Computers and Education Open 10 (2026) 100374

Contents lists available at ScienceDirect

Computers and Education Open

journal homepage: www.sciencedirect.com/journal/computers-and-education-open

LLM-generated formative feedback in education: A qualitative systematic literature review

Uwe Maier *, Moritz Seibold, Christian Klotz

University of Education Schwäbisch Gmünd, D-73525 Schwäbisch Gmünd, Oberbettringerstraße 200, Germany

ARTICLE INFO

Keywords:
Llm-generated feedback
Generative artificial intelligence
Formative feedback
Prompting strategies
Feedback literacy
Learning outcomes
Systematic review

ABSTRACT

Formative feedback is a central component of learning and a key feature of digital learning systems. Since late 2022, instruction-tuned large language models (LLMs) have enabled scalable generation of elaborated, formative feedback for text-based student assignments across domains. Existing reviews do not adequately reflect these developments. To address this gap, we conducted a qualitative systematic literature review following PRISMA guidelines, synthesizing evidence from 47 peer-reviewed empirical studies including 121 distinct research questions, published between 2022 and 2026. The review was guided by three research questions examining (1) theoretical framings and dominant research questions, (2) methodological and contextual characteristics, and (3) technological trends and empirical findings related to feedback quality, feedback processing, and learning outcomes. The reviewed literature is primarily grounded in formative feedback theory, self-regulated learning, feedback literacy, and human–AI interaction frameworks. Methodologically, studies are dominated by short-term experiments, quasi-experimental classroom studies, and expert evaluations of feedback quality, mostly situated in higher education. Technologically, most studies employ proprietary, pre-trained GPT-3.5 or GPT-4 models with context-enriched, role-based zero-shot prompting or few-shot prompting, while fine-tuned, open-source, and multi-agent approaches are beginning to emerge. Empirical findings indicate that LLM-generated formative feedback consistently outperforms no-feedback conditions and often improves revision quality, motivation, and short-term learning outcomes, sometimes approaching teacher feedback under well-designed prompting. However, recurring risks include hallucinations, over-positivity, and misclassification of student work. Emerging evidence suggests that instruction fine-tuning, grounding prompts in student artifacts, and teacher-in-the-loop oversight can mitigate these issues.

1. Introduction

Formative feedback is widely recognized as a core part of learning across behaviorist, cognitivist, and socio-constructivist perspectives and can support error correction, knowledge construction, self-regulated learning, and motivation [1–4]. Compared to summative feedback (exam scores, grading, marks), formative feedback reflects a students' ongoing learning process and is informative for further instructional adjustments. Meta-analyses consistently show that formative feedback can produce substantial learning gains [5–8], yet also highlight large variability. Some feedback improves performance, while poorly designed or individually misaligned feedback is either ignored by students or can impair motivation and learning [9,10]. Understanding when and why feedback is effective remains a challenge.

Digital formative feedback has long been explored as a scalable means to enhance learning and reduce teacher workload [5,11–13]. However, early systems were mostly limited to correctness-based responses in domains with well-structured tasks [14,15], and automated feedback for open-ended assignments required costly Natural Language Processing (NLP) pipelines and large training datasets [16,17]. Consequently, much empirical evidence and many theoretical models were grounded in studies using simple tasks and simple feedback types. The advent of large language models (LLMs) represents a fundamental shift to generate contextualized, and individualized formative feedback for open-ended written student responses [18]. This capability has triggered rapid growth in applications and research examining the quality, use, and effects of LLM-generated feedback (e.g., [19–21]). Educators and researchers are now asking whether this technology offers high-quality feedback across diverse learning domains.

However, the research field of LLM-generated formative feedback is

* Corresponding author.
E-mail address: uwe.maier@pH-gmuend.de (U. Maier).

https://doi.org/10.1016/j.caeo.2026.100374
Received 5 January 2026; Received in revised form 28 February 2026; Accepted 8 May 2026
Available online 11 May 2026

highly dynamic and heterogeneous, both in research methodology and in the technical approaches used to integrate LLMs into learning environments (e.g., [21–23,130]). Moreover, several types of research questions recur: the accuracy and pedagogical value of LLM-generated feedback (e.g., [24]), learners' perception and use of such feedback (e.g., [25]), and the cognitive and motivational effects of LLM-generated feedback (e.g., [21]). We also found that the existing qualitative literature reviews [26–29] and meta-analyses [30] on feedback and artificial intelligence (AI) do not adequately capture these developments. While these reviews provide valuable overviews of AI-based feedback more broadly, they either aggregate heterogeneous AI systems under a single umbrella, focus on pre-LLM technologies, or specific contexts such as higher education writing. Consequently, they do not systematically account for the conceptual and technological shift following the public release of general purpose, instruction-tuned LLMs in late 2022. Given the rapid acceleration of empirical studies since then, an updated and focused synthesis is needed. The goal of this systematic literature review is therefore to map research questions, research methodology and empirical evidence in the current field of LLM-generated formative feedback research. The remainder of the paper presents (2) theoretical background and open issues, (3) research questions, (4) the PRISMA-based methodology, (5) results, and (6) discussion with limitations and implications.

2. Theoretical background

2.1. Feedback and learning

Across major learning paradigms, the conceptualization of feedback has shifted substantially. In behaviorist frameworks, feedback operated primarily as reinforcement that strengthened correct behavior [1,31,32]. With the rise of cognitivism, feedback came to be understood as information that helps learners verify, correct, or reorganize knowledge structures [3,33]. Constructivist and socio-constructivist perspectives further broadened this view by positioning feedback as part of a dialogic, sense-making process through which learners actively interpret errors, regulate their learning, and co-construct meaning [34–37].

A major concept across learning theories is the productive role of errors. Research in cognitive science consistently shows that errors are not harmful when followed by corrective feedback [3,38]. To leverage errors as learning opportunities, feedback must be sufficiently informative [4,13,39,40]. Elaborated feedback provides learners with resources to correct their misconceptions either by retrieving the correct response or by deriving it from the explanation [14,4]. Importantly, instructional negative feedback that includes guidance on how to improve is significantly less demotivating than evaluative negative feedback lacking such information [6,41].

In educational contexts, the distinction between formative and summative feedback is crucial. Summative feedback provides diagnostic information on student performance at the end of a learning process, primarily for purposes of grading or certification. In contrast, formative feedback aims to support ongoing learning by offering task-specific information that helps students improve performance, correct errors, and refine strategies for subsequent tasks [2,4]. Fig. 1 presents an integrative conceptual framework of formative feedback and synthesizes the key categories of variables examined in formative feedback research, including moderators and outcome variables. The framework conceptualizes formative feedback as embedded within an instructional context and linked to explicit learning goals. Within this context, elaborated formative feedback may be generated by different agents—teachers, peers, or automated systems—typically in response to a formative assessment activity (e.g., a diagnostic task, test, or writing assignment).

A persistent problem is that many learners do not effectively process the formative feedback they receive [42,43]. Studies in digital learning environments show that low-achieving students often ignore or only superficially engage with elaborated feedback, even though they would benefit most from it [10,44]. Even when students do attend to feedback, they may misinterpret it, lack strategies for applying it, or feel overwhelmed by its complexity [37,45]. These difficulties have prompted scholars to highlight the importance of feedback literacy, which involves the dispositions learners need to appreciate feedback, make informed judgments, regulate the affective discomfort it may trigger, and translate information into action [35].

Moreover, research highlights also the role of feedback choices and

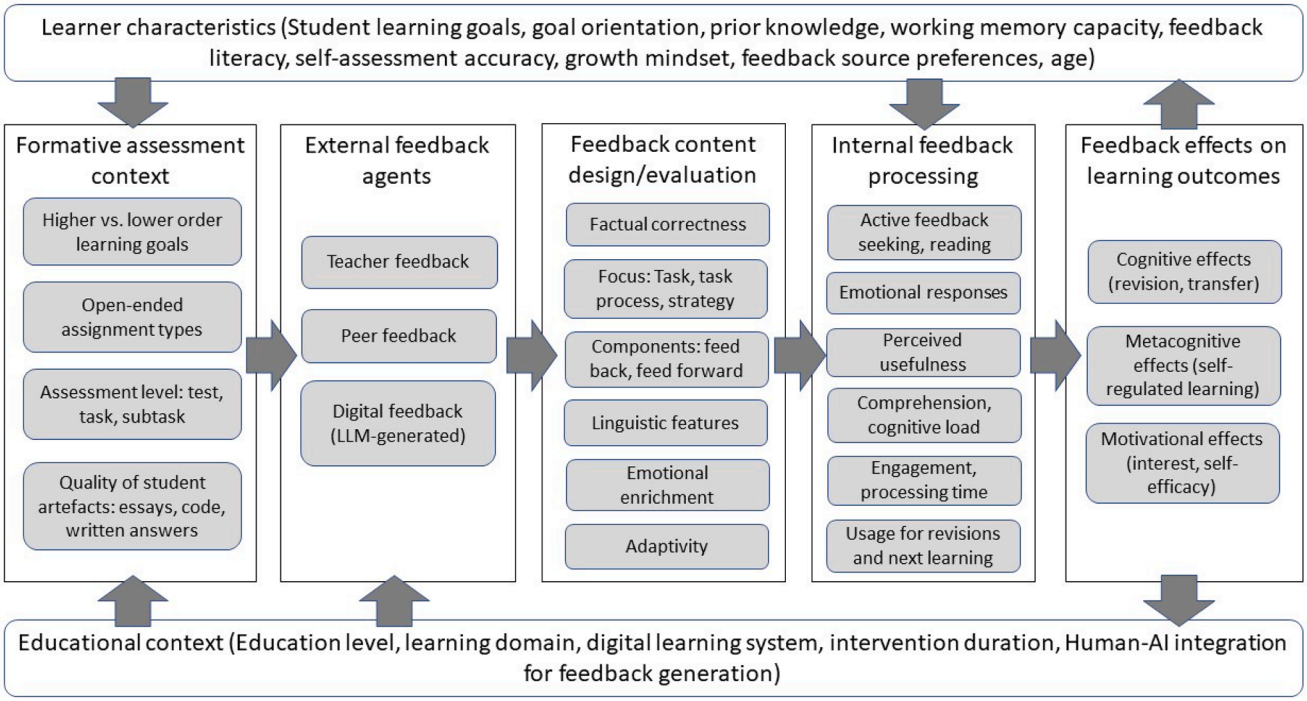

Fig. 1. Integrative conceptual framework of the elaborated formative feedback research.

feedback seeking in digital learning systems. Studies show that students with active error feedback seeking had better revision performance [46–48]. Furthermore, students' mindset to deal with negative feedback is crucial in feedback processing. Students with a growth mindset are more likely to engage with errors and subsequently perform better [49]. Research also revealed the potential of external stimuli that help students generate and learn from internal error feedback loops [50].

The Interactive Two-Feedback-Loop (ITFL) model is an empirically grounded, integrative theory framework for formative feedback in learning environments [51,52]. This model conceptualizes feedback as a dynamic regulatory process in which learning emerges from the interaction of an external feedback loop and an internal feedback loop. The external loop comprises feedback provided by an external source (e.g., teacher, digital system) that conveys information about task performance relative to predefined goals or criteria, while the internal loop refers to learners' own processes of monitoring, interpreting, and regulating their cognition, metacognition, and motivation. Central to the model is the assumption that feedback effects are not direct but mediated by learners' internal reference values, prior knowledge, and self-regulatory skills. Feedback functions are therefore multidimensional, encompassing cognitive (e.g., error identification), metacognitive (e.g., monitoring and strategy adjustment), and motivational (e.g., effort regulation) components. The model further emphasizes that discrepancies between internal and external feedback are common and that learning outcomes depend on how successfully learners resolve these discrepancies.

2.2. Feedback in digital learning technologies

Digital technologies have expanded the possibilities for providing timely, individualized, and scalable, elaborated formative feedback (see Fig. 1). Digital learning systems incorporate automated or semi-automated feedback mechanisms intended to guide learners' task performance and self-regulatory processes [53,54]. Across these systems, several core digital feedback types can be distinguished, typically aligned with Shute's [4] taxonomy: (1) Knowledge of Results (KR), indicating whether a response is correct or incorrect; (2) Knowledge of Correct Response (KCR), providing the correct solution; and (3) Elaborated Feedback (EF), which provides explanations, hints, worked examples, or strategy guidance.

Digital formative feedback is embedded in diverse learning systems that differ in the timing, granularity, and adaptivity of feedback [55]. Intelligent tutoring systems typically provide step-level feedback, enabling fine-grained corrective guidance during problem solving [40,53,56]. Learning analytics-based systems often deliver feedback at the task, course, or behavioral level by analyzing trace data and highlighting progress, engagement patterns, or self-regulatory gaps [45,57]. Automated writing evaluation (AWE) tools offer feedback on textual features, linguistic accuracy, or discourse organization [16]. Many digital learning systems also integrate formative mastery tests, where feedback is delivered after multi-item assessments and includes item-level explanations, goal-level summaries, and recommendations for next steps [12,58].

Across these systems, research has consistently shown that digital formative feedback improves learning, especially when it is informative and supports error correction (e.g., [41]). Meta-analytic work shows medium positive effects of digital feedback on learning outcomes, with elaborated feedback outperforming KR and KCR, especially for higher-order learning goals [8,11,13,39]. Brummer et al. [11] demonstrated that feedback focus is a central determinant of effectiveness: feedback targeting learning processes and underlying reasoning was consistently more effective than feedback focused solely on task correctness. At the same time, combining multiple feedback foci (e.g., task-level and process-level feedback) tended to reduce effectiveness, suggesting that excessive or poorly aligned elaboration may overload learners and diminish learning benefits.

Research also shows that metacognition, particularly calibration accuracy, monitoring, and self-regulated learning skills, mediate the effects of elaborated digital feedback on learning. Kuklick et al. [41] found that simple and more informative feedback types improved calibration accuracy, reducing over- or underestimation of performance. Improved calibration accuracy is assumed to reflect deeper engagement with errors in the feedback message resulting in deeper revisions and improved performance in subsequent learning tasks. At the meta-analytic level, research shows that elaborated feedback explicitly targeting metacognitive regulation showed particularly strong effects on learning performance [11,13].

Learners's emotional and motivational responses to corrective error feedback also play an important role in mediating the effects of elaborated feedback on learning performance. Fong et al. [6] demonstrated that instructional negative feedback, which provides guidance for improvement, was consistently associated with less negative motivational effects and outperformed non-instructional criticism. Criterion-referenced feedback was also less demotivating than normative, socially comparative feedback. Additionally, Kuklick et al. [41] found that KR had a negative effect on students' motivational state, whereas elaborated feedback had a less negative impact, suggesting that explanatory information can buffer motivational costs associated with error feedback. De Sixte et al. [133] showed that warm elaborated feedback, which combined cognitive information with motivational attributions (e.g., recognizing competence and attributing errors to controllable causes), led to substantially more revision activity than elaborated feedback alone, even though overall performance did not differ across conditions.

Characteristics of the educational context, the learning domain, and the types of the formative assessment tasks also moderate feedback effects. Both Van der Kleij et al. [13] and Brummer et al. [11] found larger effects of digital feedback in higher education than in primary education, and stronger effects in structured domains such as mathematics and science than in less structured domains. Moreover, Wang et al. [59] demonstrated that detailed explanatory feedback improved transfer performance only when students worked on constructed-response items, not multiple-choice items. These effects were mediated by longer feedback processing time, higher perceived usefulness of feedback, and increased germane cognitive load. Elaborated feedback reduced extraneous cognitive load and promoted productive cognitive effort, which in turn supported learning.

2.3. LLMs and prompt engineering

The rapid advancement of LLMs from 2017 (publication of the Transformer architecture) to late 2022 (publicly accessable general purpose, instruction-tuned LLMs) has opened new opportunities for applying AI in education, particularly for generating elaborated formative feedback. LLMs are advanced neural networks based on the Transformer architecture, trained on vast text corpora to understand, generate, and reason with human language [60]. They typically undergo two main training stages: first, pretraining on large general-purpose datasets to learn linguistic patterns and world knowledge, and second, fine-tuning on domain-specific data or via reinforcement learning from human feedback (RLHF) to specialize in particular tasks or improve alignment with user intent. These models can be deployed in different ways, either through cloud-based APIs that provide access to both proprietary and open-source LLMs, or as locally hosted or embedded components in educational tools and learning platforms.

Matarazzo and Torlone [61] grouped LLMs into major families that differ in architecture, capabilities, and accessibility. BERT represents encoder-only models designed for text comprehension, while T5 (encoder-decoder) unifies all NLP tasks in a text-to-text framework. The GPT series (decoder-only) introduced powerful generative and reasoning abilities, such as in-context learning and chain-of-thought (CoT) reasoning. LLaMA models prioritize efficiency and openness,

while Claude, Gemini, and Gemma emphasize ethical alignment, multimodality, and compact deployment. Proprietary models (e.g., GPT, Claude, Gemini) are typically accessed via APIs, whereas open-source models (e.g., LLaMA, Mistral, Gemma) can be locally hosted and customized, offering transparency and adaptability for educational use.

