how EdTech could foster meaningful pedagogical outcomes makes instructors more

reluctant to experiment with new technologies [168], and this barrier is particularly relevant with the proliferation of new LLM applications for FA that often lack evaluation metrics relevant to pedagogical outcomes of FA [3].

Another concern for educators integrating LLM in evaluating student's work is privacy and security standards. Poor EdTech security practices have led to large-scale data breaches which could have been easily prevented [169]. While this concern generally applies to cloud-based technology, it is particularly more relevant to LLM-driven products due to lots of questions surrounding the use of students' materials to train and improve the models. Related to this, another LLM-specific challenges include security vulnerabilities such as prompt-based attacks, where one can manipulate input prompts to output malicious information and sensitive information from the training data (e.g., personally identifiable information) [170,171]. These concerns make sense since student's work will likely include personally identifiable data, and if there are no privacy and security standards for LLM-integrated FA platforms to comply, malicious actors may be able to obtain student's sensitive information due to preventable LLM vulnerabilities [172,173].

Moreover, LLMs can also generate inaccuracies based on hallucinations [174], which are output information that sounds plausible but are indeed not factual. Hallucinating incorrect feedback can have negative effects on student's learning. Furthermore, LLMs can also generate toxic language like hate speech and insults [49], which may elicit negative emotions in students and hinder learning. Thus, it is important that educators, institutions, and EdTech providers work together to create a safe environment for students to learn and receive formative feedback.

5.2. Costs and Affordability

Could LLM make implementing FA more affordable? One research study identified cost factors as a limitation in implementing personalized learning in AI-integrated FA platforms [175], which is especially prevalent in low–middle income countries [176]. Furthermore, in developing countries, the cost-effectiveness of the traditional EdTech intervention of access to technology is extremely low with poor effectiveness and expensive to scale [177], where also for AI-integrated EdTech, maintaining and upgrading existing technologies can pose a financial burden in these countries [178], further increasing the gap between institutions that have sufficient resources and those who do not.

However, while LLM-integrated FA platforms may be costly, we believe that educators can experiment with the free versions of multi-purpose LLM such as ChatGPT to enhance various components of FA, from lesson planning to generating MCQs and giving feedback [3]. The important caveat here is that it relies on the educator to understand the pedagogical principles of what it means to implement FA in their classrooms and what pedagogical outcomes they are trying to achieve. As shown in our review, LLM can automate many tasks relevant to FA, making it very cost-effective for instructors.

For those wishing to build a customized LLM, recent studies have highlighted a large trade-off between using base LLMs and fine-tuning them. Although fine-tuning an LLM requires more initial compute cost than simple prompt engineering techniques, it is still far less expensive than training a model from scratch. Full-training of an LLM from scratch may require millions of compute hours, but fine-tuning is much less computationally intensive and only requires 1–2% of full-training from scratch or even less with optimized techniques [179]. Furthermore, fine-tuning can reduce model size, making it cheaper to run the model operationally [180], but the initial cost of fine-tuning may act as a deterrent for organizations who may opt for open-source LLMs instead.

5.3. Cultural Barriers to FA

Recent research demonstrates that teacher perceptions of AI in educational technology (EdTech) are generally slightly positive, as educators recognize its potential benefits, though attitudes remain mixed due to notable concerns. In the context of higher education, analysis of 99 studies revealed that a significant portion reported mixed perceptions among both students and teachers, highlighting ongoing ambivalence toward AI integration [181]. Similar patterns emerged in a survey of Saudi teachers, where AI-assisted tools were viewed as time-saving and useful for lesson planning and individualized instruction, yet concerns persisted around training demands, potential errors, and fears of job displacement [182]. Quantitative studies reinforce this trend, showing that while many teachers view AI positively, their confidence in engaging with these tools is closely linked to prior experience and the availability of institutional support [183].

Teachers' willingness to adopt AI is also shaped by how they balance perceived benefits against concerns. Across six countries, higher levels of AI knowledge and self-efficacy were strongly associated with greater perceived benefits, fewer concerns, and increased trust in AI tools [25]. Among K–12 educators, AI-specific knowledge emerged as a key predictor of trust, with findings suggesting that general digital skills alone are insufficient for fostering the confidence needed for effective AI integration [184]. Further work in this area has led to the development of survey instruments designed to measure trust in AI for EdTech use, identifying factors such as anxiety and self-efficacy as critical contributors [185]. Importantly, this mistrust can be mitigated through targeted professional development programs that explain how AI systems make decisions compared to humans and demonstrate how AI can support and enhance, rather than replace, the role of teachers [186].

Attitudes toward AI in educational technology are also significantly influenced by cultural norms, which shape both expectations around FA and the perceived role of technology in learning. Research shows that FA practices vary across cultural contexts. In Tanzanian schools, for example, Western FA ideals—such as students asking questions and engaging in group discussions—clashed with local conceptions of a "good student," which emphasized quiet listening and deference to the teacher's authority [24]. Similar cultural dynamics are reflected in studies on trust in AI for education, where significant differences emerged based on geography and cultural norms, rather than demographic variables such as age or gender [25].

Cultural orientation toward student-centered versus teacher-centered education also affects perceptions of AI in EdTech and FA. In the Czech Republic, most students were familiar with AI tools, primarily using them for information processing and comprehension [187]. These students cited time efficiency and accessibility as key advantages, while concerns included credibility and potential misuse [187]. However, student reliance on teacher assessments remains strong, especially in applied or creative disciplines that are more subjective in nature. In these cases, students viewed human feedback as more nuanced and trustworthy than AI-generated formative feedback [188].

This tension is further compounded by the nature of FA itself, which is inherently student-centered, emphasizing active participation, feedback loops, and self-reflection. These principles can conflict with traditional teacher-centered systems that prioritize rote memorization, summative assessments, and lecture-based instruction. In such environments, AI tools may be perceived as unnecessary or even disruptive to existing pedagogical norms [3]. As a result, successful integration of AI into FA may depend not only on technological capabilities but also on the alignment between AI-supported methods and the cultural values embedded within educational systems.

5.4. Keeping Pace with Rapid AI Advancements

One practical challenge with LLMs is simply keeping pace with its technological advancements, with more than 7000 publications related to LLMs in 2024, almost doubling the number of publications from 2023 [28]. Furthermore, LLM research is focusing on larger parameter sizes for training and architectural complexity [189], meaning each new model is more capable and complex than its previous versions. Moreover, within different LLMs, there is vast amounts of emerging research on input prompting techniques that educators can utilize to assist in the FA framework like with goal-setting and dynamic feedback to better help the learner understand how to reach their goal [68,190]. This is possible because LLMs' outputs are highly sensitive to changes in prompt inputs [191].

This rapid increase in LLM complexity and its research means that educational content and teacher training may need to be continually updated to leverage the latest models and prompting techniques. Here, educators may face challenges of learning about new LLM features and effective integration into their FA teaching practices, but this could be mitigated through the creation of professional development programs with monthly meetings between teachers and AI experts providing case studies on new LLM technology integration, albeit at the cost of heavy research funding and effort.

Before specific LLM-integrated EdTech tools can be procured and implemented in educational institutions, the LLMs must undergo rigorous quality and relevance testing to determine their suitability for supporting FA frameworks, as discussed in the section on standards in EdTech procurement. However, existing studies have identified significant limitations in these evaluation processes, particularly for text-based assessments, where issues with reproducibility, reliability, and consistent performance across diverse tasks and inputs have been noted [192]. Addressing these challenges requires robust evaluation protocols that incorporate standardized benchmarks and multiple trials to ensure reproducibility. For example, a pharmacy exam simulation study demonstrated that while GPT-4 achieved the highest accuracy, Llama-2-13B showed the lowest variance in answers, with all models performing better on knowledge recall than on application questions [193]. This finding underscores the need for LLM benchmarks to include higher-order reasoning problems, which align with FA's emphasis on active and metacognitive learning.

Furthermore, the criteria used to evaluate LLM quality must reflect pedagogical objectives. One study implemented automatic evaluation of educational assessments by LLMs using criteria such as avoiding incorrect statements, not revealing correct answers, providing guidance when students are stuck, identifying misunderstandings, and maintaining an encouraging tone [194]. These are elements integral to dynamic FA feedback. Despite these advances, inconsistencies in evaluation outcomes across different models indicate that further research is necessary to develop LLM-based assessment evaluation agents that are consistently aligned with students' learning goals.

6. Conclusions

This review highlights the transformative potential of LLMs in enhancing FA through applications such as storytelling, ITS, collaborative learning, game-based learning, and virtual reality-based multi-agent classrooms. LLMs can support personalized feedback, adaptive goal setting, and student engagement, while also assisting educators in instructional planning and assessment design. However, several challenges for immediate adoption still remain. These include high costs and resource constraints, cultural and pedagogical barriers within teacher-centered educational systems, privacy and security concerns, technical biases, and the rapid pace of LLM advancement. Continuous teacher training and evaluation methods might also be needed for rapidly evolving LLMs. Future research should empirically evaluate the effectiveness of LLM-based FA interventions across diverse educa-

tional contexts, explore methods to integrate multimodal data and context-based student interests, optimize collaborative and multi-agent environments, and develop standardized evaluation metrics that align with pedagogical goals.

With the proliferation of new LLM research on models and prompting techniques, teachers will need to adapt and experiment with general purpose LLMs such as ChatGPT and Google Gemini. By understanding the pedagogical principles of formative assessment and recognizing how LLMs can support its three core components, which include supporting "where learners are going," "where learners currently are," and "how to move learners forward" in the learning process, educators can effectively leverage these tools to promote meaningful educational outcomes.

Author Contributions: C.N.: Conceptualization, Methodology, Formal analysis, Investigation, Writing—Original Draft, Writing—Review and Editing, Visualization; S.J.: Conceptualization, Methodology, Formal analysis, Investigation, Writing—Review and Editing; S.P.: Conceptualization, Methodology, Formal analysis, Investigation, Resources, Data Curation, Writing—Original Draft, Writing—Review and Editing, Visualization, Supervision, Project administration, Funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: Sapolnach Prompiengchai is supported by the Rhodes Scholarship and the Warden Discretionary Grant.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: We thank the members of the Clematis Research Empowerment Hub for their support and advice to the authors in this review.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Black, P.; Wiliam, D. Developing the Theory of Formative Assessment. *Educ. Assess. Eval. Account.* **2009**, *21*, 5–31. [CrossRef]
2. Black, P.; Wiliam, D. Assessment and Classroom Learning. *Assess. Educ. Princ. Policy Pract.* **1998**, *5*, 7–74. [CrossRef]
3. Prompiengchai, S.; Narreddy, C.; Joordens, S. A Practical Guide for Supporting Formative Assessment and Feedback Using Generative AI. *arXiv* **2025**, arXiv:2505.23405. [CrossRef]
4. Panadero, E.; Andrade, H.; Brookhart, S. Fusing Self-Regulated Learning and Formative Assessment: A Roadmap of Where We Are, How We Got Here, and Where We Are Going. *Aust. Educ. Res.* **2018**, *45*, 13–31. [CrossRef]
5. Zimmerman, B.J. Investigating Self-Regulation and Motivation: Historical Background, Methodological Developments, and Future Prospects. *Am. Educ. Res. J.* **2008**, *45*, 166–183. [CrossRef]
6. Brenner, C.A. Self-Regulated Learning, Self-Determination Theory and Teacher Candidates' Development of Competency-Based Teaching Practices. *Smart Learn. Environ.* **2022**, *9*, 3. [CrossRef]
7. Perry, N.E. Young Children's Self-Regulated Learning and Contexts That Support It. *J. Educ. Psychol.* **1998**, *90*, 715–729. [CrossRef]
8. Ozan, C.; Kıncal, R. The Effects of Formative Assessment on Academic Achievement, Attitudes toward the Lesson, and Self-Regulation Skills. *Educ. Sci. Theory Pract.* **2018**, *18*, 85–118. [CrossRef]
9. Fowler, K.; Windschitl, M.; Richards, J. Exit Tickets. *Sci. Teach.* **2019**, *86*, 18–26. [CrossRef]
10. Shehzad, U.; Recker, M.; Clarke-Midura, J. Exploring the Potential of Exit Tickets as Formative Assessments of Student Affect. *Assess. Educ. Princ. Policy Pract.* **2025**, *32*, 173–191. [CrossRef]
11. Pan, Y.; Wang, L.; Zhu, Y. Strategic Questioning for Formative Assessment in TEFL: Insights from Blended Synchronous Learning Environments. *Humanit. Soc. Sci. Commun.* **2024**, *11*, 1519. [CrossRef]
12. Topping, K. Peer Assessment Between Students in Colleges and Universities. Available online: https://journals.sagepub.com/doi/epdf/10.3102/00346543068003249 (accessed on 18 July 2025).
13. Parmigiani, D.; Nicchia, E.; Murgia, E.; Ingersoll, M. Formative Assessment in Higher Education: An Exploratory Study within Programs for Professionals in Education. *Front. Educ.* **2024**, *9*, 1366215. [CrossRef]
14. Prompiengchai, S.; Baby, N.K.; Joordens, S. Bringing Learner-Centered Online Peer Assessment and Feedback to Indian and Canadian High Schools: Initial Reactions from Teachers and Students. *Soc. Sci. Humanit. Open* **2024**, *10*, 101058. [CrossRef]