At first glance, the well-documented power of feedback for learning, the time-consuming nature of formative assessment in education, and AI technologies capable of processing human-written text appear to form a natural alignment. The provision of individualized and personalized learning—specifically, feedback tailored to student needs—is one of the most prominent and widely acknowledged expectations regarding how AI will improve education and learning in the future. Although LLM technologies have sparked substantial interest and optimism among educators and researchers worldwide, most developments remain in an experimental stage and are far from production-ready learning systems. Moreover, integrating LLMs into educational contexts entails a set of interrelated technical, pedagogical, and ethical challenges that operate across model, interactional, and institutional levels [134–136]. Literature on AI in education, particularly research on AI-assisted feedback [26], identifies several ethical risks associated with LLM-generated feedback outputs:

a) Hallucinations: A central technical limitation concerns the phenomenon of hallucinations, defined as the generation of fabricated, misleading, or factually inaccurate information presented with high linguistic fluency.
b) Algorithmic bias: LLMs may reproduce or amplify biases embedded in training data, leading to unfair, culturally insensitive, or systematically skewed feedback.
c) Overreliance: Students may become overly dependent on automated suggestions, potentially reducing critical thinking, self-monitoring, and independent revision skills.
d) Insufficient pedagogical integration: AI feedback systems may be deployed without adequate alignment to instructional goals, assessment criteria, or teacher mediation.
e) Privacy and data governance: At the institutional level, privacy and data governance challenges arise from the submission of student-generated content to external AI systems.

The literature proposes several responses to these broader pedagogical and ethical challenges of LLM use in education. Specifically for LLM-generated feedback, prompt engineering and fine-tuning are considered crucial [62]. Prompting involves structuring input instructions to achieve accurate, coherent, and task-aligned responses. Foundational techniques such as zero-shot and few-shot prompting rely on task descriptions and example-based cues, while more advanced strategies, e.g., Chain-of-Thought prompting, enhance reasoning by promoting stepwise or multi-path exploration (e.g., [63]). Fine-tuning goes further and involves the additional training of a pre-trained model on curated task-specific or domain-specific data, thereby aligning the model's behavior with desired outputs [64]. Traditional supervised fine-tuning uses explicit input-output pairs, while instruction fine-tuning generalizes this approach by exposing the model to a wide range of task instructions to improve controllability (e.g., retrieval augmented generation). Empirical evidence suggests that these techniques causally reduce both hallucinations and biased outputs by grounding generation in explicit instructions or retrieved evidence [137].

2.4. Research on AI-generated feedback

The emergence of LLMs mark a potential turning point in the evolution of feedback in digital learning systems [65]. Unlike earlier NLP systems, LLMs can process text-based student answers to higher-order learning tasks and assignments. This capability allows automated, formative feedback to move beyond surface-level correctness checks toward criterion-based guidance that explains errors, recommends revisions, and supports self-regulated learning [18,19,29]. As a result, the range of potential learning domains and task types that can be incorporated into feedback research is substantially expanded [127].

The integration of LLMs into feedback systems—and the investigation of feedback uptake and effects—offers significant potential to deepen our understanding of how external and internal feedback loops interact (see ITFL model; [52]). However, our understanding of this rapidly developing field remains limited. Consequently, there is a clear need to systematically review, monitor, and integrate findings from this evolving body of research. Broader reviews (e.g., [66,67]) offer valuable overviews of AI in education but largely examine studies published before the advent of generative, LLM-based feedback systems. We found five systematic literature reviews with a more specific focus, summarizing research on AI-assisted feedback mechanisms and results (Table 1).

Shi and Aryadoust [27] provide one of the most comprehensive pre-LLM syntheses by reviewing 83 SSCI-indexed empirical studies on AI-based automated written feedback (1993 to 2022). Their analysis maps the field in terms of research contexts, system types, feedback categories, study designs, and findings. However, most of the reviewed studies focused on pre-LLM feedback technologies, e.g., form or grammar-oriented feedback in tertiary EFL/ESL writing, using systems like Pigai, Criterion, and Grammarly. Three major themes emerged: (a) system performance with mixed accuracy and validity, (b) user perceptions and engagement is generally positive but variable, and (c) effects on writing outcomes are typically positive for accuracy but less conclusive for higher-order writing. While this review provides a strong historical foundation, its scope is limited. First, it combines rule-based systems, traditional NLP methods, and emerging AI tools under a single umbrella. Second, it covers research only up to 2022, it does not systematically examine empirical work following the public availability of instruction-tuned LLMs in late 2022. Another limitation lies in the search strategy. The search terms focused on variations of "automated writing evaluation," "automated writing feedback," "automated essay evaluation," and "computer-based writing feedback." Terms such as "large language model," "LLM," "ChatGPT," or "generative AI" were not included in the core search string.

Urzúa et al. [28] conducted a systematic review of 12 empirical studies (2021–2025) examining the effects of AI-assisted feedback via generative chat on academic writing in higher education. Focusing specifically on university students' written text production, the review synthesized predominantly mixed-methods research comparing generative chat feedback with teacher or hybrid feedback models. The findings indicate largely positive effects on writing quality, self-regulation, motivation, proactivity, and reduced anxiety. At the same time, the review highlights limitations related to inaccuracies, overreliance on AI, insufficient user training, and the need for pedagogical guidance. However, this review focuses solely on higher education contexts, specifically academic writing tasks, and includes a relatively small number of studies.

Yildiz Durak and Onan [29] conducted a broader review of 91 studies (2014–2024) on AI-based feedback across educational settings, noting a sharp rise in LLM-related publications from 2023 onward. Their synthesis shows that most studies employed LLMs in higher education contexts and used experimental designs grounded in constructivist and self-regulated learning frameworks. They report generally positive effects on students' motivation, engagement, self-regulation, and academic performance, and identify challenges such as bias, emotional insensitivity, and overreliance on AI tools. However, the review remains largely descriptive regarding LLM technology. It identifies ChatGPT as dominant tools but provides little insight into how these models are prompted, fine-tuned, or evaluated. It touches only briefly on prompting approaches and does not synthesize methodological best practices or quality criteria for LLM-generated feedback.

Kaliisa et al. [30] conducted a meta-analysis comparing AI-generated and human feedback across 41 studies with 4813 students. They report

Table 1
Systematic literature reviews on AI-assisted feedback research.

Review	Time frame	Studies included	Review scope
Shi & Aryadoust [27]. A systematic review of AI-based automated written feedback research	1993 to 2022	83	Systematic qualitative review of research involving feedback from natural language processing technologies for automated writing evaluation, automated writing feedback, e.g., Criterion, Pigai, Google Docs, Writing Pal, Grammarly
Urzua et al. [28]. Effects of AI-Assisted Feedback via Generative Chat on Academic Writing in Higher Education Students: A Systematic Review of the Literature	2021 to 2025	12	Systematic qualitative literature review focusing on AI-assisted feedback technologies in higher education writing. Researched technologies specifically include ChatBots used by students for writing revisions
Yildiz Durak & Onan [29]. A systematic review of AI-based feedback in educational settings	2014 to 2024	91	Systematic qualitative review of research on AI-based feedback in education, focusing on trends, conceptualizations, methods, impacts, and challenges; feedback technologies involve machine learning, natural language processing for automated writing evaluation, and LLMs
Ba et al. [26]. Unraveling the mechanisms and effectiveness of AI-assisted feedback in education: A systematic literature review	2014 to 2023	129	Systematic qualitative literature review of research on a broad range of AI-assisted feedback technologies in education, including generative adversarial networks, large language models, machine learning models, neural networks, intelligent tutoring systems, expert systems, and chatbots. The review analyses and summarizes research from a learner-centered motivation theory perspective.
Kaliisa et al. [30]. How does artificial intelligence compare to human feedback? A meta-analysis of performance, feedback perception, and learning dispositions	2010 to 2024	41	Quantitative meta-analysis comparing cognitive, metacognitive, and motivational learning outcomes of AI-generated feedback and human (teacher) feedback; reviewed feedback technologies involve automated writing evaluation (AWE) and large language models (LLMs)
This review: Authors (2026). LLM-generated formative feedback: A qualitative systematic literature review	2022 to 2026	48	Systematic qualitative review of research on LLM-generated formative feedback in educational contexts; reviewed feedback technologies involve publicly available, general purpose, instruction-tuned large language models since late 2022

no significant differences in learning performance or feedback perceptions between AI and human feedback, recommending hybrid human–AI models that combine efficiency with empathy. Yet their analysis aggregates a wide variety of feedback technologies—from early automated writing evaluation systems to new LLM-based tools—into a single "AI feedback" category. Although the review explicitly mentions ChatGPT as an example of contemporary AI feedback, LLM-based feedback constitutes only a minor subset of the analyzed studies. Consequently, the meta-analytic effects primarily reflect older, rule-based or statistical systems and do not differentiate the distinctive affordances and risks of LLMs, such as dialogic interaction, contextual adaptability, or hallucination.

Eventually, Ba et al. [26] conducted a systematic review of 129 empirical studies (2014–2023) examining AI-assisted feedback in education. The review analyzed bibliometric trends, feedback mechanisms, and associations with learner perceptions, actions, and outcomes from a learner-centered motivation theory perspective. The included technologies encompassed a broad spectrum of AI approaches, including machine learning, deep learning, neural networks, intelligent tutoring systems, chatbots, generative adversarial networks, and large language models such as GPT and BERT. Findings indicate a sharp rise in AI-assisted feedback research after 2018 and predominantly positive effects on learner satisfaction, perceived learning, self-regulation, and academic performance. However, because this review encompasses such a broad spectrum of AI-assisted feedback technologies, it does not specifically synthesize findings from studies that explicitly investigate feedback from instruction-tuned, general-purpose LLMs that have become widely available since 2022.

3. Research questions

While systematic reviews [26–30], have summarized developments in AI-based and automated feedback more broadly, LLM-generated formative feedback systems have not been addressed systematically. These reviews provide valuable overviews but lack detailed analyses of the generation, accuracy, and cognitive or non-cognitive effectiveness of LLM-generated formative feedback since the advent of instruction-tuned LLMs in late 2022.These LLMs have enabled the generation of coherent and context-sensitive feedback texts for diverse student-written responses without requiring extensive domain-specific pre-training on annotated educational datasets. This paper, therefore, aims to critically examine the rapidly growing yet fragmented body of research on LLM-generated formative feedback in educational settings from 2022 to 2026. This time window captures the phase in which instruction-tuned LLMs became practically usable for educational feedback, coinciding with the sharp rise of publications concerning LLMs in education after 2022 [68]. We pose these research questions:

RQ1: How is research on LLM-generated formative feedback theoretically framed and what are the main lines of investigations?

- 1.1 What terminology, besides LLM-based or LLM-generated formative feedback, is used in the research field to describe feedback generated by large language models?
- 1.2 How do the reviewed studies theoretically conceptualize learning and feedback, and how is the role of LLM-generated formative feedback situated within these theoretical frameworks?
- 1.3 What are the dominant research questions within the existing literature on LLM-generated formative feedback?

RQ2: What are the methodological and contextual characteristics of research on LLM-generated formative feedback?

- 2.1 What types of research designs and methodological approaches have been implemented?
- 2.2 In which learning domains has research on LLM-generated formative feedback been conducted, and for which tasks or student assignments has such feedback been provided?

2.3 At which educational levels (primary, secondary, or tertiary education) have studies on LLM-generated formative feedback been conducted?
2.4 What types of samples are investigated, distinguishing between human participants (e.g., students, teachers) and artefactual data (e.g., essays, responses, or LLM outputs)?
2.5 In which instructional settings has research on LLM-generated formative feedback been carried out?

RQ3: What are the relevant findings of the reviewed studies?

3.1 What types of LLMs are used in these studies, and how are they configured?
3.2 What kinds of prompting strategies are employed or investigated?
3.3 What are the main empirical results along the dominant research questions?
3.4 How do studies address pedagogical and ethical challenges of LLM-generated formative feedback?

While RQ1 and RQ2 describe the methodological designs and implementation contexts of the reviewed studies, RQ3 synthesizes the reported outcomes and identifies patterns linking LLM configurations, prompting strategies, and implementation approaches to results and observed challenges. The purpose of RQ3 is to derive practice-relevant implications by examining which types of models and prompting strategies yielded which cognitive and non-cognitive effects, and under what boundary conditions problems such as hallucinations, bias, or overreliance occurred.

4. Methods

4.1. Review design

To synthesize the rapidly growing but conceptually diverse body of research on LLM-based feedback, we employ a qualitative systematic review approach [69], which retains the rigor of systematic identification and appraisal, but relies on interpretive, thematic comparison when statistical aggregation is not appropriate due to variation in study designs, outcomes, or theoretical framings. This approach is therefore well suited for LLM-based feedback research, where studies span multiple domains, rely on diverse evaluation metrics, and are not amenable to quantitative meta-analysis. We further adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guidelines [70], which provide a structured and replicable framework for documenting search, screening, and selection procedures.

4.2. Eligibility criteria

This review applied clearly defined inclusion and exclusion criteria to ensure that only empirical studies providing authentic evidence on LLM-generated feedback in educational contexts were analyzed. Studies were included if they met all of the following criteria:

(a) investigated formative feedback generated by general purpose, instruction-tuned large language models (LLMs), transformer-based generative models capable of text generation,
(b) involved real learners or teachers in authentic educational settings (K–12, higher education),
(c) reported empirical data on feedback quality, internal feedback processing or feedback effects on cognitive, metacognitive, or motivational learning outcomes,
(d) were peer-reviewed journal articles published in English between 2022 and 2026, reflecting the era when generative AI first enabled scalable, natural-language feedback with instruction-tuned LLMs,
(e) provided sufficient methodological detail for data extraction, including model specifications, prompting strategies, research design, sample characteristics, and reported outcomes.

Exclusion criteria and examples:

(a) Survey or perception studies without empirical data on actual feedback use, for example Henderson et al. [71], who examined students' general attitudes toward ChatGPT feedback without specific feedback instances as context.
(b) Non-comparative explorations of AI tools without a focus on feedback mechanisms, such as Xu et al. [72], who described ChatGPT use in tutoring contexts but not its feedback content or effects.
(c) Conceptual or theoretical papers lacking empirical results, for instance Stamper et al. [65], who proposed integrating LLM-generated feedback into intelligent tutoring systems without reporting data.
(d) Usability and system-design studies addressing interface features rather than learning or feedback outcomes, e.g., [73,74], who evaluated a GPT-based code-review tool from a usability perspective only.
(e) Specific training systems that did not isolate feedback effects, such as badminton training [75], nursing education, clinical practice for surgeons or radiologists [76].
(f) Studies relying on traditional NLP-based feedback pipelines with transformer models trained on student text corpora and rubrics, rather than generating feedback through prompt-based interaction with a general-purpose LLM (e.g., [138,139]).
(g) Multimodal, gamified, or immersive environments where LLM-generated formative feedback was not separable from other intervention elements [77–79].
(h) Studies that examined LLM-generated feedback for summative purposes, e.g., essay scoring, grading, or final examinations (e.g. [80]).
(i) Papers from research groups with consecutive results reporting across years.

4.3. Literature search strategy and selection process

The selection process followed the PRISMA flow of identification, screening, eligibility, and inclusion (Fig. 2). Five major databases were systematically searched: ERIC, Scopus, IEEE Xplore, SpringerLink and ScienceDirect. These databases cover education, psychology, and technology research. We conducted our first database search in November 2025 using the following Boolean string applied to titles, abstracts, and keywords:

("llm-based feedback" OR "llm-generated feedback" OR "large language model feedback" OR "ai-based feedback" OR "ai-generated feedback" OR "artificial intelligence" OR

"generative ai feedback" OR "gpt feedback" OR "gemini feedback") AND ("education" OR "learning" OR "instruction") AND (publication year: 2022–2026)

Filters were applied for peer-reviewed journal articles and the English language. Search results were as follows: ERIC – 111 documents, Scopus – 191 documents; IEEE Xplore – 59 documents and ScienceDirect – 62 documents. ERIC entries from the year 2021 were removed manually. All bibliographic data were exported, merged, and deduplicated, resulting in 268 unique records.