15. Mondragon-Estrada, E.; Kirschning, I.; Nolazco-Flores, J.A.; Camacho-Zuñiga, C. Fostering Digital Transformation in Education: Technology Enhanced Learning from Professors' Experiences in Emergency Remote Teaching. *Front. Educ.* **2023**, *8*, 1250461. [CrossRef]
16. Vonderwell, S.K.; Boboc, M. Promoting Formative Assessment in Online Teaching and Learning. *TechTrends* **2013**, *57*, 22–27. [CrossRef]
17. Gikandi, J.W.; Morrow, D.; Davis, N.E. Online Formative Assessment in Higher Education: A Review of the Literature. *Comput. Educ.* **2011**, *57*, 2333–2351. [CrossRef]
18. Schell, J.; Lukoff, B.; Mazur, E. Catalyzing Learner Engagement Using Cutting-Edge Classroom Response Systems in Higher Education. In *Increasing Student Engagement and Retention Using Classroom Technologies: Classroom Response Systems and Mediated Discourse Technologies*; Emerald Group Publishing Limited: Leeds, UK, 2013; Volume 6, Part E; pp. 233–261, ISBN 978-1-78190-511-1.
19. Rahman, K.A.; Hasan, M.K.; Namaziandost, E.; Ibna Seraj, P.M. Implementing a Formative Assessment Model at the Secondary Schools: Attitudes and Challenges. *Lang. Test. Asia* **2021**, *11*, 18. [CrossRef]
20. Wylie, E.C.; Lyon, C.J. The Role of Technology-Enhanced Self- and Peer Assessment in Formative Assessment. In *Classroom Assessment and Educational Measurement*; Routledge: London, UK, 2019; ISBN 978-0-429-50753-3.
21. Joordens, S.; Paré, D.E.; Walker, R.; Hewitt, J.; Brett, C. *Scaling the Development and Measurement of Transferable Skills: Assessing the Potential of Rubric Scoring in the Context of Peer Assessment*; Higher Education Quality Council of Ontario: Toronto, ON, Canada, 2019.
22. Bennett, R.E. Formative Assessment: A Critical Review. *Assess. Educ. Princ. Policy Pract.* **2011**, *18*, 5–25. [CrossRef]
23. Webb, M.E.; Prasse, D.; Phillips, M.; Kadijevich, D.M.; Angeli, C.; Strijker, A.; Carvalho, A.A.; Andresen, B.B.; Dobozy, E.; Laugesen, H. Challenges for IT-Enabled Formative Assessment of Complex 21st Century Skills. *Technol. Knowl. Learn.* **2018**, *23*, 441–456. [CrossRef]
24. Hopfenbeck, T.N.; Zhang, Z.; Sun, S.Z.; Robertson, P.; McGrane, J.A. Challenges and Opportunities for Classroom-Based Formative Assessment and AI: A Perspective Article. *Front. Educ.* **2023**, *8*, 1270700. [CrossRef]
25. Viberg, O.; Cukurova, M.; Feldman-Maggor, Y.; Alexandron, G.; Shirai, S.; Kanemune, S.; Wasson, B.; Tømte, C.; Spikol, D.; Milrad, M.; et al. What Explains Teachers' Trust in AI in Education Across Six Countries? *Int. J. Artif. Intell. Educ.* **2024**, *35*, 1288–1316. [CrossRef]
26. Naveed, H.; Khan, A.U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Akhtar, N.; Barnes, N.; Mian, A. A Comprehensive Overview of Large Language Models. *ACM Trans. Intell. Syst. Technol.* **2025**, *16*, 1–72. [CrossRef]
27. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. A Survey of Large Language Models. *arXiv* **2025**, arXiv:2303.18223.
28. Xia, Z.; Zhu, L.; Li, B.; Chen, F.; Li, Q.; Liao, C.; Wang, F.; Liu, H. Analyzing 16,193 LLM Papers for Fun and Profits. *arXiv* **2025**, arXiv:2504.08619. [CrossRef]
29. Wang, S.; Xu, T.; Li, H.; Zhang, C.; Liang, J.; Tang, J.; Yu, P.S.; Wen, Q. Large Language Models for Education: A Survey and Outlook. *arXiv* **2024**, arXiv:2403.18105. [CrossRef]
30. Gan, W.; Qi, Z.; Wu, J.; Lin, J.C.-W. Large Language Models in Education: Vision and Opportunities. *arXiv* **2023**, arXiv:2311.13160. [CrossRef]
31. Sharma, S.; Mittal, P.; Kumar, M.; Bhardwaj, V. The Role of Large Language Models in Personalized Learning: A Systematic Review of Educational Impact. *Discov. Sustain.* **2025**, *6*, 243. [CrossRef]
32. Liu, V.; Latif, E.; Zhai, X. Advancing Education through Tutoring Systems: A Systematic Literature Review. *arXiv* **2025**, arXiv:2503.09748. [CrossRef]
33. Chang, W.-L.; Sun, J.C.-Y. Evaluating AI's Impact on Self-Regulated Language Learning: A Systematic Review. *System* **2024**, *126*, 103484. [CrossRef]
34. Awalurahman, H.W.; Fathoni Aji, R.; Budi, I. Transformer and Large Language Models for Automatic Multiple-Choice Question Generation: A Systematic Literature Review. *IEEE Access* **2025**, *13*, 127100–127112. [CrossRef]
35. Mogi, K. Artificial Intelligence, Human Cognition, and Conscious Supremacy. *Front. Psychol.* **2024**, *15*, 1364714. [CrossRef] [PubMed]
36. Siemens, G.; Marmolejo-Ramos, F.; Gabriel, F.; Medeiros, K.; Marrone, R.; Joksimovic, S.; de Laat, M. Human and Artificial Cognition. *Comput. Educ. Artif. Intell.* **2022**, *3*, 100107. [CrossRef]
37. Jordan, M.I.; Mitchell, T.M. Machine Learning: Trends, Perspectives, and Prospects. *Science* **2015**, *349*, 255–260. [CrossRef]
38. Janiesch, C.; Zschech, P.; Heinrich, K. Machine Learning and Deep Learning. *Electron Mark.* **2021**, *31*, 685–695. [CrossRef]
39. Kriegeskorte, N.; Golan, T. Neural Network Models and Deep Learning. *Curr. Biol.* **2019**, *29*, R231–R236. [CrossRef]
40. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
41. Du, S.; Lee, J.; Li, H.; Wang, L.; Zhai, X. Gradient Descent Finds Global Minima of Deep Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 1675–1685.
42. Schmidhuber, J. Deep Learning in Neural Networks: An Overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef]

43. Talaei Khoei, T.; Ould Slimane, H.; Kaabouch, N. Deep Learning: Systematic Review, Models, Challenges, and Research Directions. *Neural Comput. Appl.* **2023**, *35*, 23103–23124. [CrossRef]
44. Mienye, I.D.; Swart, T.G.; Obaido, G. Recurrent Neural Networks: A Comprehensive Review of Architectures, Variants, and Applications. *Information* **2024**, *15*, 517. [CrossRef]
45. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.U.; Polosukhin, I. Attention Is All You Need. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
46. Artzy, A.B.; Schwartz, R. Attend First, Consolidate Later: On the Importance of Attention in Different LLM Layers. In Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, Miami, FL, USA, 15–16 November 2024; Belinkov, Y., Kim, N., Jumelet, J., Mohebbi, H., Mueller, A., Chen, H., Eds.; Association for Computational Linguistics: Miami, FL, USA, 2024; pp. 177–184.
47. Kumar, P. Large Language Models (LLMs): Survey, Technical Frameworks, and Future Challenges. *Artif. Intell. Rev.* **2024**, *57*, 260. [CrossRef]
48. He, H.; Su, W.J. A Law of Next-Token Prediction in Large Language Models. *arXiv* **2024**, arXiv:2408.13442. [CrossRef]
49. Patil, R.; Gudivada, V. A Review of Current Trends, Techniques, and Challenges in Large Language Models (LLMs). *Appl. Sci.* **2024**, *14*, 2074. [CrossRef]
50. Petersen, E.; Potts, C. Lexical Semantics with Large Language Models: A Case Study of English "Break". In Proceedings of the Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, 2–6 May 2023; Vlachos, A., Augenstein, I., Eds.; Association for Computational Linguistics: Dubrovnik, Croatia, 2023; pp. 490–511.
51. Rambelli, G.; Chersoni, E.; Testa, D.; Blache, P.; Lenci, A. Neural Generative Models and the Parallel Architecture of Language: A Critical Review and Outlook. *Top. Cogn. Sci.* **2024**. [CrossRef]
52. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; Burstein, J., Doran, C., Solorio, T., Eds.; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 4171–4186.
53. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv* **2023**, arXiv:1910.10683.
54. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models Are Unsupervised Multitask Learners; OpenAI, 2019. Available online: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (accessed on 2 October 2025).
55. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models Are Few-Shot Learners. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020; Curran Associates Inc.: Red Hook, NY, USA, 2020; pp. 1877–1901.
56. Chen, K.; Shao, A.; Burapacheep, J.; Li, Y. Conversational AI and Equity through Assessing GPT-3's Communication with Diverse Social Groups on Contentious Topics. *Sci. Rep.* **2024**, *14*, 1561. [CrossRef] [PubMed]
57. Xie, J.; Chen, Z.; Zhang, R.; Wan, X.; Li, G. Large Multimodal Agents: A Survey. *arXiv* **2024**, arXiv:2402.15116. [CrossRef]
58. Berrar, D.; Dubitzky, W. Learning, Supervised. In *Encyclopedia of Systems Biology*; Springer: New York, NY, USA, 2013; p. 1122, ISBN 978-1-4419-9863-7.
59. van Engelen, J.E.; Hoos, H.H. A Survey on Semi-Supervised Learning. *Mach. Learn.* **2020**, *109*, 373–440. [CrossRef]
60. Rani, V.; Nabi, S.T.; Kumar, M.; Mittal, A.; Kumar, K. Self-Supervised Learning: A Succinct Review. *Arch. Comput. Methods Eng.* **2023**, *30*, 2761–2775. [CrossRef]
61. Kaelbling, L.P.; Littman, M.L.; Moore, A.W. Reinforcement Learning: A Survey. *arXiv* **1996**, arXiv:cs/9605103. [CrossRef]
62. Qiang, W.; Zhongli, Z. Reinforcement Learning Model, Algorithms and Its Application. In Proceedings of the 2011 International Conference on Mechatronic Science, Electric Engineering and Computer (MEC), Jilin, China, 19–22 August 2011; pp. 1143–1146.
63. Lin, F.; Morland, R.; Yan, H. QuizMaster: An Adaptive Formative Assessment System. In *Generative Intelligence and Intelligent Tutoring Systems*; Sifaleras, A., Lin, F., Eds.; Springer Nature: Cham, Switzerland, 2024; pp. 55–67.
64. Kaufmann, T.; Weng, P.; Bengs, V.; Hüllermeier, E. A Survey of Reinforcement Learning from Human Feedback. *arXiv* **2024**, arXiv:2312.14925. [CrossRef]
65. Zhang, S.; Dong, L.; Li, X.; Zhang, S.; Sun, X.; Wang, S.; Li, J.; Hu, R.; Zhang, T.; Wu, F.; et al. Instruction Tuning for Large Language Models: A Survey. *arXiv* **2024**, arXiv:2308.10792. [CrossRef]
66. Parthasarathy, V.B.; Zafar, A.; Khan, A.; Shahid, A. The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities. *arXiv* **2024**, arXiv:2408.13296. [CrossRef]

67. Carpenter, D.; Min, W.; Lee, S.; Ozogul, G.; Zheng, X.; Lester, J. Assessing Student Explanations with Large Language Models Using Fine-Tuning and Few-Shot Learning. In Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024), Mexico City, Mexico, 20–21 June 2024; Kochmar, E., Bexte, M., Burstein, J., Horbach, A., Laarmann-Quante, R., Tack, A., Yaneva, V., Yuan, Z., Eds.; Association for Computational Linguistics: Mexico City, Mexico, 2024; pp. 403–413.
68. Chen, B.; Zhang, Z.; Langrené, N.; Zhu, S. Unleashing the Potential of Prompt Engineering for Large Language Models. *Patterns* **2025**, *6*, 101260. [CrossRef] [PubMed]
69. Sahoo, P.; Singh, A.K.; Saha, S.; Jain, V.; Mondal, S.; Chadha, A. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. *arXiv* **2025**, arXiv:2402.07927.
70. Wang, W.; Zheng, V.W.; Yu, H.; Miao, C. A Survey of Zero-Shot Learning: Settings, Methods, and Applications. *ACM Trans. Intell. Syst. Technol.* **2019**, *10*, 13:1–13:37. [CrossRef]
71. Chamieh, I.; Zesch, T.; Giebermann, K. LLMs in Short Answer Scoring: Limitations and Promise of Zero-Shot and Few-Shot Approaches. In Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024), Mexico City, Mexico, 20–21 June 2024; Kochmar, E., Bexte, M., Burstein, J., Horbach, A., Laarmann-Quante, R., Tack, A., Yaneva, V., Yuan, Z., Eds.; Association for Computational Linguistics: Mexico City, Mexico, 2024; pp. 309–315.
72. Kong, A.; Zhao, S.; Chen, H.; Li, Q.; Qin, Y.; Sun, R.; Zhou, X.; Wang, E.; Dong, X. Better Zero-Shot Reasoning with Role-Play Prompting. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Mexico City, Mexico, 16–21 June 2024; Duh, K., Gomez, H., Bethard, S., Eds.; Association for Computational Linguistics: Mexico City, Mexico, 2024; Volume 1: Long Papers, pp. 4099–4113.
73. Wang, Y.; Yao, Q.; Kwok, J.T.; Ni, L.M. Generalizing from a Few Examples: A Survey on Few-Shot Learning. *ACM Comput. Surv.* **2020**, *53*, 63:1–63:34. [CrossRef]
74. Zhou, Y.; Li, J.; Xiang, Y.; Yan, H.; Gui, L.; He, Y. The Mystery of In-Context Learning: A Comprehensive Survey on Interpretation and Analysis. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, FL, USA, 12–16 November 2024; Al-Onaizan, Y., Bansal, M., Chen, Y.-N., Eds.; Association for Computational Linguistics: Vienna, Austria, 2024; pp. 14365–14378.
75. Nguyen, H.; Park, S. Providing Automated Feedback on Formative Science Assessments: Uses of Multimodal Large Language Models. In Proceedings of the 15th International Learning Analytics and Knowledge Conference, Dublin, Ireland, 3–7 March 2025; Association for Computing Machinery: New York, NY, USA, 2025; pp. 803–809.
76. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv* **2023**, arXiv:2201.11903.
77. Lee, G.-G.; Latif, E.; Wu, X.; Liu, N.; Zhai, X. Applying Large Language Models and Chain-of-Thought for Automatic Scoring. *Comput. Educ. Artif. Intell.* **2024**, *6*, 100213. [CrossRef]
78. Diao, S.; Wang, P.; Lin, Y.; Pan, R.; Liu, X.; Zhang, T. Active Prompting with Chain-of-Thought for Large Language Models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand, 11–16 August 2024; Ku, L.-W., Martins, A., Srikumar, V., Eds.; Association for Computational Linguistics: Vienna, Austria, 2024; Volume 1: Long Papers, pp. 1330–1350.
79. Cohn, C.; Hutchins, N.; Le, T.; Biswas, G. A Chain-of-Thought Prompting Approach with LLMs for Evaluating Students' Formative Assessment Responses in Science. *AAAI* **2024**, *38*, 23182–23190. [CrossRef]
80. Hu, T.; Zhou, X.-H. Unveiling LLM Evaluation Focused on Metrics: Challenges and Solutions. *arXiv* **2024**, arXiv:2404.09135. [CrossRef]
81. Christen, P.; Hand, D.J.; Kirielle, N. A Review of the F-Measure: Its History, Properties, Criticism, and Alternatives. *ACM Comput. Surv.* **2023**, *56*, 73:1–73:24. [CrossRef]
82. Mendonça, P.C.; Quintal, F.; Mendonça, F. Evaluating LLMs for Automated Scoring in Formative Assessments. *Appl. Sci.* **2025**, *15*, 2787. [CrossRef]
83. Pearson, K. VII. Mathematical Contributions to the Theory of Evolution.—III. Regression, Heredity, and Panmixia. *Phil. Trans. R. Soc. Lond. A* **1896**, *187*, 253–318. [CrossRef]
84. Pack, A.; Barrett, A.; Escalante, J. Large Language Models and Automated Essay Scoring of English Language Learner Writing: Insights into Validity and Reliability. *Comput. Educ. Artif. Intell.* **2024**, *6*, 100234. [CrossRef]
85. Maity, S.; Deroy, A.; Sarkar, S. Can Large Language Models Meet the Challenge of Generating School-Level Questions? *Comput. Educ. Artif. Intell.* **2025**, *8*, 100370. [CrossRef]
86. An, Y.; Liu, J.; Acharya, N.; Hashmi, R. Enhancing Student Learning with LLM-Generated Retrieval Practice Questions: An Empirical Study in Data Science Courses. *arXiv* **2025**, arXiv:2507.05629. [CrossRef]
87. Oli, P.; Banjade, R.; Olney, A.M.; Rus, V. Can LLMs Identify Gaps and Misconceptions in Students' Code Explanations? *arXiv* **2024**, arXiv:2501.10365.