For the revision of this paper in February 2026, we decided to broaden our search strategy to include additional relevant publications. We therefore combined the terms "LLM," "large language model," "AI," and "artificial intelligence" with "feedback" in the title, and "education," "learning," or "instruction" in the title, abstract, or keywords. This expanded search substantially increased the number of retrieved records across databases: ScienceDirect (151 documents), ERIC (148

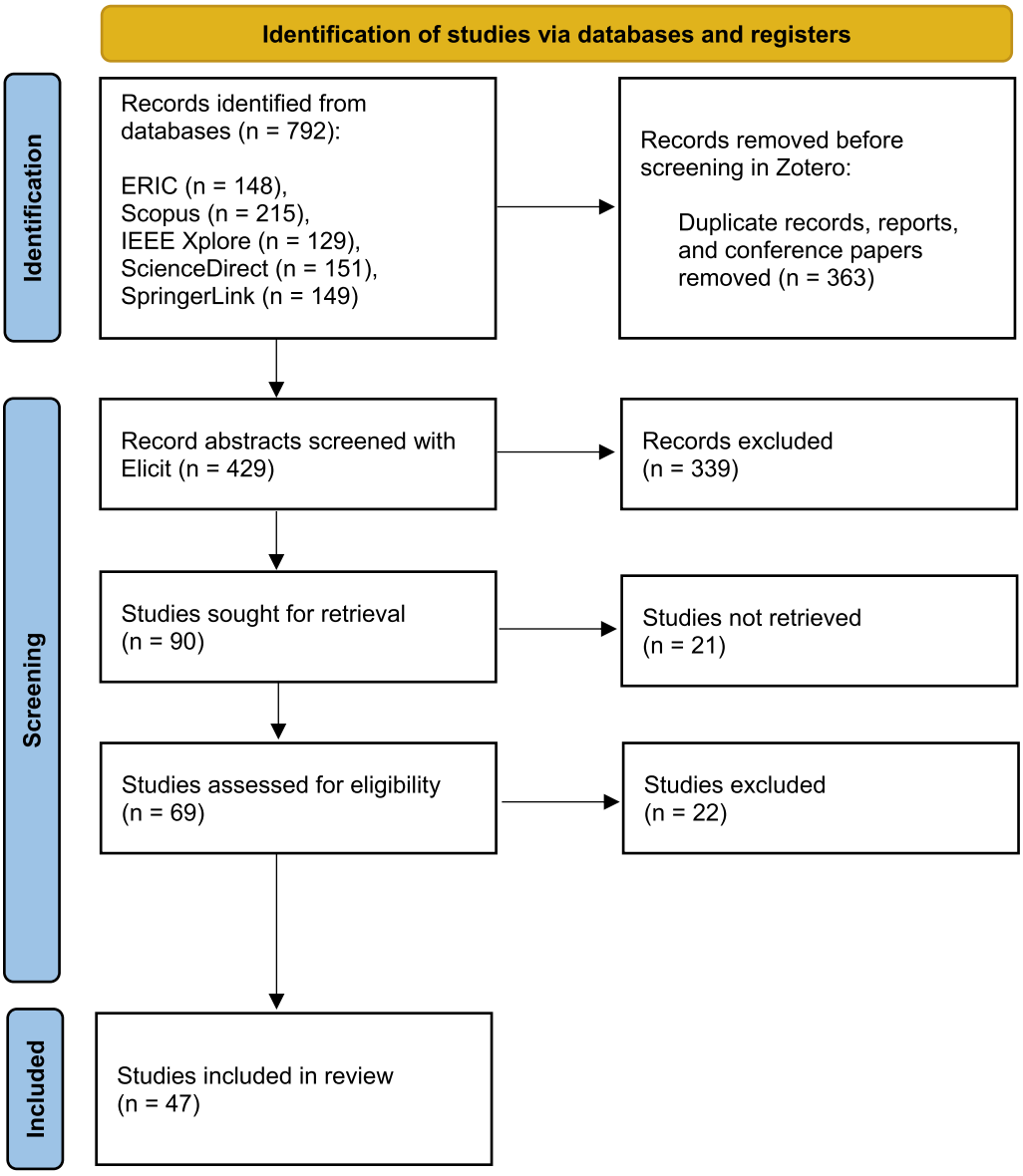

Fig. 2. PRISMA flow diagram of article search, screening, and selection.

documents), Scopus (215 documents; limited to Social Sciences, Arts and Humanities, Psychology, and final peer-reviewed articles in English), IEEE Xplore (129 documents), and SpringerLink (149 documents; limited to subject areas including AI, instructional design, higher education, e-learning, learning and instruction, NLP, language teaching and learning, education science, educational research, and pedagogy).

All database records were exported and imported into a Zotero collection for deduplication. After removing duplicates, 638 unique documents remained. Following the exclusion of conference papers and other publication types that did not meet our inclusion criteria, the dataset was reduced to 429 peer-reviewed journal articles.

These documents were then uploaded to a collection in the Elicit Pro version (elicit.org) to conduct a semi-automated screening of abstracts. We prompted Elicit to identify all papers meeting our predefined inclusion criteria. Proof-of-concept studies suggest that Elicit performs comparably to human reviewers in literature search and data extraction tasks [81,82]. We downloaded the resulting Excel file generated by Elicit, which provided detailed information on whether each abstract met the inclusion criteria. Subsequently, we manually reviewed all 90 papers identified by Elicit as eligible and ultimately selected 69 studies for which the full texts were retrieved and examined in detail.

After detailed full-text evaluation, 22 articles were excluded for lacking empirical feedback data, focusing on non-educational domains, or using synthetic datasets (see exclusion criteria). In some cases, we found papers from research groups with consecutive results reporting across years (e.g., [24,83]). We then used the most recent or comprehensive report. The final 47 studies were retained for qualitative synthesis (see Table 1). The publication year of the included studies span from 2023 to 2026 (2023: 3 studies, 2024: 13 studies, 2025: 24 studies, 2026: 7 studies). The reviewed research was published from research groups across the world, with most papers coming from USA, China, and Germany.

4.4. *Data extraction and qualitative content analyses*

Data extraction proceeded in several stages. First, the lead author created a preliminary extraction table based on a subset of studies, aligned with the research questions. Extracted fields covered study metadata, theoretical background, research questions, technical configuration, research design, learning domain, education level, sample characteristics, instructional setting, LLM specification, prompting approach, empirical findings, ethical risks and challenges (see Appendix

A). Second, the two co-authors independently audited 30 % of entries to verify accuracy, completeness, and consistency. Third, the team discussed the required depth of extraction (e.g., level of methodological detail, specificity of prompting information, and granularity of reported outcomes). Based on this consensus, the lead author completed extraction for all remaining studies. Co-authors again checked 30 % of these entries. This process yielded the comprehensive data extraction table, which served as the basis for inductive synthesis. The 47 empirical studies included in sum 121 distinct research questions. Theory background, methodological characteristics, and technological specifications were extracted on the study level. Table 1 presents a condensed overview of all includes studies and their key characteristics. We extracted and analyzed research results of the reviewed studies on the level of research questions (see Appendix A).

We applied a combination of deductive and inductive qualitative content analysis following the principles outlined by Mayring [84], depending on the nature of each research question. For research questions based on predefined or clearly structured dimensions (e.g., terminology used to describe LLM-generated feedback, educational levels, types of LLMs), we applied deductive coding. In these cases, we either counted occurances (terminology) or categories were established prior to analysis, and the procedure primarily involved systematic data extraction and classification of studies into these predefined categories. For research questions that were exploratory in nature and not grounded in an a priori coding framework (e.g., theoretical conceptualizations, research results), we employed inductive qualitative content analysis. Here, categories were developed iteratively from the extracted data to capture recurring patterns and meaningful thematic structures across studies.

4.5. Ethical considerations

This study involved the synthesis of published research and did not include human participants or the collection of primary data. Therefore, institutional ethical approval was not required. Nonetheless, ethical considerations arise in relation to the topic itself. Many of the included studies involved learners and teachers interacting with LLMs, and we report their findings with careful attention to issues of privacy, fairness, bias, and potential harms. Throughout the review process, no copyrighted or identifiable human data were used beyond what was available in published sources.

5. Results

The results are structured according to the research questions. For RQ1.1, RQ2.2, and RQ2.3, we report descriptive data. For all other research questions, we present the categories that emerged from our inductive qualitative content analysis. Within each category, we provide illustrative examples of studies that, in our assessment, most clearly and comprehensively represent the respective category. These examples are intended to exemplify the identified patterns rather than to offer an exhaustive listing of all studies assigned to a given category. Readers interested in more detailed information or in specific studies are referred to Table 1 and Appendix A.

5.1. Terminology, theoretical framing and main lines of investigations (RQ 1)

5.1.1. Terminology

What terminology is used in the research field to describe feedback generated by large language models (RQ 1.1)? Table 2 summarizes the frequencies of the used terms in the reviewed studies' titles, abstracts, and keywords. The terms reflect our search string. The most frequent terms combine "artificial intelligence" or "large language model" with "generated" and "feedback". However, many studies more specifically used ChatGPT or GPT in combination with feedback, reflecting that most studies used this specific LLM in their research (Table 3).

5.1.2. Theoretical background of the reviewed studies

How do the reviewed studies conceptualize learning, and how is the role of LLM-generated feedback situated within their theoretical frameworks (RQ 1.2)? Across the reviewed studies, we found 5 broad categories of theoretical frameworks for LLM-based feedback research. Most studies implemented several of these theoretical frameworks in their research papers.

a) Formative feedback: The most prevalent theoretical foundation across the reviewed studies is general formative feedback theory [2,4]. These models define effective formative feedback as information that helps learners answer three core questions: Where am I going? How am I going? and Where to next? Many studies used these questions as a coding scheme or evaluation rubric for assessing the pedagogical value of LLM-generated formative feedback [24,85-87,132]. Other works combined Hattie and Timperley's structure with formative assessment theory to justify the potential of LLMs to provide adaptive, individualized learning support [88–90].
b) Self-regulated learning and motivation: A second major category links feedback to self-regulated learning (SRL) and motivational–affective theories. These perspectives explain how formative feedback triggers internal regulation processes, sustains engagement, and shapes learners' emotions. Studies grounded in SRL theory [34,91] conceptualize feedback as part of a cyclical system in which learners monitor their progress and adjust strategies accordingly [25,89,92,131]. Related frameworks such as self-determination theory [93], expectancy-value theory [94], and control-value theory of achievement emotions [95] provide a motivational lens, emphasizing LLM-generated feedback as an autonomy-supportive and emotionally meaningful input that can enhance enjoyment, pride, and perceived competence [21,96,140]. Some studies draw on the concept of feedback literacy [35] which also emphasized self-regulated learning and shifts attention from the feedback message to the learner's ability and willingness to engage productively with it (e.g., [23,97,141]).
c) General learning theories: Some papers also referenced general learning theories, such as the Zone of Proximal development framework from Vygotski which is often referred to as socio-constructivist learning or development theory (e.g., [142–144]). Feedback is seen within this theory framework as an important ingredient in scaffolding and challenging learners to get to their next level of competence which is slightly above the actual level of knowledge of competence. Other studies (e.g., [89,98]) referred cognitive learning perspectives such as cognitive load theory [99] or self-explanation theory [100]. These theories justify the design of elaborated formative LLM feedback that minimizes extraneous processing and deepens comprehension.
d) Domain-specific frameworks: Several studies integrate domain-specific theories of learning to contextualize how LLM-generated formative feedback operates in particular subjects. In writing and language learning, researchers often reference Second Language Acquisition (SLA) and Written Corrective Feedback (WCF) theories, emphasizing interactional feedback, noticing, and learner uptake [101–103,143,145]. Farrokhnia et al. [132], for example, grounded their study in Toulmin's model of argumentation, which structures argumentative writing around components such as claims, arguments, counterarguments, evidence, and conclusions. In science learning, studies such as Seßler et al. [24] or Wan and Chen [104] rely on constructivist learning and deliberate practice frameworks, conceptualizing feedback as a means of conceptual change through active problem solving and reflection.
e) Technology acceptance and Human-AI frameworks: Other studies complemented pedagogical theories with technology acceptance and evaluation frameworks drawn from human–computer interaction

Table 2
Studies included in the review.

Study	Country	Learning domain and assignments	Education level	Human sample	LLMs with specifications
Algobaei & Alzain [143]	Saudi Arab	EFL: Writing short academic paragraphs	Tertiary	32 students; 3 experienced EFL instructor raters	GPT-4
Alsaiari et al. [140]	Australia	Web design: Assessment tasks	Tertiary	381 students	ChatGPT 3.5 via integration into the RiPPLE platform
Asadi et al. [142]	Iran	EFL: Argumentative essay writing	Tertiary	68 students; 8 teachers interviewed	ChatGPT (no specification in paper, most likely GPT 4.0)
Banihashem et al. [148]	Nether-lands	Food sciences: Argumentative essay writing	Tertiary	74 students	ChatGPT (no model version specified)
Baral et al. [88]	USA	Math: Open-ended problem-solving tasks	K-12	About 5000 student responses	GOAT model fine-tuned, SBERT-Canberra, and GPT-4 Turbo
Çağlar-Özhan [150]	Turkey	Computer technology and instructional design	Tertiary	48 students, 24 instructors	ChatGPT 3.5
Chan et al. [96]	Hong Kong	English language education: Academic writing	Tertiary	918 students, 16 students interviewed	GPT-3.5-turbo
Dai et al. [85]	Australia	Data science education: Writing project proposals	Tertiary	103 students	ChatGPT (no model version specified)
Er et al. [114]	Türkiye	Computer science education, Java programming	Tertiary	35 students	ChatGPT-4
Escalante et al. [101]	USA	EFL: Academic writing	Tertiary	48 students	ChatGPT-4
Farrokhnia et al. [132]	Iran	Educational sciences: Argumentative essay writing	Tertiary	70 students and one experienced teacher	ChatGPT-4o
Hao et al. [147]	China	EFL: Argumentative essay writing	Tertiary	64 students, 2 experienced writing teachers	ChatGPT 4.0
Hofmann et al. [146]	Germany	Teacher education: Reflective writing about pedagogical case studies	Tertiary	85 pre-service teachers	ChatGPT 4.0
Jacobsen et al. [109]	Germany	Teacher education: Writing about learning goals in education	Tertiary	20 / 153 pre-service teachers	ChatGPT-4, Claude 3, Gemini Advanced
Jamshed et al. [102]	India	EFL writing and grammar skills	K-12	132 students	ChatGPT 3.5 via mobile app
Jia et al. [86]	USA	Programming education: Code refactoring, feature implementation, and testing	Tertiary	484 students	BART-large-CNN model, ChatGPT-4 via OpenAI API
Jia et al. [20]	USA	Programming education: Code refactoring, feature implementation, and testing	Tertiary	82 students, 28 survey responses	BART-large-CNN model, ChatGPT-4 via OpenAI API
Jovic et al. [120]	USA	Communication course: Argumentative and reflective essay writing	Tertiary	12 students	ChatGPT-4.5, Claude-3.7 Sonnet
Jürgensmeier & Skiera [106]	Germany	Marketing analytics: Project with coding, statistical analysis, and economic reasoning	Tertiary	243 / 20 / 43 students	ChatGPT-3.5 and ChatGPT–4–8k, Claude 3 Opus, Claude 3 Sonnet, Gemini 1.0 Pro, Gemini 1.5 Pro, Mistral Large
Kerslake et al. [115]	New Zealand	Programming education: Code comprehension through Explain in Plain English (EiPE) tasks	Tertiary	21 students	ChatGPT-3.5 via API access with automated prompting
Kinder et al. [89]	Germany	Teacher education: Justifying pedagogical decisions about fictitious students	Tertiary	269 pre-service teachers	GPT–4–1106-preview via API
Li et al. [149]	Different regions	English language arts and Science: Open-ended short-answer responses	K-12	56 educators	ChatGPT 4.0, BERT-base-uncased model
Liebenow et al. [131]	Germany	EFL: Argumentative essays based on TOEFL-style prompts	K-12	459 students	GPT-3.5-turbo via API
Liu et al. [141]	China	EFL: Writing academic literature reviews	Tertiary	71 students	ERNIE Bot 4.5 Turbo
Lo et al. [110]	Hong Kong	EFL: Argumentative essay writing	Tertiary	1102 students, 18 students interviewed	ChatGPT-3.5
Maqbali et al. [144]	Oman	EFL: Essay writing	Tertiary	70 students	Copilot AI
McNichols et al. [22]	USA	Math: Open-ended linear equations problems from a middle school math ITS	K-12	Problem solutions from 26,845 students	Mistral-7B-v0.1, ChatGPT-3.5-turbo-1106 (fine-tuning enabled)
Meyer et al. [21]	Germany	EFL: Argumentative essay writing	K-12	459 students	ChatGPT-3.5-turbo via GPT Playground
Na Nongkhai et al. [129]	Thailand	Programming education: Coding exercises	Tertiary	87 students	GPT-4 via API integration within a LangChain-based architecture
Nazaretsky et al. [116]	Switzer-land	Educational technology and pedagogy: Writing tasks	Tertiary	457 students	ChatGPT (no model version specified)
Nguyen et al. [98]	USA	Math: Decimal numbers, decimal magnitude and place value	K-12	117 students	ChatGPT-3.5 via web interface and API automation
Noroozi et al. [151]	Nether-lands	Food science: Writing argumentative essays on controversial topics	Tertiary	50 students	ChatGPT (no model version specified), Grammarly
Rüdian et al. [23]	Germany	Teacher education: Argumentative writing on a poem and its implications for language and grammar	Tertiary	32 students	Llama 3 70B (open-source foundation model)
Ruwe et al. [97]	Germany	Teacher education: Argumentative writing and reasoning	Tertiary	169 students	ChatGPT (no model version specified)
Scholz et al. [117]	Germany	Programming education: Python code exercises	K-12	8 computer science teachers	ChatGPT-4o and GPT-4o mini (2024 versions) in a multi-agent system

(*continued on next page*)

Table 2 (*continued*)

Study	Country	Learning domain and assignments	Education level	Human sample	LLMs with specifications
Seßler et al. [24]	Germany	Science education: Experimentation protocols	K-12	37 students, 11 science teachers, 5 science education experts	GPT-3.5-turbo-0125, accessed via API
Soyoof et al. [130]	Iran	EFL: Grammar, article usage	K-12	40 students	ChatGPT 3.5
Steiss et al. [87]	USA	History education: Students wrote source-based argumentative essays in history	K-12	198 students	ChatGPT 3.5
Sun et al. [92]	China	Programming education: Python exercises.	Tertiary	63 students	GPT-3.5 integrated via API
Thomas et al. [111]	USA	Teacher education: Core tutoring competencies	Tertiary	885 college-student tutors	GPT-3.5-turbo
Tran [145]	Vietnam	EFL: Academic opinion essay writing	Tertiary	14 students, one writing instructor	Gemini (model version not specified)
Wan & Chen [104]	USA	Science education: Conceptual understanding of Newton's laws	Tertiary	85 student, 4 instructors	GPT-3.5 Turbo
Weidlich et al. [25]	Switzer-land	Educational sciences: Scientific writing and argumentation	Tertiary	90 students	ChatGPT GPT-4 (2024) via custom GPT
Xavier et al. [128]	Brazil	Science education: Chemical concepts assessed through open-ended short-answer questions	K-12	60 students	DeepSeek V-3 integrated into the Tutoria platform
Yu & Xie [103]	China	EFL argumentative essay writing	K-12	60 students, 4 English teachers	ChatGPT-4o
Zhang et al. [119]	USA	Programming education: Java programming assignments in an introductory course	Tertiary	58 students	ChatGPT via API
Zhuang et al. [112]	China	EFL: Picture-cued descriptive writing tasks	K-12	386 students, 7 schools	Fine-tuned multimodal LLM (not specified)

Table 3
Frequency of LLM-based terminology in the reviewed studies.