88. Popović, M. chrF: Character n-Gram F-Score for Automatic MT Evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation, Lisbon, Portugal, 17–18 September 2015; Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Hokamp, C., Huck, M., Logacheva, V., Pecina, P., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2015; pp. 392–395.
89. Banerjee, S.; Lavie, A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, MI, USA, 29 June 2005; Goldstein, J., Lavie, A., Lin, C.-Y., Voss, C., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA; pp. 65–72.
90. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. BERTScore: Evaluating Text Generation with BERT. *arXiv* **2020**, arXiv:1904.09675. [CrossRef]
91. Cer, D.; Yang, Y.; Kong, S.; Hua, N.; Limtiaco, N.; John, R.S.; Constant, N.; Guajardo-Cespedes, M.; Yuan, S.; Tar, C.; et al. Universal Sentence Encoder. *arXiv* **2018**, arXiv:1803.11175. [CrossRef]
92. Jaganov, T.; Blake, J.; Villegas, J.; Carr, N. Large Language Model-Driven Dynamic Assessment of Grammatical Accuracy in English Language Learner Writing. *arXiv* **2025**, arXiv:2505.00931. [CrossRef]
93. Hew, K.F.; Huang, W.; Wang, S.; Luo, X.; Gonda, D.E. Towards a Large-Language-Model-Based Chatbot System to Automatically Monitor Student Goal Setting and Planning in Online Learning. *Educ. Technol. Soc.* **2025**, *28*, 112–132. [CrossRef]
94. McHugh, M.L. Interrater Reliability: The Kappa Statistic. *Biochem. Med.* **2012**, *22*, 276–282. [CrossRef]
95. Sridhar, P.; Doyle, A.; Agarwal, A.; Bogart, C.; Savelka, J.; Sakr, M. Harnessing LLMs in Curricular Design: Using GPT-4 to Support Authoring of Learning Objectives. *arXiv* **2023**, arXiv:2306.17459. [CrossRef]
96. Ng, C.; Fung, Y. Educational Personalized Learning Path Planning with Large Language Models. *arXiv* **2024**, arXiv:2407.11773. [CrossRef]
97. Seßler, K.; Bewersdorff, A.; Nerdel, C.; Kasneci, E. Towards Adaptive Feedback with AI: Comparing the Feedback Quality of LLMs and Teachers on Experimentation Protocols. *arXiv* **2025**, arXiv:2502.12842. [CrossRef]
98. Hu, B.; Zhu, J.; Pei, Y.; Gu, X. Exploring the Potential of LLM to Enhance Teaching Plans through Teaching Simulation. *npj Sci. Learn.* **2025**, *10*, 7. [CrossRef]
99. Fan, H.; Chen, G.; Wang, X.; Peng, Z. LessonPlanner: Assisting Novice Teachers to Prepare Pedagogy-Driven Lesson Plans with Large Language Models. *arXiv* **2024**, arXiv:2408.01102.
100. Cuéllar, Ó.; Contero, M.; Hincapié, M. Personalized and Timely Feedback in Online Education: Enhancing Learning with Deep Learning and Large Language Models. *Multimodal Technol. Interact.* **2025**, *9*, 45. [CrossRef]
101. Fisher, R.A. Statistical Methods for Research Workers. In *Breakthroughs in Statistics. Methodology and Distribution*, 1st ed.; Springer Series in Statistics; Springer: New York, NY, USA, 1992; pp. 66–70, ISBN 978-0-387-94039-7.
102. Cohen, J. A Power Primer. *Psychol. Bull.* **1992**, *112*, 155–159. [CrossRef]
103. Ge, W.; Sun, Y.; Wang, Z.; Zheng, H.; He, W.; Wang, P.; Zhu, Q.; Wang, B. SRLAgent: Enhancing Self-Regulated Learning Skills through Gamification and LLM Assistance. *arXiv* **2025**, arXiv:2506.09968. [CrossRef]
104. Shimorina, A. Human vs Automatic Metrics: On the Importance of Correlation Design. *arXiv* **2021**, arXiv:1805.11474. [CrossRef]
105. Katuka, G.A.; Gain, A.; Yu, Y.-Y. Investigating Automatic Scoring and Feedback Using Large Language Models. *arXiv* **2024**, arXiv:2405.00602. [CrossRef]
106. Chu, Z.; Wang, S.; Xie, J.; Zhu, T.; Yan, Y.; Ye, J.; Zhong, A.; Hu, X.; Liang, J.; Yu, P.S.; et al. LLM Agents for Education: Advances and Applications. *arXiv* **2025**, arXiv:2503.11733.
107. Kjell, O.N.E.; Sikström, S.; Kjell, K.; Schwartz, H.A. Natural Language Analyzed with AI-Based Transformers Predict Traditional Subjective Well-Being Measures Approaching the Theoretical Upper Limits in Accuracy. *Sci. Rep.* **2022**, *12*, 3918. [CrossRef]
108. Hosna, A.; Merry, E.; Gyalmo, J.; Alom, Z.; Aung, Z.; Azim, M.A. Transfer Learning: A Friendly Introduction. *J. Big Data* **2022**, *9*, 102. [CrossRef]
109. Sung, C.; Dhamecha, T.I.; Mukhi, N. Improving Short Answer Grading Using Transformer-Based Pre-Training. In Proceedings of the Artificial Intelligence in Education, Chicago, IL, USA, 25–29 June 2019; Isotani, S., Millán, E., Ogan, A., Hastings, P., McLaren, B., Luckin, R., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 469–481.
110. Lee, J.; Koh, E. Teamwork Dimensions Classification Using BERT. In Proceedings of the Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky, Tokyo, Japan, 3–7 July 2023; Wang, N., Rebolledo-Mendez, G., Dimitrova, V., Matsuda, N., Santos, O.C., Eds.; Springer Nature: Cham, Switzerland, 2023; pp. 254–259.
111. Beltagy, I.; Peters, M.E.; Cohan, A. Longformer: The Long-Document Transformer. *arXiv* **2020**, arXiv:2004.05150. [CrossRef]
112. Botarleanu, R.-M.; Dascalu, M.; Allen, L.K.; Crossley, S.A.; McNamara, D.S. Multitask Summary Scoring with Longformers. In Proceedings of the Artificial Intelligence in Education, Virtual, 27–31 July 2022; Rodrigo, M.M., Matsuda, N., Cristea, A.I., Dimitrova, V., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 756–761.

113. Tran, N.; Pierce, B.; Litman, D.; Correnti, R.; Matsumura, L.C. Utilizing Natural Language Processing for Automated Assessment of Classroom Discussion. In Proceedings of the Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky, Tokyo, Japan, 3–7 July 2023; Wang, N., Rebolledo-Mendez, G., Dimitrova, V., Matsuda, N., Santos, O.C., Eds.; Springer Nature: Cham, Switzerland, 2023; pp. 490–496.
114. Cochran, K.; Cohn, C.; Rouet, J.F.; Hastings, P. Improving Automated Evaluation of Student Text Responses Using GPT-3.5 for Text Data Augmentation. In Proceedings of the Artificial Intelligence in Education, Tokyo, Japan, 3–7 July 2023; Wang, N., Rebolledo-Mendez, G., Matsuda, N., Santos, O.C., Dimitrova, V., Eds.; Springer Nature: Cham, Switzerland, 2023; pp. 217–228.
115. Jürgensmeier, L.; Skiera, B. Generative AI for Scalable Feedback to Multimodal Exercises. *Int. J. Res. Mark.* **2024**, *41*, 468–488. [CrossRef]
116. Baytak, A. The Content Analysis of the Lesson Plans Created by ChatGPT and Google Gemini. *Res. Soc. Sci. Technol.* **2024**, *9*, 329–350. [CrossRef]
117. Gubelmann, R.; Burkhard, M.; Ivanova, R.V.; Niklaus, C.; Bermeitinger, B.; Handschuh, S. Exploring the Usefulness of Open and Proprietary LLMs in Argumentative Writing Support. In Proceedings of the Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky, Tokyo, Japan, 3–7 July 2023; Olney, A.M., Chounta, I.-A., Liu, Z., Santos, O.C., Bittencourt, I.I., Eds.; Springer Nature: Cham, Switzerland, 2024; pp. 175–182.
118. Kumar, S.S.; Lones, M.A.; Maarek, M.; Zantout, H. Investigating the Proficiency of Large Language Models in Formative Feedback Generation for Student Programmers. In Proceedings of the 2024 IEEE/ACM International Workshop on Large Language Models for Code (LLM4Code), Lisbon, Portugal, 20 April 2024; pp. 88–93.
119. Lyu, C.; Wu, M.; Wang, L.; Huang, X.; Liu, B.; Du, Z.; Shi, S.; Tu, Z. Macaw-LLM: Multi-Modal Language Modeling with Image, Audio, Video, and Text Integration. *arXiv* **2023**, arXiv:2306.09093.
120. Olney, A. Generating Multiple Choice Questions from a Textbook: LLMs Match Human Performance on Most Metrics. In Proceedings of the Workshop on Empowering Education with LLMs—The Next-Gen Interface and Content Generation at the AIED'23 Conference, Tokyo, Japan, 7 July 2023.
121. Chung, H.W.; Hou, L.; Longpre, S.; Zoph, B.; Tai, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. Scaling Instruction-Finetuned Language Models. *arXiv* **2022**, arXiv:2210.11416. [CrossRef]
122. Caccavale, F.; Gargalo, C.L.; Gernaey, K.V.; Krühne, U. FermentAI: Large Language Models in Chemical Engineering Education for Learning Fermentation Processes. In *Computer Aided Chemical Engineering, Proceedings of the 34 European Symposium on Computer Aided Process Engineering/15 International Symposium on Process Systems Engineering, Florence, Italy, 2–6 June 2024*; Manenti, F., Reklaitis, G.V., Eds.; Elsevier: Amsterdam, The Netherlands, 2024; Volume 53, pp. 3493–3498.
123. Ding, N.; Qin, Y.; Yang, G.; Wei, F.; Yang, Z.; Su, Y.; Hu, S.; Chen, Y.; Chan, C.-M.; Chen, W.; et al. Parameter-Efficient Fine-Tuning of Large-Scale Pre-Trained Language Models. *Nat. Mach. Intell.* **2023**, *5*, 220–235. [CrossRef]
124. Wang, X.; Sun, J.; Qi, C. CEDA-TQA: Context Enhancement and Domain Adaptation Method for Textbook QA Based on LLM and RAG. In Proceedings of the 2024 International Conference on Networking and Network Applications (NaNA), Yinchuan City, China, 9–12 August 2024; pp. 263–268.
125. Suzuki, W.A.; Feliú-Mójer, M.I.; Hasson, U.; Yehuda, R.; Zarate, J.M. Dialogues: The Science and Power of Storytelling. *J. Neurosci.* **2018**, *38*, 9468–9470. [CrossRef] [PubMed]
126. Chhun, C.; Suchanek, F.M.; Clavel, C. Do Language Models Enjoy Their Own Stories? Prompting Large Language Models for Automatic Story Evaluation. *Trans. Assoc. Comput. Linguist.* **2024**, *12*, 1122–1142. [CrossRef]
127. Hitzler, P.; Eberhart, A.; Ebrahimi, M.; Sarker, M.K.; Zhou, L. Neuro-Symbolic Approaches in Artificial Intelligence. *Natl. Sci. Rev.* **2022**, *9*, nwac035. [CrossRef]
128. Farrell, R.; Ware, S.G. Large Language Models as Narrative Planning Search Guides. *IEEE Trans. Games* **2025**, *17*, 419–428. [CrossRef]
129. Simon, N.; Muise, C. TattleTale—Storytelling with Planning and Large Language Models. In Proceedings of the ICAPS Workshop on Scheduling and Planning Applications (SPARK 2022), Singapore, 7–12 June 2022.
130. Alkhatlan, A.; Kalita, J. Intelligent Tutoring Systems: A Comprehensive Historical Survey with Recent Developments. *arXiv* **2018**, arXiv:1812.09628. [CrossRef]
131. Lin, C.-C.; Huang, A.Y.Q.; Lu, O.H.T. Artificial Intelligence in Intelligent Tutoring Systems toward Sustainable Education: A Systematic Review. *Smart Learn. Environ.* **2023**, *10*, 41. [CrossRef]
132. Steinert, S.; Avila, K.E.; Ruzika, S.; Kuhn, J.; Küchemann, S. Harnessing Large Language Models to Enhance Self-Regulated Learning via Formative Feedback. *Smart Learn. Environ.* **2024**, *11*, 62. [CrossRef]
133. Yadav, G.; Tseng, Y.-J.; Ni, X. Contextualizing Problems to Student Interests at Scale in Intelligent Tutoring System Using Large Language Models. *arXiv* **2023**, arXiv:2306.00190. [CrossRef]