Terms in titles, keywords, and abstracts	Frequency
AI chatbot-assisted feedback	1
AI feedback	8
AI-driven feedback	1
AI-generated feedback	17
AI-mediated feedback	1
ChatGPT feedback	5
ChatGPT-provided feedback	1
ChatGPT-delivered feedback	1
ChatGPT-generated feedback	6
GPT-3.5-turbo-generated feedback	1
GenAI feedback	8
GenAI-generated feedback	2
GenAI-assisted feedback	1
GenAI-based feedback	1
LLM feedback	3
LLM-based feedback	4
LLM-driven feedback	1
LLM-assisted feedback	1
LLM-generated feedback	13

and AI ethics. The Technology Acceptance Model (TAM) [105] and subsequent extensions are used to assess perceived usefulness, ease of use, and attitudes toward AI feedback systems in both higher education and language learning contexts [102,106]. Ethical perspectives on AI transparency, bias, and human-in-the-loop validation frequently accompany these frameworks to ensure that automation does not undermine fairness or pedagogical integrity [24,86]. From a technological angle, researchers also use Human-AI collaboration and AI literacy frameworks [107,108] to discuss prompt engineering and educator oversight as critical competencies for ensuring pedagogically sound AI feedback [109].

5.1.3. Categories of research questions

What are the dominant research questions within the existing literature on LLM-generated feedback (RQ 1.3)? We found that research questions on LLM-generated formative feedback cluster into five main categories. We also saw that most research papers include several research questions and therefore may be assigned to multiple categories in this section.

a) Effects on learning outcomes: A majority of the reviewed studies examine cognitive, metacognitive or motivational effects of LLM-generated formative feedback. In writing and language learning, it is asked whether AI feedback enhances text quality and task motivation [21,96,110,145], or grammar skills (e.g., [102,130]). In teacher education and STEM, studies investigated how LLM-generated feedback supports reasoning and problem-solving [89,111,112]. Many of the reviewed papers studied metacognitive or motivational effects, e.g., learners' self-assessment accuracy [131] or critical reflection skills [146]. Most of the studies compare effects of LLM-generated feedback with other feedback sources, e.g., teachers or peers (e.g., [141,146]), or other feedback conditions, e.g. no feedback, only scoring feedback (e.g., [21]), or other types of adaptive digital feedback [129].
b) LLM-generated feedback quality: These studies evaluate feedback quality and reliability compared to human benchmarks, focusing on specificity, constructiveness, tone, clarity, and rubric alignment (e.g., [24,85,87,113]). Most of the research questions in this category are focusing on the differences between teacher-generated and AI-generated feedback (e.g., [132]). Fewer studies, e.g. Banihashem et al. [148], investigated the quality of LLM-generated feedback vs. peer feedback.
c) Learner and teacher perceptions and usage of LLM-generated feedback: These studies ask how students and teachers perceive and use LLM-generated feedback, focusing on usefulness, fairness, trust, and engagement (e.g., [101,114]). Alsaiari et al. [140], for example, investigated how emotionally enriched AI-generated feedback affects student engagement with the feedback messages. Other studies in this category investigate how learners read and apply feedback [89,115,151]. Few studies in this category explored differences between students who deliberately chose to use LLM-generated feedback compared to those who rejected using it [111,151].
d) Feedback source and disclosure effects: A smaller line of research examines how learners react to knowing or not knowing the feedback source. These studies ask if students can distinguish AI from human feedback and how disclosure of the feedback source influences trust and fairness [23,97,116].
e) Human-AI feedback systems: Studies that fall in this category focus on integrating LLMs into educational systems and balancing automation with teacher oversight [88,106,115,117]. Li et al. [149], for

example, asked to what extent LLM-based assessment insights support educators in writing their own feedback to students' written answers. These works explore conditions under which LLMs serve as pedagogically credible feedback partners.

5.2. Methodological and contextual characteristics (RQ 2)

5.2.1. Types of research design and study procedures

What types of research designs and methodological approaches have been implemented (RQ 2.1)?

a) Experimental or quasi-experimental studies comparing LLM-generated feedback to human feedback: Multiple classroom-based RCTs examined causal effects of LLM-generated formative feedback on revision quality, writing performance, reasoning accuracy, and affective outcomes [21,89,114,118,131]. Additional quasi-experimental studies implemented structured pre–post or cluster-randomized designs in authentic educational settings [144, 146].
b) Experimental or quasi-experimental model comparisons: Some studies compared different LLM systems without including human feedback as a control [22,88]. These analyses often relied on large-scale corpus comparisons. Other work audited hallucinations or contrasted prompt-driven and fine-tuned systems [86]. Multi-condition quasi-experimental designs also compared adaptive, GenAI-based, and hybrid feedback modes across parallel intervention groups [129].
c) Comparing different prompting strategies: Several studies experimentally manipulated prompt design within the same LLM. Within-subjects and between-subjects designs compared general, grammar-focused, or EFL-tailored prompts, as well as neutral versus emotionally enriched feedback [140,143]. Other research contrasted zero-shot and chain-of-thought prompting against teacher feedback [132].
d) Perception and disclosure experiments: A set of studies examined how source transparency and provider information influenced feedback perception. Blind–disclosure designs assessed changes in perceived helpfulness and motivational quality once feedback was revealed as AI-generated [23]. Factorial experiments further analyzed trust, credibility, and fairness when manipulating feedback provider and transparency [97,116].
e) Comparative feedback quality studies: Several contributions focused on evaluating the pedagogical quality of LLM-generated feedback relative to human feedback. Large-scale grading accuracy and fluency analyses were conducted using labeled datasets [98]. Other studies employed content-analytic and rubric-based evaluations with blinded raters to assess dimensions such as accuracy, constructiveness, tone, and clarity [24,85,87].
f) Human–AI interaction studies: Finally, qualitative and mixed-method studies investigated how learners and educators interpret and use AI-generated feedback. Analyses of programming lab logs and reflective surveys explored student uptake of LLM-generated hints [115]. Surveys and thematic analyses captured preferences and perceived usefulness [119]. Experimental designs also examined how AI-based assessment insights influenced educators' feedback practices [149].

The reviewed studies demonstrate substantial methodological diversity, combining controlled intervention research with comparative system analyses and socio-cognitive investigations of feedback perception and use.

5.2.2. Learning domains and student assignments

In which learning domains has research on LLM-generated feedback been conducted, and for which tasks or student assignments has such feedback been provided (RQ 2.2)? We assigned the reviewed studies to five learning domain categories.

a) English (EFL, ESL) writing and academic writing: This is the largest domain and focuses primarily on argumentative, reflective, descriptive, and literature review writing assignments in English as a Foreign Language (e.g., [21,96,102,103,110,112,120,130,131,141, 144,148]). Tasks typically involve essay drafting and revision, with feedback targeting coherence, grammar, lexical resource, argumentation, and task achievement. Several studies also draw on standardized formats such as IELTS or TOEFL prompts.
b) Teacher education and educational sciences: These studies are situated in teacher education and focus on diagnostic reasoning, reflective writing, argumentation, learning goal formulation, tutoring competencies, and pedagogical case analysis. Writing tasks often require justification of pedagogical decisions, integration of multiple data sources, or reflection on classroom scenarios (e.g., [23,25,89, 97,109,111,146]).
c) Computer science and programming education: This domain includes introductory and intermediate programming courses (CS1, Python, Java, C#) and focuses on code comprehension, recursion, conditional logic, debugging, and software development tasks. Feedback targets code correctness, explanation quality, reasoning, and programming practices (e.g., [92,114,115,117,119,129]).
d) Mathematics and science education: Studies in mathematics examine open-ended problem solving, conceptual understanding (e.g., decimal magnitude, linear equations, ratios), and written mathematical reasoning. Tasks frequently involve short-answer explanations rather than purely numerical responses (e.g., [22,88,98]). This category also includes science disciplines where students respond to conceptual or inquiry-based tasks. Assignments involve open-ended explanations, experimental reasoning, laboratory protocols, or conceptual understanding of foundational principles (e,g., [24,104,114, 149]).
e) Other technical domains: Some studies focus on domain-specific technical writing and analytical reasoning, such as project reports in software engineering, data science proposals, or marketing analytics assignments involving statistical modeling and coding (e.g., [85,106]).

Across the reviewed studies, LLM-generated feedback has been investigated most extensively in English academic writing contexts, followed by teacher education and computer science/programming. Other domains—including mathematics, science, data science, and applied design—are represented but comparatively less frequently.

5.2.3. Educational level

At which educational levels and educational institutions have studies on LLM-based feedback been conducted (RQ 2.3)? Our analyses clearly showed that the majority of the reviewed studies (33 out of 47) were conducted in tertiary education, including undergraduate and graduate university contexts (e.g., [20,23,89,96,119,144]). About one-third-of the studies (14 out of 47) focus on secondary school settings, mostly high schools, in writing, science, and mathematics [21,22,24,87,88,98, 130]. A few projects involve teacher-training and educational psychology programs situated at the university level [89,97,111,146]. We found no projects that involved primary school students.

5.2.4. Sample types

What types of samples are investigated, distinguishing between human participants (students, teachers) and artefactual data, e.g., essays, responses, or LLM outputs (RQ 2.4)?

In synthesizing the reviewed studies, it is important to distinguish between different types of samples that underpin the empirical evidence. Research on LLM-generated feedback draws on diverse sample configurations, including human participants (such as students, teachers, tutors, and expert raters) as well as artefactual samples (such as

essays, code submissions, short-answer responses, and feedback texts). These sample types differ not only in scale but also in their analytical role: human samples allow investigation of learning processes, perceptions, and decision-making, whereas artefact and feedback samples enable evaluation of feedback quality, model performance, and alignment with expert judgment. Differentiating between these configurations is therefore essential for understanding the methodological foundations and interpretive scope of the field.

a) Student-only intervention samples form the largest group of studies. In these designs, students are the primary participants and unit of analysis. Data typically include essays, programming assignments, revision drafts, exams, surveys, or system logs, and the focus lies on learning outcomes, revision quality, or feedback perception (e.g., [21,101,132,146]).
b) A second group combines student samples with human raters or expert evaluators. In these studies, student artefacts such as essays, experimentation protocols, or programming solutions are assessed by teachers, trained coders, or domain experts using rubrics or annotation schemes (e.g., [24,103]).
c) Another set of studies employs mixed student–teacher designs in which both groups are active participants. Teachers may provide comparison feedback, participate in interviews, code reflections, or act as evaluators alongside AI systems (e.g., [20,142]).
d) A smaller but distinct group centers primarily on teacher or educator samples. In these studies, educators are the main participants, and the focus lies on grading practices, perception of AI-supported assessment, or instructional decision-making (e.g., [117]).
e) Finally, several studies rely predominantly on large-scale artefact or log-based datasets. The primary data consist of thousands of student responses, code submissions, or tutoring system logs, sometimes complemented by targeted human annotation (e.g., [22,86,88]).

Additionally, it is important to consider how samples were recruited and structured. A large proportion of studies rely on convenience samples (e.g., [23,101,119,130,146]. In contrast, several studies implemented randomized controlled trials or structured experimental group assignments to strengthen internal validity and causal inference (e.g., [21,96]),

5.2.5. Educational context and instructional setting of the studies

In which instructional settings has research on LLM-based formative feedback been carried out (RQ 2.5)? The reviewed studies were conducted across a diverse range of educational contexts and instructional settings, which can be grouped into four categories:

a) Studies conducted in authentic classroom settings integrated LLM-generated feedback directly into ongoing instruction, allowing researchers to observe its functioning within typical learning environments. For example, students revised essays during scheduled class periods or submitted programming tasks as part of their coursework, receiving LLM-based or comparative feedback while following standard instructional routines (e.g., [85,89,141,144,146]).
b) Studies in controlled laboratory or research-lab environments isolated specific aspects of LLM-generated feedback under standardized conditions. Participants typically completed writing or reasoning tasks within fixed time slots, enabling precise evaluation of feedback quality or revision effects without classroom variability (e.g., [21]).
c) Online and hybrid course contexts leveraged digital platforms that naturally support scalable, asynchronous feedback delivery. Students engaged with LLM-generated comments through learning management systems or online assessment tools while completing regular course activities. Examples include programming labs, writing modules, or simulation-based training environments where LLM feedback supplemented automated testing or human review (e.g., [113,115,149]).
d) Post-hoc or non-interactive studies analyzed existing student work without delivering feedback to learners directly. These investigations mainly focused on evaluating feedback quality (e.g., [24,87,120]).

5.3. Technological trends and empirical results (RQ 3)

5.3.1. Types of LLMs used in the studies and LLM specifications

What types of LLMs are used in these studies, and how are they configured (RQ 3.1)?

a) Proprietary, pre-trained models: Most studies relied on pre-trained GPT-3.5 or GPT-4 variants, accessed via API, web interface, or platform integration, typically without fine-tuning (e.g., [96,98,130,146]). Other studies employed more advanced models, such as GPT-4o and GPT-4.5, to generate adaptive or domain-specific feedback with higher linguistic and analytic quality (e.g., [89,109]). OpenAI with its GPT models was the most frequent LLM-provider in the reviewed studies. Only a minority of studies utilized other commercial models such as Gemini from Google [145], ERNIE Bot from Baidu [141], or Claude from Anthropic [109].
b) Open-source models: A second group of studies used open-source foundation models, often for reasons of transparency or local deployment. Llama-based systems were tested for student perceptions (e.g., [23,74]). Other studies used Mistral, Qwen, or DeepSeek for evaluating code, generating structured feedback, or supporting science and chemistry tasks (e.g., [22,114]). Additional architectures such as Bert, SBERT or a fine-tuned GOAT model appeared in comparative benchmarking (e.g., [88,149]).
c) Domain-specific fine-tuned models: A smaller number of studies employed fine-tuned models, often trained on domain-specific feedback datasets. Jia et al. [20,86] used BART-large variants fine--tuned on hundreds of teacher-student feedback pairs, enabling close alignment with instructor tone and disciplinary norms. One study used a fine-tuned multimodal vision–language model adapted with Low-rank Adaptation (LoRA) to evaluate experimental protocols integrating text and images [112].
d) Agentic systems: Few studies implemented multi-LLM pipelines or agent-based systems, where distinct models handled error analysis, feedback generation, and validation. This included coordinated GPT-4o, GPT-4o-mini, or expert-role LLM agents (e.g., [117]) as well as open-source multi-agent assemblies integrating Llama, Qwen, and related models for local deployment (e.g., [74]).

5.3.2. Prompting strategies explored in the reviewed studies

What kinds of prompting strategies are employed or investigated (RQ 3.2)?

a) Instructional, role-based, zero-short prompts: Most studies relied on structured instructional prompts, which defined the LLM's role (e.g., teacher or tutor), specified feedback criteria, and required formatting or tone. Such structured and role-based prompting was used widely in writing, programming, and subject-matter feedback studies (e.g., [21,25,96,101,103,113,114,147]).
b) Context-enriched, few-shot prompting: A second major strategy involved few-shot prompting, in which researchers embedded exemplar student–feedback pairs to enforce consistent tone, structure, and feedback style (e.g., [23,86,89,104,111]). Studies also incorporated contextualized prompts enriched with rubrics, correct solutions, model answers, common misconceptions, test-case outputs, or student performance data (e.g., [74,98,104,119]).
c) Multi-step prompting, Chain-of-Thought prompting: This group of studies implemented multi-step or process-oriented prompting, guiding the LLM through staged actions such as reformulating the student response, explaining errors, providing corrective guidance,

and summarizing key points (e.g., [24,88,92]). Farrokhnia et al. [132], for example, compared teacher feedback on student written essays to feedback generated with zero-shot prompting and Chain-of-Thought (CoT) prompting. The CoT prompt included a step-by-step reasoning instruction and an explicit worked example, guiding the model to evaluate essay components sequentially.