134. Cai, Z.; Park, S.; Nixon, N.; Doroudi, S. Advancing Knowledge Together: Integrating Large Language Model-Based Conversational AI in Small Group Collaborative Learning. In Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, New York, NY, USA, 11–16 May 2024; Association for Computing Machinery: New York, NY, USA, 2024; pp. 1–9.
135. Zhang, Z.; Sun, B.; An, P. Breaking Barriers or Building Dependency? Exploring Team-LLM Collaboration in AI-Infused Classroom Debate. *arXiv* **2025**, arXiv:2501.09165.
136. Chiang, C.-W.; Lu, Z.; Li, Z.; Yin, M. Enhancing AI-Assisted Group Decision Making through LLM-Powered Devil's Advocate. In Proceedings of the 29th International Conference on Intelligent User Interfaces, Greenville, SC, USA, 18–21 March 2024; Association for Computing Machinery: New York, NY, USA, 2024; pp. 103–119.
137. Wang, Q.; Wang, Z.; Su, Y.; Tong, H.; Song, Y. Rethinking the Bounds of LLM Reasoning: Are Multi-Agent Discussions the Key? *arXiv* **2024**, arXiv:2402.18272. [CrossRef]
138. Küçüktütüncü, E.; Izzouzi, L. "Let's Ask What AI Thinks Then!": Using LLMs for Collaborative Problem-Solving in Virtual Environments. In Proceedings of the 2024 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), Bellevue, WA, USA, 21–25 October 2024; pp. 193–198.
139. Chen, S.; Taveekitworachai, P.; Xia, Y.; Li, X.; Gursesli, M.C.; Lanata, A.; Guazzini, A.; Thawonmas, R. Don't Do That! Reverse Role Prompting Helps Large Language Models Stay in Personality Traits. In *Interactive Storytelling*; Murray, J.T., Reyes, M.C., Eds.; Springer Nature: Cham, Switzerland, 2025; pp. 101–114.
140. Smiderle, R.; Rigo, S.J.; Marques, L.B.; Peçanha de Miranda Coelho, J.A.; Jaques, P.A. The Impact of Gamification on Students' Learning, Engagement and Behavior Based on Their Personality Traits. *Smart Learn. Environ.* **2020**, *7*, 3. [CrossRef]
141. da Rocha Seixas, L.; Gomes, A.S.; de Melo Filho, I.J. Effectiveness of Gamification in the Engagement of Students. *Comput. Hum. Behav.* **2016**, *58*, 48–63. [CrossRef]
142. Li, M.; Ma, S.; Shi, Y. Examining the Effectiveness of Gamification as a Tool Promoting Teaching and Learning in Educational Settings: A Meta-Analysis. *Front. Psychol.* **2023**, *14*, 1253549. [CrossRef]
143. Lampropoulos, G.; Sidiropoulos, A. Impact of Gamification on Students' Learning Outcomes and Academic Performance: A Longitudinal Study Comparing Online, Traditional, and Gamified Learning. *Educ. Sci.* **2024**, *14*, 367. [CrossRef]
144. Plass, J.L.; Homer, B.D.; Kinzer, C.K. Foundations of Game-Based Learning. *Educ. Psychol.* **2015**, *50*, 258–283. [CrossRef]
145. Videnovik, M.; Vold, T.; Kiønig, L.; Madevska Bogdanova, A.; Trajkovik, V. Game-Based Learning in Computer Science Education: A Scoping Literature Review. *Int. J. STEM Educ.* **2023**, *10*, 54. [CrossRef]
146. Goslen, A.; Kim, Y.J.; Rowe, J.; Lester, J. LLM-Based Student Plan Generation for Adaptive Scaffolding in Game-Based Learning Environments. *Int. J. Artif. Intell. Educ.* **2025**, *35*, 533–558. [CrossRef]
147. Huber, S.E.; Kiili, K.; Nebel, S.; Ryan, R.M.; Sailer, M.; Ninaus, M. Leveraging the Potential of Large Language Models in Education Through Playful and Game-Based Learning. *Educ. Psychol. Rev.* **2024**, *36*, 25. [CrossRef]
148. Chang, H.-T.; Lin, C.-Y. Applying Competition-Based Learning to Stimulate Students' Practical and Competitive AI Ability in a Machine Learning Curriculum. *IEEE Trans. Educ.* **2024**, *67*, 256–265. [CrossRef]
149. McGuire, M. Impact of Competition-Based Learning on Student Engagement and Performance. *Int. J. Constr. Educ. Res.* **2025**, 1–30. [CrossRef]
150. Burguillo, J.C. Using Game Theory and Competition-Based Learning to Stimulate Student Motivation and Performance. *Comput. Educ.* **2010**, *55*, 566–575. [CrossRef]
151. Cagiltay, N.E.; Ozcelik, E.; Ozcelik, N.S. The Effect of Competition on Learning in Games. *Comput. Educ.* **2015**, *87*, 35–41. [CrossRef]
152. Cipresso, P.; Giglioli, I.A.C.; Raya, M.A.; Riva, G. The Past, Present, and Future of Virtual and Augmented Reality Research: A Network and Cluster Analysis of the Literature. *Front. Psychol.* **2018**, *9*, 2086. [CrossRef]
153. Stracke, C.M.; Bothe, P.; Adler, S.; Heller, E.S.; Deuchler, J.; Pomino, J.; Wölfel, M. Immersive Virtual Reality in Higher Education: A Systematic Review of the Scientific Literature. *Virtual Real.* **2025**, *29*, 64. [CrossRef]
154. Campos, E.; Hidrogo, I.; Zavala, G. Impact of Virtual Reality Use on the Teaching and Learning of Vectors. *Front. Educ.* **2022**, *7*, 965640. [CrossRef]
155. Mendes Neto, F.M.; De Almeida, L.M.B.; De Freitas Lopes, E.G.; De Araújo Pontes, V.M.; Chagas, J.F.S.; Alves, F.H. The Development and Evaluation of a Multi-Agent System for Supporting Flipped Classroom. *Creat. Educ.* **2018**, *9*, 1667–1679. [CrossRef]
156. Zhang, Z.; Zhang-Li, D.; Yu, J.; Gong, L.; Zhou, J.; Hao, Z.; Jiang, J.; Cao, J.; Liu, H.; Liu, Z.; et al. Simulating Classroom Education with LLM-Empowered Agents. *arXiv* **2024**, arXiv:2406.19226. [CrossRef]
157. Wang, T.; Zhan, Y.; Lian, J.; Hu, Z.; Yuan, N.J.; Zhang, Q.; Xie, X.; Xiong, H. LLM-Powered Multi-Agent Framework for Goal-Oriented Learning in Intelligent Tutoring System. In Proceedings of the Companion Proceedings of the ACM on Web Conference 2025, Sydney, Australia, 28 April–2 May 2025; Association for Computing Machinery: New York, NY, USA, 2025; pp. 510–519.