5.3.3. Results of the reviewed studies

What are the main empirical results of LLM-based feedback research along the dominant research questions (RQ 3.3)? For this core questions of our review, we grouped research results in six categories aligned with the conceptual framework for elaborated formative feedback research (see Fig. 1). Since most studies include several research questions, we assigned some example studies to several categories.

a) Effects of LLM-generated feedback on cognitive learning outcomes: Across studies comparing LLM feedback to no feedback, most report clear learning benefits, particularly in writing, programming, and diagnostic reasoning tasks. LLM-based feedback improved revision quality, essay scores, justification quality, code performance, and post-test outcomes in secondary and higher education contexts [21,89,92,96,110-112]. Effect sizes are typically small to moderate but consistent. When compared directly to teacher feedback, findings are more mixed: some studies report equal or superior outcomes for human feedback [25,101,114,118,144], while others find no significant performance differences between AI and teacher feedback [104,114,146]. In secondary EFL contexts, teacher feedback led to stronger and more durable grammar gains than ChatGPT feedback [130]. Hybrid approaches appear particularly promising: combining ChatGPT with teacher feedback enhanced revision performance beyond teacher-only conditions [142], and integrating AI with an existing AWE system significantly improved higher-order thinking compared to AWE alone [147].

b) Effects on metacognitive learning outcomes: Research indicates that LLM-generated formative feedback can support metacognitive development. Several intervention studies report improvements in metacognitive regulation, planning, monitoring, and higher-order thinking when AI feedback is integrated into writing or programming instruction [92,147]. More specifically, Liu et al. [141] found significant gains in overall feedback literacy and across all five feedback literacy dimensions in the LLM-feedback group compared to peer feedback. However, other studies report mixed effects. Liebenow et al. [131] reported no overall main effect of LLM feedback on self-assessment accuracy, although students with initially low self-assessment accuracy improved significantly more when receiving feedback. Moreover, teacher feedback may better support personalized scaffolding and affective trust, which can indirectly influence metacognitive engagement [130].

c) Effects on emotional and motivational learning outcomes: Across the reviewed studies, evidence for sustained motivational effects of LLM-generated feedback is generally positive but modest and context-dependent. Randomized and quasi-experimental studies report increases in task motivation and interest over the course of interventions when students received LLM-based feedback compared to no feedback [21,96,110]. In programming contexts, formative LLM-generated feedback was associated with higher reported interest and motivational across the learning period [89,92]. Some studies also indicate gains in confidence or writing self-efficacy during multi-week writing interventions [102]. However, motivational effects were not universally stronger than those of teacher feedback: in some contexts, AI feedback did not significantly outperform human feedback in motivational dimensions [101,118].

d) Internal processing of LLM-generated formative feedback: Across the reviewed studies, LLM-generated feedback is generally perceived as clear, structured, and easy to understand, with many students describing it as specific, fast, and helpful, and often rating it as at least as useful as human feedback in formative contexts [101,102,106,110-112,119]. It frequently supports active revision behavior, increasing revision frequency and prompting both local and global changes, particularly when feedback is detailed and tailored [103,115,145]. However, uptake often concentrates on surface-level corrections rather than deeper conceptual revisions [103,151], and learning benefits depend on learners' active engagement rather than mere access [111]. Perceptions and preferences are nuanced: some learners favor AI feedback for its immediacy and systematic structure, while others prefer teacher feedback for empathy, contextual understanding, and relational trust [25,101,103]. Disclosure of the feedback source plays a critical role—when students learn that feedback is AI-generated, trust, credibility, and feelings of appreciation can decline, even if the identical feedback was previously rated positively [23,116]. In secondary EFL contexts, however, teacher feedback elicited higher trust due to personalized scaffolding and affective support, whereas ChatGPT feedback was perceived as more mechanical and less adaptive, reducing perceived effectiveness [130].

e) Feedback content: Studies that evaluated the content of LLM-generated feedback through expert or trained rater judgments conclude that LLM feedback is pedagogically usable and often close to human quality, while still showing systematic limitations. In science education, LLM feedback on experimentation protocols was rated as largely comparable to teacher and expert feedback in structure and pedagogical soundness, though weaker in context-specific error explanation and depth [24]. In history writing, human evaluators rated teacher feedback higher in clarity, accuracy, prioritization, and tone [87]. In mathematics, expert comparisons showed that fine-tuned models aligned better with teacher grading patterns than similarity-based systems, and zero-shot GPT-4 often misaligned with implicit grading standards [88]. In data science writing, ChatGPT feedback was rated as more readable and structurally coherent than instructor feedback, but showed inflated positivity and weaker alignment on evaluative dimensions [85]. Prompting strategies strongly influenced expert-rated quality: domain-specific prompts and conversational or chain-of-thought prompting significantly improved structural quality, specificity, and developmental guidance compared to zero-shot outputs [109,120,132,143]. However, hallucination analyses revealed that a substantial proportion of feedback sentences were unfaithful to the source text, indicating risks that may not always be captured by surface-level quality ratings [86].

f) Role of learner characteristics: Research indicates that learner characteristics meaningfully moderate both the processing of and the effects produced by LLM-generated feedback. Learners with initially low self-assessment accuracy benefited more from LLM feedback in terms of improved calibration, whereas students with already accurate self-assessments showed little additional gain [131]. Similarly, feedback quality and its developmental potential appear to scale with the quality of students' initial work in some contexts, meaning that stronger writers may receive more elaborated or structurally refined AI feedback, which can widen performance differences if not carefully scaffolded [87,132]. Motivational orientation and feedback literacy shapes feedback uptake. Intrinsically motivated students are more likely to engage deeply with LLM feedback and translate it into revision behavior [25], while active voluntary use of AI feedback predicts learning gains more strongly than mere access [111]. Students with higher feedback literacy perceive feedback differently and may benefit more from teacher feedback, whereas those with lower literacy sometimes rely more heavily on structured AI guidance [25,141]. Language proficiency may also play a role: while some studies find no systematic disadvantage for English learners in AI feedback quality [87,110], linguistic context can influence how feedback is interpreted and how well models perform in underrepresented languages [132].

g) Role of learning context: Research suggests that the formative assessment context and the broader instructional design substantially shape the quality, processing, and effects of LLM-generated feedback. First, integration within structured formative designs enhances both feedback quality and learning impact. When LLM feedback is embedded in scaffolded revision cycles, adaptive systems, or hybrid teacher–AI models, effects on performance and higher-order thinking are stronger than when AI operates in isolation [129,142,147]. Second, the comparison condition within the formative context matters. LLM feedback shows clearer benefits when compared to no feedback [21,110], whereas differences relative to teacher feedback are often small, context-dependent, or favor teachers on nuanced dimensions [87,114]. Third, the nature of the task and disciplinary domain moderates effects. LLM-generated formative feedback performs well for structured writing, programming exercises, and rubric-driven tasks, especially in early drafts or lower-quality work [24,117]. However, limitations emerge in complex reasoning, highly contextualized critique, or unfamiliar error types [22,86].

5.3.4. Risks and challenges of LLM-generated formative feedback

How do studies address issues of hallucinations, biases, and fairness in pedagogical and ethical challenges of LLM-generated feedback (RQ 3.4)? Across the reviewed corpus, studies converge on three main categories of concerns, while also pointing to emerging mitigation strategies and remaining gaps.

First, multiple studies document inaccurate, ungrounded, or misclassified feedback, including hallucinated references to missing content, erroneous validation of student work, and rubric-inconsistent judgments. For example, LLM feedback sometimes referenced absent elements in software engineering project reports [20], occasionally reinforced misconceptions in programming when superficially plausible outputs aligned with test cases [115], and misclassified incorrect mathematical reasoning as correct due to keyword matching rather than conceptual evaluation [98]. Similar issues were reported in extended classroom use, where fabricated or contradictory feedback occurred and limitations in discipline-specific accuracy were acknowledged [141].

Second, studies identify positivity bias as a recurring pattern: LLM-generated feedback can be overly polite, inflated, or insufficiently critical, which risks masking errors and weakening instructional precision. This was evident when ChatGPT produced systematically more positive evaluations than instructors on dimensions such as novelty or clarity in project proposals [85], used generic encouragement even for incorrect responses in decimal reasoning tasks [98], and delivered supportive but sometimes non-specific praise in grammar-oriented writing feedback [102]. Relatedly, some learners reported difficulty interpreting AI feedback and raised concerns about overreliance, indicating an ethical risk when users treat AI guidance as authoritative without sufficient critical evaluation [147].

Third, fairness and bias analyses are comparatively sparse but suggest mixed implications. Some studies found no systematic disadvantage for English learners relative to fluent writers in feedback quality, implying no detectable linguistic bias in the evaluated context [87], and no differences in AI feedback responsiveness across proficiency levels in a large writing intervention [110]. Other comparative work suggests that LLM feedback may apply more uniform standards than humans, potentially reducing disparities arising from differential feedback practices [120]. However, most studies do not systematically test demographic bias beyond limited subgroup comparisons, and language coverage remains a concern: work conducted in Persian highlights that model behavior may vary in linguistically underrepresented languages, with feedback quality scaling more strongly with input quality [132].

6. Discussion

This review corroborates several central findings of the five previous literature reviews on AI-assisted feedback [26,27,29,30]. Consistent with these earlier syntheses, the evidence indicates that AI-generated feedback generally produces positive effects on revision quality and short-term learning outcomes when compared to no-feedback conditions, and in many cases approaches the effectiveness of human feedback [28–30]. As previously reported, research is heavily concentrated in higher education, particularly in English academic writing and EFL/ESL contexts [26–28], with comparatively fewer studies in school settings or non-language domains. Across reviews, learners tend to perceive AI-generated feedback as clear and useful, while still valuing teachers for contextual sensitivity and affective support [28,30]. Furthermore, recurring concerns—such as hallucinations, bias, overreliance, and insufficient pedagogical integration—are echoed in the present synthesis [26,29], reinforcing the broader conclusion that AI feedback should be embedded within carefully designed instructional frameworks and, ideally, combined with teacher oversight in hybrid models.

At the same time, this review expands the existing knowledge base in several important ways. First, unlike broader reviews that aggregate diverse AI technologies—including rule-based systems, traditional NLP tools, and LLMs [26,27,30]—this review focuses exclusively on instruction-tuned, general-purpose LLMs published between 2022 and 2026, thereby isolating the post-ChatGPT technological shift. Second, it provides a more fine-grained analysis of model configurations and prompting strategies, which are only briefly addressed in earlier reviews [28,29]. Third, by embedding findings in an integrative conceptual framework of formative feedback, the review systematically differentiates feedback content, internal feedback processing, and cognitive, metacognitive, and motivational outcomes—an analytical distinction not developed in the prior syntheses (Fig. 1). From this perspective, the reviewed literature can be interpreted as an expansion of the established discussion on elaborated digital feedback in the context of formative assessment in educational contexts. We therefore discuss the findings of the literature review along this framework.

6.1. Discussion of the review results

6.1.1. External feedback agents

The motivation for this review is the emergence of LLMs as a new class of external feedback agents. Accordingly, many studies used experimental or quasi-experimental designs to compare LLM-generated feedback with teacher or peer feedback (e.g., [21,25,101,114,148]). Overall, LLM feedback is often broadly comparable in quality to human feedback, although human feedback remains stronger for nuanced contextualization and prioritization in some settings [25,87,114]. Given rapid model development and the scalability of instant feedback, generative AI may increasingly match or exceed human capacity for elaborated feedback on open-ended formative assignments.

In contrast, fewer studies systematically compare different LLMs or prompting strategies. Most rely on GPT models, while only a smaller subset examines open-source models, fine-tuning, or prompt variations [22,86,88,109]. Yet available evidence shows that model type and prompt engineering substantially shape feedback quality: prompting can improve structural quality (e.g., CoT vs. zero-shot), but higher-quality messages do not necessarily translate into stronger revision outcomes, highlighting the role of contextual sensitivity, pedagogical adaptivity, and learner uptake [103,132,151]. Where stronger models are paired with structured prompting, feedback tends to be more coherent and criterion-aligned, but evidence remains limited [109,120]. Looking ahead, more explicit reasoning models, multimodal and tool-augmented systems, and retrieval- and rubric-grounded pipelines may improve accuracy and reduce unfaithful statements [86,74,112], shifting the research frontier from LLM-human comparisons toward uptake and regulation in authentic classroom conditions.

6.1.2. Feedback content

Across the reviewed studies, research on LLM-generated formative feedback places particularly strong emphasis on feedback content. First, factual correctness is the most prominent quality criterion, reflecting concerns that fluent feedback may nevertheless be inaccurate or ungrounded. Although LLM feedback is mostly described as correct in the reviewed studies, some studies also document that LLM feedback can include hallucinated statements or misclassifications, such as referencing missing elements in student artefacts that are actually present [20], validating incorrect reasoning in code comprehension tasks [115], or overestimating correctness in mathematical explanations through keyword heuristics rather than conceptual evaluation [98]. Importantly, the evidence suggests that these problems vary by learning domain and model configuration and are rarely addressed systematically across full classroom-scale deployments. At the same time, a small set of studies points toward mitigation through verification and evaluation pipelines [86,98,115]. This direction aligns with the broader argument that hallucinations are not fixed model properties but can be reduced through deliberate model selection, prompt design, and human-in-the-loop procedures [121,137].

Second, many studies operationalize pedagogical quality using established formative feedback dimensions such as Feed Up/Feed Back/Feed Forward [21,24]. Other studies applied formative feedback models primarily as analytic lenses to characterize the focus of LLM feedback, e.g., by coding feedback for task, process, and self-regulation levels and showing that LLM feedback tends to concentrate on task and process while under-representing self-regulation prompts [85]. A notable research gap is therefore that relatively few studies examine whether LLM feedback consistently targets higher-level process or strategy guidance, despite evidence on elaborated feedback that process- and self-regulation-focused feedback is often more beneficial than purely task-level correctness feedback [2,4,7].

Third, beyond accuracy, multiple studies highlight systematic distortions in feedback content. Positivity bias is repeatedly observed, with LLMs tending to produce overly polite or inflated evaluations that can reduce diagnostic value [85,98,102]. Moreover, an algorithmic "input-quality sensitivity" emerges as a potential fairness-relevant mechanism: in some studies, feedback quality correlates with the quality of students' initial drafts, raising the risk of reinforcing achievement gaps if weaker submissions trigger less useful or less actionable feedback [132]. This concern aligns with earlier evidence from digital feedback contexts that low-achieving learners may receive more error information yet engage with it less effectively, potentially due to cognitive load and self-protective processing [12,152]. The implication is that future work should examine prompt and system mechanisms that compensate for low-quality inputs (e.g., stronger diagnosis, prioritization, staged guidance) rather than merely mirroring student performance.

Fourth, the review indicates that linguistic and cultural generalizability remains an underexplored dimension of feedback content. Only a small number of studies explicitly address language coverage, but work in linguistically underrepresented contexts suggests that model behavior may differ across languages and writing conventions, which may partly explain observed links between input quality and feedback quality [132]. This highlights a broader limitation of the evidence base: despite the global relevance of formative feedback, most evaluations are conducted in English or in English-learning contexts, leaving open questions about feedback validity, fairness, and usefulness across non-English educational settings.

Finally, adaptivity remains a central "missing link" in feedback content research. While LLM feedback is often elaborated and can be structured using formative principles, most systems remain artefact-centric: they generate feedback solely from the submitted product without systematically incorporating learner goals, prior knowledge, misconception profiles, or motivational and affective states. This stands in contrast to decades of evidence from adaptive learning and tutoring systems showing that personalization based on learner models can enhance both effectiveness and equity [12,53,55].

6.1.3. Internal feedback processing

Our review demonstrates that internal feedback processing is a key area of LLM-generated formative feedback research. This focus aligns with established feedback process models that conceptualize feedback as a learner-centered cycle of interpretation, affect regulation, judgment, and action rather than a one-way transmission of information [2,37,43,51]. Our review shows that LLM-feedback research clearly extends this tradition. A large share of studies does not only evaluate outcomes, but explicitly examines how learners perceive, understand, and use LLM-generated feedback, often treating uptake as a central dependent variable [23,96,103,119,151].

A first recurring theme is that LLM-generated feedback is typically elaborated and therefore places high demands on learners' selection and sense-making processes. Several studies describe LLM-generated formative feedback as clear, structured, and actionable, and learners often report that it helps them orient their revision efforts [96,102,110,112]. At the same time, uptake evidence suggests that elaboration can interact with learners' limited processing capacities: students frequently prioritize surface-level edits (grammar, wording) over deeper conceptual revision even when higher-order suggestions are available [103,151]. This pattern mirrors earlier findings from digital feedback research that learners may not fully process dense feedback, may selectively attend to low-effort fixes, or may avoid engaging with challenging information for self-protective reasons [51,152]. This suggests that LLM feedback systems should be designed not only to generate helpful messages but also to structure learners' engagement with those messages in ways that sustain autonomy and reflective processing [37,43].

Internal processing is not only a mediator but also a gatekeeper: learning effects depend on learners' active engagement with feedback. One study explicitly operationalized feedback seeking and showed that the availability of LLM-generated formative feedback alone was insufficient. Learning gains emerged only for participants who actively chose to use the LLM feedback [111]. This resonates with broader evidence that feedback seeking and willingness to confront corrective information are central determinants of feedback effectiveness [8,43] and with research showing that learners' choice to seek feedback can explain more variance in performance than revision behavior alone [46]. These findings indicate that future LLM-feedback research should treat feedback seeking not as context variables but as central constructs, and should test interventions that explicitly foster productive engagement (e.g., reflective prompts, self-explanation requirements, or dialogic feedback interfaces).

6.1.4. Feedback effects on learning outcomes

Across the reviewed studies, the overall pattern is consistent with decades of evidence on elaborated formative feedback: providing learners with informative, explanatory feedback is generally more effective than providing no feedback or minimal corrective information [8,13]. In line with this, multiple experimental and quasi-experimental studies report that LLM-generated feedback improves cognitive learning outcomes, most prominently the quality of revisions and near-term task performance in writing, reasoning, and programming contexts [21,89,92,96,110,112]. When LLM feedback is compared to no-feedback or weaker digital feedback conditions, effects are typically positive and robust, reinforcing the conclusion that LLM-generated formative feedback can function as an effective formative support under realistic instructional constraints [21,110].

Comparisons with human feedback are more nuanced: depending on domain and task, LLM feedback performs similarly, sometimes better, and sometimes worse than teacher feedback, with teachers often retaining an advantage for complex tasks requiring contextual prioritization or nuanced guidance [25,101,114,130], while other studies find no significant performance differences [114,104,146]. Evidence also

suggests that hybrid approaches can be particularly productive, for example when AI feedback complements teacher feedback or when AI is integrated with other feedback systems, yielding improvements beyond single-source feedback in some designs [129,142,147].

At the same time, the reviewed literature indicates that most learning effects are proximal and revision-focused. A large share of studies uses short-term, single-task designs (one assignment and one revision cycle), which makes it difficult to draw conclusions about durable learning, transfer to new tasks, or cumulative "feedback spirals" across time [21, 132]. This pattern echoes broader feedback research: elaboration can support immediate correction, but durable learning requires repeated opportunities to apply feedback, build evaluative judgment, and develop self-regulation and feedback literacy [8,37]. Future work should therefore move beyond "single-shot" revision effects and examine how LLM-generated feedback operates across multiple cycles, assignments, and contexts.

6.1.5. *Formative assessment context*

The integrative framework also highlights that LLM-generated feedback is embedded in specific formative assessment tasks and classroom routines, and its effectiveness depends on this proximal assessment context. The reviewed studies demonstrate that LLM-generated formative feedback is already being applied across a range of open-ended assignments, including argumentative and reflective essays, project reports, programming tasks, and short-answer explanations [20,21,89,96, 119]. However, the empirical landscape remains strongly concentrated in a limited set of assessment formats, most notably EFL/ESL academic writing in higher education and introductory programming exercises, with fewer studies addressing other domains such as mathematics, science, or history writing [22,24,87,88,98]. Secondary education contexts are represented but less common, often focusing on writing, grammar, or selected STEM topics [21,104,130].

This concentration limits generalizability because formative assessment varies substantially across subjects in task structure, epistemic aims, and what counts as a "good" response. An urgent research need is therefore to expand beyond domains where LLMs naturally excel in fluent text production and into more diverse assessment types, particularly those requiring precise conceptual diagnosis and domain-specific reasoning, such as complex mathematical problem solving or lower-secondary topics like fractions and foundational computation where LLMs have documented weaknesses. Existing studies in mathematics already suggest that performance can be uneven, especially when tasks require robust conceptual evaluation rather than surface plausibility [22,98], underscoring the importance of domain-sensitive research that tests how LLM feedback functions across different formative assessment traditions and knowledge representations.

6.1.6. *Learner and context characteristics*

Several reviewed studies provide direct evidence that initial learner status moderates outcomes. For example, self-assessment accuracy functions as a clear moderator: learners with low initial calibration accuracy benefited more from LLM feedback than those who were already accurate, suggesting that AI feedback can support metacognitive correction particularly for weaker self-monitors [131]. Relatedly, feedback literacy and motivational orientation influence how feedback is perceived and acted upon: students with higher feedback literacy tend to perceive teacher feedback as fairer and benefit more from it, whereas intrinsically motivated students show higher willingness to revise after LLM feedback, implying that learners' dispositions shape both trust and uptake [25].

A particularly critical learner-related risk emerging from the review is input-quality sensitivity, or a Matthew-effect mechanism. In some studies, the quality of AI-generated feedback is significantly correlated with the quality of students' initial submissions, meaning that stronger drafts elicit more coherent, relevant, and structurally refined feedback while weaker drafts may trigger more generic, less actionable, or less pedagogically adaptive responses [132]. In digital feedback research more broadly, low-achieving learners tend to receive more error information but can process less of it, contributing to selective uptake and reduced benefits [152], and similar dynamics may occur in LLM-feedback contexts when elaboration becomes cognitively or emotionally taxing.

The reviewed studies also emphasize the importance of the surrounding learning context, including classroom norms, and the characteristics of the digital learning systems LLM-feedback is embedded in In line with broader AIEd work, several studies implicitly or explicitly support a human-in-the-loop perspective: educators remain responsible for aligning feedback with curricular goals, ensuring appropriateness for the learner group, and mitigating inaccuracies or unintended pedagogical consequences [114]. Similarly, work that uses LLM-based "insights" to support teachers' feedback writing shows that LLM support can improve the quality of educators' written feedback compared to simpler AI aids, indicating a productive division of labor where AI augments teacher judgment rather than replacing it [149].

However, most studies do not systematically examine demographic bias, data privacy, or governance arrangements for classroom deployment. Instead, the few fairness-related analyses focus on linguistic proficiency and suggest that LLM feedback can be relatively uniform across learner language backgrounds, whereas human teachers may differentiate more strongly [87,110,120]. While this uniformity could imply fairness advantages, it co-occurs with recurring positivity bias and overpraise, which can inflate evaluations and weaken diagnostic precision [85,98,102]. Moreover, hallucination and misclassification risks raise immediate instructional governance questions: who is responsible for validating feedback, how errors are detected, and how accountability is handled when learners act on incorrect guidance [20,86]. In sum, LLM-generated feedback should therefore be conceptualized as a socio-technical component of formative assessment systems, where effectiveness and ethical acceptability depend on teacher oversight, platform design, transparency practices, and institutional safeguards, not merely on model capability.

6.2. *Implications for future research and development*

Building on the strengths and limitations identified in the reviewed corpus, future research should move beyond demonstrating that LLM-generated feedback "works" in narrow settings and instead clarify when, for whom, and under what conditions it is instructionally effective, trustworthy, and equitable. The following directions briefly synthesize methodological, pedagogical, and ethical priorities emerging from the literature.

Future research should …

- broaden sampling beyond convenience cohorts to strengthen external validity. Future studies should replicate findings across multiple institutions and use more diverse recruitment strategies so that results generalize beyond readily accessible university seminars and courses.
- expand in under-researched learning domains and educational contexts. Because current evidence is heavily concentrated in higher education (EFL writing, academic writing in teacher education, programming), more work is needed in K-12 school settings where formative feedback demands and learners' self-regulation capacities differ substantially. Future studies should also test LLM-generated feedback in a wider range of subjects and formative tasks, e.g., complex and foundational mathematics contexts where LLM reasoning is to date less reliable.
- develop and test hybrid human–AI feedback models rather than replacement scenarios. Research should identify optimal divisions of labor between teachers, peers, and LLMs, including when teacher oversight is essential and when AI can safely automate parts of the process [101,111]. We also question whether the field requires

additional disclosure experiments. While these studies have yielded valuable insights, AI-in-education research should move beyond the binary comparison of human versus AI feedback. In realistic future educational settings, teachers will increasingly collaborate with AI assistants, forming a "human-AI hybrid intelligence" [122] in which the boundaries between "human" and "AI" contributions are not always clearly distinguishable, but pedagogically co-constructed.

- embed LLM feedback in iterative "feedback spirals" across multiple drafts and assignments. Longitudinal designs should test cumulative learning, feedback literacy development, and transfer effects beyond single revision cycles. Jamshed et al. [102] or Sun et al. [92] are examples of longer interventions that allow establishing long-term effects of LLM-generated formative feedback integration into authentic classroom contexts. Interestingly, Sun et al. [26,92] described their quasi-experimental 10-week LLM-feedback intervention within a Python programming course as "relatively short duration". Researchers in the field probably should clarify their understanding of the external validity standards of LLM-generated feedback research.
- should build on decades of elaborated formative feedback research and harness the potential of LLMs to produce elaborated feedback messages with different foci (e.g., [92]: task-level, process-level, self-regulation-level, self-level) or different normative references (e.g., self-referenced, criterion-referenced, norm-referenced, group--referenced) and establish how these different types work in new learning domains and with new formative assessment types. Combined with the idea to use student models (state of knowledge acquisition, motivational state) and learning analytics data from intelligent tutoring systems (behavioral data) for feedback adaptivity, this opens possibilities to adapt feedback to student's learning history, their misconceptions in a domain, or their motivational characteristics (e.g., [12,55,57]).
- investigate internal feedback processing across diverse learning domains and instructional settings. The review shows that LLM-generated feedback research already treats feedback uptake, processing, perceptions, and usage as central mediator variables between feedback messages and learning outcomes (see Fig. 1). However, future work should more explicitly build on established feedback process models (e.g., [37,43]), which conceptualize internal processing as the critical bottleneck of feedback effectiveness—particularly in contexts where formative tasks become more complex, student responses involve multi-step reasoning, and feedback messages are increasingly structured and lengthy. Research could, for example, operationalize active feedback seeking and feedback choice as independent variables, testing whether giving learners control over feedback amount, timing, and level of detail enhances engagement and learning compared to passive receipt [46,111]. In addition, future studies should employ fine-grained behavioral measures of engagement with feedback, such as clickstream logging, time-on-feedback, scrolling patterns, or eye-tracking, to identify which feedback segments are actually attended to and why. Such evidence could inform more adaptive formatting and scaffolding of elaborated feedback messages.
- develop and evaluate mechanisms for hallucination mitigation and faithfulness assurance. Research should test retrieval- and rubric-grounded prompting, justification requirements linked to student artefacts, and automated verification pipelines to reduce unfaithful feedback [74,86]. Moreover, research on fairness and bias should go beyond linguistic proficiency and include diverse populations. Future work should test demographic fairness (e.g., gender, socio-economic background) and contextual bias systematically, rather than inferring fairness from limited subgroup analyses, e.g., expand research in non-English and underrepresented language contexts. Lastly, research should address "Matthew effects" and input-quality sensitivity in feedback generation. Studies should test mechanisms that compensate for low-quality submissions so that weaker learners receive feedback that is equally actionable and supportive [132].
- harness advances in LLM development beyond fluent text generation, including enhanced reasoning, grounding, and tool integration capabilities. Chain-of-Thought (CoT) prompting has shown that LLMs can generate explicit intermediate reasoning steps, substantially improving multi-step problem solving [123]. Retrieval-augmented generation (RAG) architectures further strengthen reliability by grounding outputs in retrieved documents, thereby reducing hallucinations and increasing faithfulness [124]. In addition, tool-augmented models can decide when to call external tools such as calculators, search engines, or code interpreters to verify results and support more accurate reasoning [125]. These technical developments are highly relevant for educational applications, especially formative feedback on complex, open-ended assignments. Future research on LLM-generated formative feedback should therefore systematically explore these technological advancements. Some reviewed studies already move in this direction (e.g., [86,115,132]). However, these efforts remain partial and exploratory. A systematic integration of advanced reasoning, grounding, and tool use into LLM-based formative feedback systems represents a key frontier for increasing pedagogical precision, reliability, and alignment with assessment criteria.

6.3. Limitations of this review

This review is subject to several limitations that must be considered when interpreting its findings.

Most importantly, the scope was deliberately restricted to empirical studies in which LLM-generated feedback constituted the core intervention. As a result, adjacent research areas—such as game-based learning, augmented and virtual reality, educational chatbots, and other AI-supported learning environments—were excluded, even though LLM-based feedback mechanisms may also play an important role within these contexts. Consequently, relevant technological developments and design insights that embed LLM feedback as part of broader, multimodal, or immersive systems may not be fully represented. This focused approach strengthens the internal coherence and interpretability of the findings by isolating feedback-specific effects, but it also limits the generalizability of the results across the wider ecosystem of AI-enhanced educational technologies. Future reviews could adopt a broader lens to examine how LLM-generated feedback functions in combination with other instructional features and learning environments.

A second limitation concerns the rapidly evolving and inconsistent terminology used to describe LLM-based feedback. Although a systematic search strategy was applied across five major databases using a comprehensive Boolean string, it remains possible that relevant studies were not retrieved because they employed nonstandard, emergent, or domain-specific labels for LLM-driven feedback. In particular, some studies may have framed LLM feedback under broader terms such as *AI-assisted assessment, automated formative guidance, intelligent tutoring,* or *adaptive support* without explicitly referencing LLMs, GPT, or generative AI in titles, abstracts, or keywords.

Third, publication bias is likely, as early positive or promising findings may be more readily published during periods of technological novelty [126]. In addition, some studies lacked sufficient transparency in reporting model configurations, prompt engineering procedures, hyperparameters, or versioning details. These omissions hinder reproducibility, limit comparability across studies, and introduce the possibility of selective reporting of effective prompt variants or model outputs.

Lastly, a limitation of this review concerns its temporal scope in a rapidly evolving technological landscape. We defined our inclusion window from 2022 onward, marking the broader advancement and public availability of general-purpose, instruction-tuned large language

models (LLMs). This decision reflects a technological turning point at which LLMs became widely accessible and practically usable for generating context-sensitive feedback in authentic educational settings without extensive task-specific fine-tuning. However, restricting the review to the period 2022–2026 inevitably excludes earlier research on precursor LLM technologies. As a result, the review does not capture the full historical trajectory of LLM development in education and may underrepresent conceptual continuities with earlier AI-based feedback systems. Furthermore, because model capabilities have evolved rapidly within the selected time frame itself, the synthesized evidence spans multiple generations of instruction-tuned systems with differing performance characteristics. This temporal concentration therefore enhances conceptual coherence but may limit historical completeness and comparability across technological generations.

7. Conclusions

This qualitative systematic review examined 47 empirical studies with 121 research questions published since the emergence of GPT-3.5 to map the rapidly expanding field of LLM-generated feedback in education. The synthesis shows that current research is conceptually grounded in formative feedback theory, self-regulated learning, and feedback literacy, yet remains methodologically fragmented and technologically heterogeneous. Empirically, LLM-generated feedback generally improves revision quality and short-term learning outcomes compared with no feedback and, under well-structured prompting, can approach the effectiveness of teacher feedback. At the same time, challenges such as hallucinations, positivity bias, and reduced trust when feedback is disclosed as AI-generated constrain its reliability and pedagogical uptake. Most existing systems generate feedback solely from the learner's artefact, with little adaptation to prior knowledge, goals, or affective states, indicating substantial room for advancing personalization through richer learner models and more sophisticated prompting or fine-tuning approaches. Eventually, the evidence supports LLMs as scalable contributors to formative feedback ecosystems, particularly within hybrid teacher-in-the-loop designs. Future progress will depend on rigorous longitudinal research, systematic evaluation of model configurations and prompting strategies, and theoretically informed development of adaptive and equitable AI-supported feedback practices.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethical statement

This study is a systematic review and did not involve human participants, animals, or identifiable personal data; therefore, ethical approval and informed consent were not required.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used ChatGPT (OpenAI) in multiple writing and revision cycles to support language refinement, improve readability, and check for language and content consistency throughout the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Data availability

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

CRediT authorship contribution statement

Uwe Maier: Writing – review & editing, Writing – original draft, Supervision, Methodology, Data curation, Conceptualization. **Moritz Seibold:** Writing – review & editing, Methodology, Data curation. **Christian Klotz:** Writing – review & editing, Methodology, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.caeo.2026.100374.

References

[1] Bangert-Drowns RL, Kulik C-LC, Kulik JA, Morgan M. The instructional effect of feedback in test-like events. Rev Educ Res 1991;61(2):213–38. https://doi.org/10.3102/00346543061002213.
[2] Hattie J, Timperley H. The Power of Feedback. Rev Educ Res 2007;77(1):81–112. https://doi.org/10.3102/003465430298487.
[3] Metcalfe J. Learning from errors. Annu Rev Psychol 2017;68(1):465–89. https://doi.org/10.1146/annurev-psych-010416-044022.
[4] Shute VJ. Focus on formative feedback. Rev Educ Res 2008;78(1):153–89. https://doi.org/10.3102/0034654307313795.
[5] Azevedo R, Bernard RM. A meta-analysis of the effects of feedback in computer-based instruction. Journal of Educational Computing Research 1995;13(2):111–27. https://doi.org/10.2190/9LMD-3U28-3A0G-FTQT.
[6] Fong CJ, Patall EA, Vasquez AC, Stautberg S. A meta-analysis of negative feedback on intrinsic motivation. Educ Psychol Rev 2019;31(1):121–62. https://doi.org/10.1007/s10648-018-9446-6.
[7] Kluger AN, DeNisi A. The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychol Bull 1996;119(2):254–84. https://doi.org/10.1037/0033-2909.119.2.254.
[8] Wisniewski B, Zierer K, Hattie J. The power of feedback revisited: a meta-analysis of educational feedback research. Front Psychol 2020;10:3087. https://doi.org/10.3389/fpsyg.2019.03087.
[9] Grundmann F, Scheibe S, Epstude K. When ignoring negative feedback is functional: presenting a model of motivated feedback disengagement. Curr Dir Psychol Sci 2021;30(1):3–10. https://doi.org/10.1177/0963721420969386.
[10] Maier U, Klotz C. Students ignore their mistakes: elaborated error feedback processing in a digital learning system. Contemp Educ Psychol 2025:102395. https://doi.org/10.1016/j.cedpsych.2025.102395.
[11] Brummer L, De Boer H, Mouw JM, Strijbos J-W. A meta-analysis of the effects of context, content, and task factors of digitally delivered instructional feedback on learning performance. Learn Environ Res 2024;27(3):453–76. https://doi.org/10.1007/s10984-024-09501-4.
[12] Maier U, Klotz C. Personalized feedback in digital learning environments: classification framework and literature review. Computers and Education: Artificial Intelligence 2022;3:100080. https://doi.org/10.1016/j.caeai.2022.100080.
[13] Van Der Kleij FM, Feskens RCW, Eggen TJHM. Effects of feedback in a computer-based learning environment on students' Learning outcomes: a meta-analysis. Rev Educ Res 2015;85(4):475–511. https://doi.org/10.3102/0034654314564881.
[14] Kuklick L, Lindner MA. Computer-based knowledge of results feedback in different delivery modes: effects on performance, motivation, and achievement emotions. Contemp Educ Psychol 2021;67:102001. https://doi.org/10.1016/j.cedpsych.2021.102001.
[15] Tärning B. Review of feedback in digital applications – Does the feedback they provide support learning? Journal of Information Technology Education: Research 2018;17:247–83. https://doi.org/10.28945/4104.
[16] Fleckenstein J, Liebenow LW, Meyer J. Automated feedback and writing: a multi-level meta-analysis of effects on students' performance. Frontiers in Artificial Intelligence 2023;6:1162454. https://doi.org/10.3389/frai.2023.1162454.
[17] Tate TP, Steiss J, Bailey D, Graham S, Moon Y, Ritchie D, Tseng W, Warschauer M. Can AI provide useful holistic essay scoring? Computers and Education: Artificial Intelligence 2024;7:100255. https://doi.org/10.1016/j.caeai.2024.100255.
[18] Wang S, Wang F, Zhu Z, Wang J, Tran T, Du Z. Artificial intelligence in education: a systematic literature review. Expert Syst Appl 2024;252:124167. https://doi.org/10.1016/j.eswa.2024.124167.