158. Amini, H.; Mia, M.J.; Saadati, Y.; Imteaj, A.; Nabavirazavi, S.; Thakker, U.; Hossain, M.Z.; Fime, A.A.; Iyengar, S.S. Distributed LLMs and Multimodal Large Language Models: A Survey on Advances, Challenges, and Future Directions. *arXiv* **2025**, arXiv:2503.16585. [CrossRef]
159. Song, Y.; Wu, K.; Ding, J. Developing an Immersive Game-Based Learning Platform with Generative Artificial Intelligence and Virtual Reality Technologies—"LearningverseVR". *Comput. Educ. X Real.* **2024**, *4*, 100069. [CrossRef]
160. Brito, I.A.; Dollis, J.S.; Färber, F.B.; Ribeiro, P.S.F.B.; Sousa, R.T.; Filho, A.R.G. Integrating Personality into Digital Humans: A Review of LLM-Driven Approaches for Virtual Reality. *arXiv* **2025**, arXiv:2503.16457.
161. Lu, D.; Xie, Y.-N. The Application of Educational Technology to Develop Problem-Solving Skills: A Systematic Review. *Think. Ski. Creat.* **2024**, *51*, 101454. [CrossRef]
162. Dron, J. Educational Technology: What It Is and How It Works. *AI Soc.* **2022**, *37*, 155–166. [CrossRef]
163. Hillman, V. *Edtech Procurement Matters: It Needs a Coherent Solution, Clear Governance and Market Standards*; Department of Social Policy, London School of Economics and Political Science: London, UK, 2022.
164. Ali, H.; Prompiengchai, S.; Joordens, S. Educational Technology Procurement at Canadian Colleges and Universities: An Environmental Scan. *Standards* **2024**, *4*, 1–24. [CrossRef]
165. Morrison, J.; Ross, S.; Corcoran, R.; Reid, A.J. *Fostering Market Effciency in K-12 Ed-Tech Procurement*; Johns Hopkins University: Baltimore, MD, USA, 2014.
166. Macgilchrist, F.; Potter, J.; Williamson, B. Challenging the Inequitable Impacts of Edtech. *Learn. Media Technol.* **2024**, *49*, 147–150. [CrossRef]
167. Crompton, H.; Chigona, A.; Jordan, K.; Myers, C. *Inequalities in Girls' Learning Opportunities via EdTech: Addressing the Challenge of Covid-19*; EdTech Hub: Victoria, UK, 2021.
168. Czerniewicz, L.; Cronin, C. *Higher Education for Good: Teaching and Learning Futures*; Open Book Publishers: Cambridge, UK, 2023; ISBN 978-1-80511-127-6.
169. Fouad, N.S. The Security Economics of EdTech: Vendors' Responsibility and the Cybersecurity Challenge in the Education Sector. *Digit. Policy Regul. Gov.* **2022**, *24*, 259–273. [CrossRef]
170. Yao, Y.; Duan, J.; Xu, K.; Cai, Y.; Sun, Z.; Zhang, Y. A Survey on Large Language Model (LLM) Security and Privacy: The Good, The Bad, and The Ugly. *High-Confid. Comput.* **2024**, *4*, 100211. [CrossRef]
171. Das, B.C.; Amini, M.H.; Wu, Y. Security and Privacy Challenges of Large Language Models: A Survey. *ACM Comput. Surv.* **2025**, *57*, 152:1–152:39. [CrossRef]
172. Regan, P.M.; Jesse, J. Ethical Challenges of Edtech, Big Data and Personalized Learning: Twenty-First Century Student Sorting and Tracking. *Ethics Inf. Technol.* **2019**, *21*, 167–179. [CrossRef]
173. Kousa, P.; Niemi, H. AI Ethics and Learning: EdTech Companies' Challenges and Solutions. *Interact. Learn. Environ.* **2023**, *31*, 6735–6746. [CrossRef]
174. Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* **2025**, *43*, 42:1–42:55. [CrossRef]
175. Vorobyeva, K.I.; Belous, S.; Savchenko, N.V.; Smirnova, L.M.; Nikitina, S.A.; Zhdanov, S.P. Personalized Learning through AI: Pedagogical Approaches and Critical Insights. *Contemp. Educ. Technol.* **2025**, *17*, ep574. [CrossRef] [PubMed]
176. Alamri, H.; Lowell, V.; Watson, W.; Watson, S.L. Using Personalized Learning as an Instructional Approach to Motivate Learners in Online Higher Education: Learner Self-Determination and Intrinsic Motivation. *J. Res. Technol. Educ.* **2020**, *52*, 322–352. [CrossRef]
177. Rodriguez-Segura, D. EdTech in Developing Countries: A Review of the Evidence. *World Bank Res. Obs.* **2022**, *37*, 171–203. [CrossRef]
178. Al Tal, S. Educational Technology and Its Impact on Learning. In *Technology for Societal Transformation: Exploring the Intersection of Information Technology and Societal Development*; Yesufu, L.O., Nohuddin, P.N.E., Eds.; Springer Nature: Singapore, 2025; pp. 29–44, ISBN 978-981-96-1721-0.
179. Wang, L.; Chen, S.; Jiang, L.; Pan, S.; Cai, R.; Yang, S.; Yang, F. Parameter-Efficient Fine-Tuning in Large Language Models: A Survey of Methodologies. *Artif. Intell. Rev.* **2025**, *58*, 227. [CrossRef]
180. Micallef, K.; Borg, C. MELABenchv1: Benchmarking Large Language Models against Smaller Fine-Tuned Models for Low-Resource Maltese NLP. *arXiv* **2025**, arXiv:2506.04385.
181. Wu, F.; Dang, Y.; Li, M. A Systematic Review of Responses, Attitudes, and Utilization Behaviors on Generative AI for Teaching and Learning in Higher Education. *Behav. Sci.* **2025**, *15*, 467. [CrossRef]
182. Alwaqdani, M. Investigating Teachers' Perceptions of Artificial Intelligence Tools in Education: Potential and Difficulties. *Educ. Inf. Technol.* **2025**, *30*, 2737–2755. [CrossRef]
183. Bergdahl, N.; Sjöberg, J. Attitudes, Perceptions and AI Self-Efficacy in K-12 Education. *Comput. Educ. Artif. Intell.* **2025**, *8*, 100358. [CrossRef]

184. Lucas, M.; Zhang, Y.; Bem-haja, P.; Vicente, P.N. The Interplay between Teachers' Trust in Artificial Intelligence and Digital Competence. *Educ. Inf. Technol.* **2024**, *29*, 22991–23010. [CrossRef]
185. Nazaretsky, T.; Cukurova, M.; Alexandron, G. An Instrument for Measuring Teachers' Trust in AI-Based Educational Technology. In Proceedings of the LAK22: 12th International Learning Analytics and Knowledge Conference, Online, 21–25 March 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 56–66.
186. Nazaretsky, T.; Ariely, M.; Cukurova, M.; Alexandron, G. Teachers' Trust in AI-Powered Educational Technology and a Professional Development Program to Improve It. *Br. J. Educ. Technol.* **2022**, *53*, 914–931. [CrossRef]
187. Dobrovská, D.; Vaněček, D.; Yorulmaz, Y.I. Students' Attitudes towards AI in Teaching and Learning. *Int. J. Eng. Pedagog. (iJEP)* **2024**, *14*, 88–106. [CrossRef]
188. Alamäki, A.; Khan, U.A.; Kauttonen, J.; Schlögl, S. An Experiment of AI-Based Assessment: Perspectives of Learning Preferences, Benefits, Intention, Technology Affinity, and Trust. *Educ. Sci.* **2024**, *14*, 1386. [CrossRef]
189. Shahzad, T.; Mazhar, T.; Tariq, M.U.; Ahmad, W.; Ouahada, K.; Hamam, H. A Comprehensive Review of Large Language Models: Issues and Solutions in Learning Environments. *Discov. Sustain.* **2025**, *6*, 27. [CrossRef]
190. Vatsal, S.; Dubey, H. A Survey of Prompt Engineering Methods in Large Language Models for Different NLP Tasks. *arXiv* **2024**, arXiv:2407.12994. [CrossRef]
191. Razavi, A.; Soltangheis, M.; Arabzadeh, N.; Salamat, S.; Zihayat, M.; Bagheri, E. Benchmarking Prompt Sensitivity in Large Language Models. In Proceedings of the Advances in Information Retrieval, Lucca, Italy, 6–10 April 2025; Hauff, C., Macdonald, C., Jannach, D., Kazai, G., Nardini, F.M., Pinelli, F., Silvestri, F., Tonellotto, N., Eds.; Springer Nature: Cham, Switzerland, 2025; pp. 303–313.
192. Laskar, M.T.R.; Alqahtani, S.; Bari, M.S.; Rahman, M.; Khan, M.A.M.; Khan, H.; Jahan, I.; Bhuiyan, A.; Tan, C.W.; Parvez, M.R.; et al. A Systematic Survey and Critical Review on Evaluating Large Language Models: Challenges, Limitations, and Recommendations. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, FL, USA, 12–16 November 2024; Al-Onaizan, Y., Bansal, M., Chen, Y.-N., Eds.; Association for Computational Linguistics: Vienna, Austria, 2024; pp. 13785–13816.
193. Yang, H.; Hu, M.; Most, A.; Hawkins, W.A.; Murray, B.; Smith, S.E.; Li, S.; Sikora, A. Evaluating Accuracy and Reproducibility of Large Language Model Performance on Critical Care Assessments in Pharmacy Education. *Front. Artif. Intell.* **2025**, *7*, 1514896. [CrossRef]
194. Seo, H.; Hwang, T.; Jung, J.; Kang, H.; Namgoong, H.; Lee, Y.; Jung, S. Large Language Models as Evaluators in Education: Verification of Feedback Consistency and Accuracy. *Appl. Sci.* **2025**, *15*, 671. [CrossRef]