[19] Atkinson J, Palma D. An LLM-based hybrid approach for enhanced automated essay scoring. Sci Rep 2025;15(1). https://doi.org/10.1038/s41598-025-87862-3.

[20] Jia Q, Cui J, Du H, Rashid P, Xi R, Li R, Gehringer E. LLM-generated feedback in real classes and beyond: perspectives from students and instructors. In: *Proceedings of the 17th International* Conference on Educational Data Mining (EDM 2024); 2024. p. 862–7. https://doi.org/10.5281/ZENODO.12729974.

[21] Meyer J, Jansen T, Schiller R, Liebenow LW, Steinbach M, Horbach A, Fleckenstein J. Using LLMs to bring evidence-based feedback into the classroom: aI-generated feedback increases secondary students' text revision, motivation, and positive emotions. Computers and Education: Artificial Intelligence 2024;6:100199. https://doi.org/10.1016/j.caeai.2023.100199.

[22] McNichols H, Lee J, Fancsali S, Ritter S, Lan A. *Can large language models replicate ITS feedback on open-ended math questions?* (Version 2). arXiv 2024. https://doi.org/10.48550/ARXIV.2405.06414.

[23] Rüdian S, Podelo J, Kužílek J, Pinkwart N. Feedback on Feedback: student's perceptions for Feedback from teachers and few-shot LLMs. In: Proceedings of the 15th International Learning Analytics and Knowledge Conference; 2025. p. 82–92. https://doi.org/10.1145/3706468.3706479.

[24] Seßler K, Bewersdorff A, Nerdel C, Kasneci E. *Towards adaptive feedback with AI: comparing the feedback quality of LLMs and teachers on experimentation protocols* (Version 1). arXiv 2025. https://doi.org/10.48550/ARXIV.2502.12842.

[25] Weidlich J, Gotsch F, Schudel K, Marusic-Würscher C, Mazzarella J, Bolten H, Bütler D, Luger S, Wohlfender B, Merki KM. Teacher, peer, or AI? Comparing effects of feedback sources in higher education. Computers and Education Open 2025;9:100300. https://doi.org/10.1016/j.caeo.2025.100300.

[26] Ba S, Yang L, Yan Z, Looi CK, Gašević D. Unraveling the mechanisms and effectiveness of AI-assisted feedback in education: a systematic literature review. Computers and Education Open 2025;9:100284. https://doi.org/10.1016/j.caeo.2025.100284.

[27] Shi H, Aryadoust V. A systematic review of AI-based automated written feedback research. ReCALL 2024;36(2):187–209. https://doi.org/10.1017/S0958344023000265.

[28] Urzúa CAC, Ranjan R, Saavedra EEM, Badilla-Quintana MG, Lepe-Martínez N, Philominraj A. Effects of AI-assisted feedback via generative chat on academic writing in higher education students: a systematic review of the literature. Education Sciences 2025;15(10):1396. https://doi.org/10.3390/educsci15101396.

[29] Yildiz Durak H, Onan A. A systematic review of AI-based feedback in educational settings. Journal of Computational Social Science 2025;8(4):96. https://doi.org/10.1007/s42001-025-00428-1.

[30] Kaliisa R, Misiejuk K, López-Pernas S, Saqr M. How does artificial intelligence compare to human feedback? A meta-analysis of performance, feedback perception, and learning dispositions. Educ Psychol (Lond) 2025:1–32. https://doi.org/10.1080/01443410.2025.2553639.

[31] Anderson RC, Kulhavy RW, Andre T. Feedback procedures in programmed instruction. J Educ Psychol 1971;62(2):148–56. https://doi.org/10.1037/h0030766.

[32] Leeder TM. Behaviorism, Skinner, and operant conditioning: considerations for sport coaching practice. Strategies 2022;35(3):27–32. https://doi.org/10.1080/08924562.2022.2052776.

[33] Fyfe ER, DeCaro MS, Rittle-Johnson B. When feedback is cognitively-demanding: the importance of working memory capacity. Instr Sci 2015;43(1):73–91. https://doi.org/10.1007/s11251-014-9323-8.

[34] Butler DL, Winne PH. Feedback and self-regulated learning: a theoretical synthesis. Rev Educ Res 1995;65(3):245–81. https://doi.org/10.3102/00346543065003245.

[35] Carless D, Boud D. The development of student feedback literacy: enabling uptake of feedback. Assessment & Evaluation in Higher Education 2018;43(8):1315–25. https://doi.org/10.1080/02602938.2018.1463354.

[36] Timms M, DeVelle S, Lay D. Towards a model of how learners process feedback: a deeper look at learning. Australian Journal of Education 2016;60(2):128–45. https://doi.org/10.1177/0004944116652912.

[37] Winstone NE, Nash RA. Toward a cohesive psychological science of effective feedback. Educ Psychol 2023;58(3):111–29. https://doi.org/10.1080/00461520.2023.2224444.

[38] Fyfe ER. Providing feedback on computer-based algebra homework in middle-school classrooms. Comput Human Behav 2016;63:568–74. https://doi.org/10.1016/j.chb.2016.05.082.

[39] Mertens U, Finn B, Lindner MA. Effects of computer-based feedback on lower- and higher-order learning outcomes: a network meta-analysis. J Educ Psychol 2022;114(8):1743–72. https://doi.org/10.1037/edu0000764.

[40] Narciss S. The impact of informative tutoring feedback and self-efficacy on motivation and achievement in concept learning. Exp Psychol 2004;51(3):214–28. https://doi.org/10.1027/1618-3169.51.3.214.

[41] Kuklick L, Greiff S, Lindner MA. Computer-based performance feedback: effects of error message complexity on cognitive, metacognitive, and motivational outcomes. Comput Educ 2023;200:104785. https://doi.org/10.1016/j.compedu.2023.104785.

[42] Lui AM, Andrade HL. The next Black Box of formative assessment: a model of the internal mechanisms of feedback processing. Frontiers in Education 2022;7. https://doi.org/10.3389/feduc.2022.751548.

[43] Panadero E, Lipnevich AA. A review of feedback models and typologies: towards an integrative model of feedback elements. Educational Research Review 2022;35:100416. https://doi.org/10.1016/j.edurev.2021.100416.

[44] Timmers C, Veldkamp B. Attention paid to feedback provided by a computer-based assessment for learning on information literacy. Comput Educ 2011;56(3):923–30. https://doi.org/10.1016/j.compedu.2010.11.007.

[45] Lim L-A, Dawson S, Gaševic D, Joksimovic S, Pardo A, Fudge A, Gentili S. Students' Perceptions of, and emotional responses to, personalised learning analytics-based feedback: an exploratory study of four courses. Assessment & Evaluation in Higher Education 2021;46(3):339–59.

[46] Cutumisu M, Blair KP, Chin DB, Schwartz DL. Assessing whether students seek constructive criticism: the design of an automated feedback system for a graphic design task. International Journal of Artificial Intelligence in Education 2017;27(3):419–47. https://doi.org/10.1007/s40593-016-0137-5.

[47] Finkelstein SR, Fishbach A. Tell me what I did wrong: experts seek and respond to negative feedback. Journal of Consumer Research 2012;39(1):22–38. https://doi.org/10.1086/661934.

[48] Ober TM, Cheng Y, Carter MF, Liu C. Leveraging performance and feedback-seeking indicators from a digital learning platform for early prediction of students' learning outcomes. Journal of Computer Assisted Learning 2024;40(1):219–40. https://doi.org/10.1111/jcal.12870.

[49] Cutumisu M, Lou NM. The moderating effect of mindset on the relationship between university students' critical feedback-seeking and learning. Comput Human Behav 2020;112:106445. https://doi.org/10.1016/j.chb.2020.106445.

[50] Narciss S, Prescher C, Khalifah L, Körndle H. Providing external feedback and prompting the generation of internal feedback fosters achievement, strategies and motivation in concept learning. Learn Instr 2022;82:101658. https://doi.org/10.1016/j.learninstruc.2022.101658.

[51] Narciss S. Feedback strategies for interactive learning tasks. In: Spector JM, Merrill D, van Merrienboer JJG, Driscoll M, editors. Handbook of research on educational communications and technology. LEA; 2008. p. 125–44.

[52] Narciss S, Sosnovsky S, Schnaubert L, Andrès E, Eichelmann A, Goguadze G, Melis E. Exploring feedback and student characteristics relevant for personalizing feedback strategies. Comput Educ 2014;71:56–76. https://doi.org/10.1016/j.compedu.2013.09.011.

[53] Aleven V, McLaughlin EA, Glenn RA, Koedinger KR. Instruction based on adaptive learning technologies. In: Mayer RE, Alexander PA, editors. Handbook of research on learning and instruction. 0 ed. Routledge; 2016. https://doi.org/10.4324/9781315736419.

[54] Cavalcanti AP, Barbosa A, Carvalho R, Freitas F, Tsai Y-S, Gašević D, Mello RF. Automatic feedback in online learning environments: a systematic literature review. Computers and Education: Artificial Intelligence 2021;2:100027. https://doi.org/10.1016/j.caeai.2021.100027.

[55] Van Schoors R, Elen J, Raes A, Depaepe F. An overview of 25 years of research on digital personalised learning in primary and secondary education: a systematic review of conceptual and methodological trends. British Journal of Educational Technology 2021;52(5):1798–822. https://doi.org/10.1111/bjet.13148.

[56] VanLehn K, Milner F, Banerjee C, Wetzel J. A step-based tutoring system to teach underachieving students how to construct algebraic models. International Journal of Artificial Intelligence in Education 2024;34(2):224–46. https://doi.org/10.1007/s40593-023-00328-3.

[57] Pardo A, Jovanovic J, Dawson S, Gašević D, Mirriahi N. Using learning analytics to scale the provision of personalised feedback. British journal of educational technology, 50. Education Source; 2019. p. 128–38.

[58] Sales AC, Pane JF. The role of mastery learning in an intelligent tutoring system: principal stratification on a latent variable. Ann Appl Stat 2019;13(1). https://doi.org/10.1214/18-AOAS1196.

[59] Wang Z, Gong S-Y, Xu S, Hu X-E. Elaborated feedback and learning: examining cognitive and motivational influences. Comput Educ 2019;136:130–40. https://doi.org/10.1016/j.compedu.2019.04.003.

[60] Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, Min Y, Zhang B, Zhang J, et al. A survey of large language models (Version 16). arXiv 2023. https://doi.org/10.48550/ARXIV.2303.18223.

[61] Matarazzo A, Torlone R. A survey on large language models with some insights on their capabilities and limitations (Version 2). arXiv 2025. https://doi.org/10.48550/ARXIV.2501.04040.

[62] Debnath T, Siddiky MNA, Rahman ME, Das P, Guha AK, Rahman MR, Kabir HMD. A comprehensive survey of prompt engineering techniques in large language models. Preprints 2025. https://doi.org/10.36227/techrxiv.174140719.96375390/v2.

[63] Hewing M, Leinhos V. The prompt canvas: a literature-based practitioner guide for creating effective prompts in large language models. arXiv 2024. https://doi.org/10.48550/ARXIV.2412.05127.

[64] Parthasarathy VB, 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 (Version 3). arXiv 2024. https://doi.org/10.48550/ARXIV.2408.13296.

[65] Stamper J, Xiao R, Hou X. Enhancing LLM-based feedback: insights from intelligent tutoring Systems and the Learning Sciences. arXiv 2024. https://doi.org/10.48550/ARXIV.2405.04645.

[66] Martin F, Zhuang M, Schaefer D. Systematic review of research on artificial intelligence in K-12 education (2017–2022). Computers and Education: Artificial Intelligence 2024;6:100195. https://doi.org/10.1016/j.caeai.2023.100195.

[67] Zhai X, Chu X, Chai CS, Jong MSY, Istenic A, Spector M, Liu J-B, Yuan J, Li Y. A review of artificial intelligence (AI) in education from 2010 to 2020. Complexity, 2021;2021(1):8812542. https://doi.org/10.1155/2021/8812542.

[68] Oprea S-V, Bâra A. Transforming education with large language models: trends, themes, and untapped potential. IEEE Access 2025;13:87292–312. https://doi.org/10.1109/ACCESS.2025.3570649.

[69] Snyder H. Literature review as a research methodology: an overview and guidelines. J Bus Res 2019;104:333–9. https://doi.org/10.1016/j.jbusres.2019.07.039.

[70] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, Moher D. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021:n71. https://doi.org/10.1136/bmj.n71.

[71] Henderson M, Bearman M, Chung J, Fawns T, Buckingham Shum S, Matthews KE, De Mello Heredia J. Comparing generative AI and teacher feedback: student perceptions of usefulness and trustworthiness. Assessment & Evaluation in Higher Education 2025:1–16. https://doi.org/10.1080/02602938.2025.2502582.

[72] Xu S, Su Y, Liu K. Investigating student engagement with AI-driven feedback in translation revision: a mixed-methods study. Education and Information Technologies 2025;30(12):16969–95. https://doi.org/10.1007/s10639-025-13457-0.

[73] Lee D-K, Joe I. A GPT-based code review system with accurate feedback for programming education. IEEE Access 2025;13:105724–37. https://doi.org/10.1109/ACCESS.2025.3581139.

[74] Zhao R, Bobrov A, Li J, Aloisi C, He Y. *LearnLens: lLM-enabled personalised, curriculum-grounded feedback with educators in the loop* (Version 4). arXiv 2025. https://doi.org/10.48550/ARXIV.2507.04295.

[75] Qiu Y. The impact of LLM hallucinations on motor skill learning: a case study in badminton. IEEE Access 2024;12:139669–82. https://doi.org/10.1109/ACCESS.2024.3444783.

[76] Jarry Trujillo C, Vela Ulloa J, Escalona Vivas G, Grasset Escobar E, Villagrán Gutiérrez I, Achurra Tirado P, Varas Cohen J. Surgeons vs ChatGPT: assessment and feedback performance based on real surgical scenarios. J Surg Educ 2024:S1931720424001582. https://doi.org/10.1016/j.jsurg.2024.03.012.

[77] Coban A, Dzsotjan D, Küchemann S, Durst J, Kuhn J, Hoyer C. AI support meets AR visualization for Alice and Bob: personalized learning based on individual ChatGPT feedback in an AR quantum cryptography experiment for physics lab courses. EPJ Quantum Technology 2025;12(1):15. https://doi.org/10.1140/epjqt/s40507-025-00310-z.

[78] Hirschi K, Kang O, Yang M, Hansen JHL, Beloin K. Artificial intelligence-generated feedback for second language intelligibility: an exploratory intervention study on effects and perceptions. Lang Learn 2025;75:204–41.

[79] Wang W-S, Lin C-J, Lee H-Y, Huang Y-M, Wu T-T. Enhancing self-regulated learning and higher-order thinking skills in virtual reality: the impact of ChatGPT-integrated feedback aids. Education and information technologies, 30. Scopus; 2025. p. 19419–45. https://doi.org/10.1007/s10639-025-13557-x.

[80] Abiad A, Richter BS, Padilla MSTL, Flores MA, Thompson JC, Lechner PG, Edwards KD, Link RD. AI-assisted grading increases quality feedback and grading efficiency in undergraduate chemistry courses. J Chem Educ 2025;102(11):4936–43. https://doi.org/10.1021/acs.jchemed.5c00664.

[81] Bernard N, Sagawa Jr Y, Bier N, Lihoreau T, Pazart L, Tannou T. Using artificial intelligence for systematic review: the example of elicit. BMC Med Res Methodol 2025;25(1):75. https://doi.org/10.1186/s12874-025-02528-y.

[82] Hilkenmeier F, Pelzer M, Stierle C, Fink-Lamotte J. Evaluating the AI tool "Elicit" as a semi-automated second reviewer for data extraction in systematic reviews: a proof-of-concept. Soc Sci Comput Rev 2025:08944393251404052. https://doi.org/10.1177/08944393251404052.

[83] Bewersdorff A, Seßler K, Baur A, Kasneci E, Nerdel C. Assessing student errors in experimentation using artificial intelligence and large language models: a comparative study with human raters. Computers and Education: Artificial Intelligence 2023;5:100177. https://doi.org/10.1016/j.caeai.2023.100177.

[84] Mayring P. Qualitative content analysis: theoretical background and procedures. In: Bikner-Ahsbahs A, Knipping C, Presmeg N, editors. Approaches to qualitative research in mathematics education. Springer Netherlands; 2015. p. 365–80. https://doi.org/10.1007/978-94-017-9181-6_13.

[85] Dai W, Lin J, Jin F, Li T, Tsai Y-S, Gasevic D, Chen G. Can large language models provide feedback to students? A case study on ChatGPT. EdArXiv 2023. https://doi.org/10.35542/osf.io/hcgzj.

[86] Jia Q, Cui J, Xi R, Liu C, Rashid P, Li R, Gehringer E. On assessing the faithfulness of LLM-generated feedback on student assignments. In: Proceedings of the 17th International Conference on Educational Data Mining; 2024. p. 491–9. https://doi.org/10.5281/ZENODO.12729867.

[87] Steiss J, Tate T, Graham S, Cruz J, Hebert M, Wang J, Moon Y, Tseng W, Warschauer M, Olson CB. Comparing the quality of human and ChatGPT feedback of students' writing. Learn Instr 2024;91:101894. https://doi.org/10.1016/j.learninstruc.2024.101894.

[88] Baral S, Worden E, Lim W-C, Luo Z, Santorelli C, Gurung A, Heffernan N. Automated feedback in math education: a comparative analysis of LLMs for open-ended responses (Version 1). arXiv 2024. https://doi.org/10.48550/ARXIV.2411.08910.

[89] Kinder A, Briese FJ, Jacobs M, Dern N, Glodny N, Jacobs S, Leßmann S. Effects of adaptive feedback generated by a large language model: a case study in teacher education. Computers and Education: Artificial Intelligence 2025;8:100349. https://doi.org/10.1016/j.caeai.2024.100349.

[90] Zhang K. Enhancing critical writing through AI feedback: a randomized control study. Behavioral sciences, 15. Scopus; 2025. https://doi.org/10.3390/bs15050600.

[91] Schunk DH, Zimmerman BJ. Influencing children's self-efficacy and self-regulation of reading and writing through modeling. Reading & Writing Quarterly 2007;23(1):7–25. https://doi.org/10.1080/10573560600837578.

[92] Sun D, Xu P, Zhang J, Liu R, Zhang J. How self-regulated learning is affected by feedback based on large language models: data-driven sustainable development in computer programming learning. Electronics (Basel) 2025;14(1):194. https://doi.org/10.3390/electronics14010194.

[93] Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist 2000;55(1):68–78. https://doi.org/10.1037/0003-066X.55.1.68.

[94] Wigfield A, Eccles JS. Expectancy–Value theory of achievement motivation. Contemp Educ Psychol 2000;25(1):68–81. https://doi.org/10.1006/ceps.1999.1015.

[95] Pekrun R. The control-value theory of achievement emotions: assumptions, corollaries, and implications for educational research and practice. Educ Psychol Rev 2006;18(4):315–41. https://doi.org/10.1007/s10648-006-9029-9.

[96] Chan STS, Lo NPK, Wong AMH. Enhancing university level English proficiency with generative AI: empirical insights into automated feedback and learning outcomes. Contemporary Educational Technology 2024;16(4):ep541. https://doi.org/10.30935/cedtech/15607.

[97] Ruwe T, Mayweg-Paus E. Embracing LLM feedback: the role of feedback providers and provider information for feedback effectiveness. Frontiers in Education 2024;9:1461362. https://doi.org/10.3389/feduc.2024.1461362.

[98] Nguyen HA, Stec H, Hou X, Di S, McLaren BM. Evaluating ChatGPT's decimal skills and feedback generation in a digital learning game. In: Viberg O, Jivet I, Muñoz-Merino PJ, Perifanou M, Papathoma T, editors. Responsive and sustainable educational futures. Responsive and sustainable educational futures, 14200. Springer Nature Switzerland; 2023. p. 278–93. https://doi.org/10.1007/978-3-031-42682-7_19.

[99] Sweller J, Van Merriënboer JJG, Paas F. Cognitive architecture and instructional design: 20 years later. Educ Psychol Rev 2019;31(2):261–92. https://doi.org/10.1007/s10648-019-09465-5.

[100] Chi M. Eliciting self-explanations improves understanding. Cogn Sci 1994;18(3):439–77. https://doi.org/10.1016/0364-0213(94)90016-7.

[101] Escalante J, Pack A, Barrett A. AI-generated feedback on writing: insights into efficacy and ENL student preference. International Journal of Educational Technology in Higher Education 2023;20(1):57. https://doi.org/10.1186/s41239-023-00425-2.

[102] Jamshed M, Ahmed Abu Saleh Md Manjur, Sarfaraj Md, Warda Wahaj Unnisa. The impact of ChatGPT on English language learners' Writing skills: an assessment of AI feedback on mobile. International Journal of Interactive Mobile Technologies (iJIM) 2024;18(19):18–36. https://doi.org/10.3991/ijim.v18i19.50361.

[103] Yu H, Xie Q. Generative AI vs. Teachers: feedback quality, feedback uptake, and revision. Language Teaching Research Quarterly 2025;47:113–37. https://doi.org/10.32038/ltrq.2025.47.07.

[104] Wan T, Chen Z. Exploring generative AI assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning. Physical Review Physics Education Research 2024;20(1):010152. https://doi.org/10.1103/PhysRevPhysEducRes.20.010152.

[105] Davis FD. Perceived usefulness, Perceived ease of use, and user acceptance of information technology. MIS Quarterly 1989;13(3):319–40. https://doi.org/10.2307/249008.

[106] Jürgensmeier L. Generative AI for scalable feedback to multimodal exercises. International Journal of Research in Marketing 2024;41(3):468–88.

[107] Ifenthaler D, Majumdar R, Gorissen P, Judge M, Mishra S, Raffaghelli J, Shimada A. Artificial intelligence in education: implications for policymakers, researchers, and practitioners. Technology, Knowledge and Learning 2024;29(4):1693–710. https://doi.org/10.1007/s10758-024-09747-0.

[108] Long D, Magerko B. What is AI literacy? Competencies and design considerations. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems; 2020. p. 1–16. https://doi.org/10.1145/3313831.3376727.

[109] Jacobsen LJ, Rohlmann J, Weber KE. AI feedback in education: the impact of prompt design and Human expertise on LLM performance. Open Science Framework 2025. https://doi.org/10.31219/osf.io/fx5qz.

[110] Lo N, Wong A, Chan S. The impact of generative AI on essay revisions and student engagement. Computers and Education Open 2025;9:100249. https://doi.org/10.1016/j.caeo.2025.100249.

[111] Thomas, D.R., Borchers, C., Bhushan, S., Gatz, E., Gupta, S., & Koedinger, K.R. (2025). *LLM-generated feedback supports learning if learners choose to use it* (No. arXiv:2506.17006). arXiv. https://doi.org/10.48550/arXiv.2506.17006.

[112] Zhuang Y, Zhao R, Xie Z, Yu PLH. Enhancing language learning through generative AI feedback on picture-cued writing tasks. Computers and Education: Artificial Intelligence 2025;9:100450. https://doi.org/10.1016/j.caeai.2025.100450.

[113] Banihashem SK, Kerman NT, Noroozi O, Moon J, Drachsler H. Feedback sources in essay writing: peer-generated or AI-generated feedback? International Journal of Educational Technology in Higher Education 2024;21(1):23. https://doi.org/10.1186/s41239-024-00455-4.

[114] Er E, Akçapınar G, Bayazıt A, Noroozi O, Banihashem SK. Assessing student perceptions and use of instructor versus AI -generated feedback. British Journal of Educational Technology 2025;56(3):1074–91. https://doi.org/10.1111/bjet.13558.

[115] Kerslake C, Denny P, Smith DH, Leinonen J, MacNeil S, Luxton-Reilly A, Becker BA. Exploring student reactions to LLM-generated feedback on explain in plain English problems. Proceedings of the 56th ACM Technical Symposium on Computer Science Education V 2025;1:575–81. https://doi.org/10.1145/3641554.3701934.

[116] Nazaretsky T, Mejia-Domenzain P, Swamy V, Frej J, Käser T. AI or Human? Evaluating student feedback perceptions in higher education. In: Ferreira Mello R, Rummel N, Jivet I, Pishtari G, Ruipérez Valiente JA, editors. Technology enhanced learning for inclusive and equitable quality education. Technology enhanced learning for inclusive and equitable quality education, 15159. Springer Nature Switzerland; 2024. p. 284–98. https://doi.org/10.1007/978-3-031-72315-5_20.
[117] Scholz N, Nguyen MH, Singla A, Nagashima T. *Partnering with AI: a pedagogical feedback system for LLM integration into programming education* (Version 3). arXiv 2025. https://doi.org/10.48550/ARXIV.2507.00406.
[118] Çağlar-Özhan Ş, Tekeli P, Arkün-Kocadere S. Comparison of AI-generated and instructor feedback: no significant difference in perceived feedback quality and neither on performance. Journal of Computer Assisted Learning 2025;41(5): e70134. https://doi.org/10.1111/jcal.70134.
[119] Zhang Z, Dong Z, Shi Y, Price T, Matsuda N, Xu D. Students' Perceptions and preferences of generative artificial intelligence feedback for programming. Proceedings of the AAAI Conference on Artificial Intelligence 2024;38(21): 23250–8. https://doi.org/10.1609/aaai.v38i21.30372.
[120] Jovic M, Papakonstantinidis S, Kirkpatrick R. From red ink to algorithms: investigating the use of large language models in academic writing feedback. Language Testing in Asia 2025;15(1):59. https://doi.org/10.1186/s40468-025-00389-2.
[121] Heo S, Son S, Park H. HaluCheck: explainable and verifiable automation for detecting hallucinations in LLM responses. Expert Syst Appl 2025;272:126712. https://doi.org/10.1016/j.eswa.2025.126712.
[122] Cukurova M. The interplay of learning, analytics and artificial intelligence in education: a vision for hybrid intelligence. British Journal of Educational Technology 2025;56(2):469–88. https://doi.org/10.1111/bjet.13514.
[123] 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 (Version 6). arXiv 2022. https://doi.org/10.48550/ARXIV.2201.11903.
[124] Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, Dai Y, Sun J, Wang M, Wang H. Retrieval-augmented generation for large language models: a survey (Version 5). arXiv 2023. https://doi.org/10.48550/ARXIV.2312.10997.
[125] Schick T, Dwivedi-Yu J, Dessì R, Raileanu R, Lomeli M, Zettlemoyer L, Cancedda N, Scialom T. Toolformer: language models can teach themselves to use tools (Version 1). arXiv 2023. https://doi.org/10.48550/ARXIV.2302.04761.
[126] Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan A-W, Cronin E, Decullier E, Easterbrook PJ, Von Elm E, Gamble C, Ghersi D, Ioannidis JPA, Simes J, Williamson PR. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. PLoS ONE 2008;3(8):e3081. https://doi.org/10.1371/journal.pone.0003081.
[127] Hui V, Guan S, Feng X. Development and quasi-experimental evaluation of a large language model-based automated feedback system for nursing innovation pitches. Nurse Educ Pract 2026;90:104672. https://doi.org/10.1016/j.nepr.2025.104672.
[128] Xavier C, Da Costa NT, Valdo AK, Alves G, Rodrigues L, Rodrigues LF, Silva M, Neto R, Falcão TP, Gasevic D, Mello RF. Human teacher vs. LLM-generated feedback in secondary education: a comparative study on student perceptions. In: Tammets K, Sosnovsky S, Mello RFerreira, Pishtari G, Nazaretsky T, editors. Two decades of TEL. from lessons learnt to challenges ahead. Two decades of TEL. from lessons learnt to challenges ahead, 16063. Springer Nature Switzerland; 2026. p. 534–48. https://doi.org/10.1007/978-3-032-03870-8_36.
[129] Na Nongkhai L, Wang J, Wynn A, Mendori T. Evaluating adaptive and generative AI-based feedback and recommendations in a knowledge-graph-integrated programming learning system. CAEAI 2026;10:100526. https://doi.org/10.1016/j.caeai.2025.100526.
[130] Soyoof A, Reynolds BL, Rassaei E, Kao C-W, Van Ha X. From teachers to chatbots: Scaffolded corrective feedback and student trust in online L2 English classrooms. CAEAI 2026;10:100530. https://doi.org/10.1016/j.caeai.2025.100530.
[131] Liebenow LW, Schmidt FTC, Meyer J, Fleckenstein J. Self-assessment accuracy in the age of artificial Intelligence: Differential effects of LLM-generated feedback. Comput Educ 2025;237:105385. https://doi.org/10.1016/j.compedu.2025.105385.
[132] Farrokhnia M, Latifi S, Papadopoulos PM, Hogenkamp L, Gijlers H, Khosravi H, Noroozi O. Generative AI offers more, but students revise less: Comparing the effects of teacher and AI feedback on student essay revisions. Int J Educ Technol High Educ 2026;23(1). https://doi.org/10.1186/s41239-026-00579-9.
[133] De Sixte R, Mañá A, Ávila V, Sánchez E. Warm elaborated feedback. Exploring its benefits on post-feedback behaviour. Educ Psyc 2020;40(9):9.
[134] Yan L, Sha L, Zhao L, Li Y, Martinez-Maldonado R, Chen G, Li X, Jin Y, Gašević D. Practical and ethical challenges of large language models in education: A systematic scoping review. Br J Educ Technol 2024;55(1):90–112. https://doi.org/10.1111/bjet.13370.
[135] Lee J, Hicke Y, Yu R, Brooks C, Kizilcec RF. The life cycle of large language models in education: A framework for understanding sources of bias. Br J Educ Technol 2024;55(5):1982–2002. https://doi.org/10.1111/bjet.13505.
[136] Delikoura, I., Fung, Yi. R., Hui, P., From Superficial Outputs to Superficial Learning: Risks of Large Language Models in Education (Version 2). (2025) arXiv. https://doi.org/10.48550/ARXIV.2509.21972.
[137] Lin Z, Guan S, Zhang W, Zhang H, Li Y, Zhang H. Towards trustworthy LLMs: A review on debiasing and dehallucinating in large language models. Artif Intell Rev 2024;57(9):243. https://doi.org/10.1007/s10462-024-10896-y.
[138] Atkinson RK, Derry SJ, Renkl A, Wortham D. Learning from Examples: Instructional Principles from the Worked Examples Research. Rev Educ Res 2000; 70(2):181–214. https://doi.org/10.3102/00346543070002181.
[139] Chen S, Chen W. AI-Driven Intelligent Feedback System for Enhancing Self-Assessment Accuracy in Higher Education Writing. ES 2026;43(1):e70184. https://doi.org/10.1111/exsy.70184.
[140] Alsaiari O, Baghaei N, Lahza H, Lodge JM, Boden M, Khosravi H. Emotionally enriched AI-generated feedback: Supporting student well-being without compromising learning. Comput Educ 2025;239:105363. https://doi.org/10.1016/j.compedu.2025.105363.
[141] Liu J, Shi Z, Li W. Developing students' feedback literacy in disciplinary academic writing through generative artificial intelligence. Assmt writing 2026;68:101030. https://doi.org/10.1016/j.asw.2026.101030.
[142] Asadi M, Ebadi S, Mohammadi L. The impact of integrating ChatGPT with teachers' feedback on EFL writing skills. TSC 2025;56:101766. https://doi.org/10.1016/j.tsc.2025.101766.
[143] Algobaei F, Alzain E. Prompt engineering for non-native English learners: A generative AI approach to personalised language feedback. J Stud Soc Sci Humanit 2026;13:102341. https://doi.org/10.1016/j.ssaho.2025.102341.
[144] Maqbali, A.A., Al-Hanshi, S., Ghaithi, A.A., Behforouz, B., Reimagining Writing Feedback: The Effect of Copilot Artificial Intelligence (AI) on English Foreign Language (EFL) Students' Written Performance. Int J Learn Teach Educ Res, (2026) 25(1), 94–110. Scopus. doi:10.26803/ijlter.25.1.5.
[145] Tran, T.T.T., Enhancing EFL Writing Revision Practices: The Impact of AI- and Teacher-Generated Feedback and Their Sequences. Educ Sci, (2025) 15(2). doi:10.3390/educsci15020232.
[146] Hofmann F, Daunicht T-M, Plößl L, Gläser-Zikuda M. Promoting Reflection Skills of Pre-Service Teachers—The Power of AI-Generated Feedback. Educ Sci 2025;15 (10). https://doi.org/10.3390/educsci15101315.
[147] Hao H, Razali AB, Zuo R. Exploring the Impact of Integrated AWE and Generative AI Feedback on Chinese EFL Undergraduates' Higher-Order Thinking in Argumentative Writing. SAGE Open 2026;16(1). https://doi.org/10.1177/21582440251413884.
[148] Banihashem SK, Kerman NT, Noroozi O, Moon J, Drachsler H. Feedback sources in essay writing: Peer-generated or AI-generated feedback? Int J Educ Technol High Educ 2024;21(1):23. https://doi.org/10.1186/s41239-024-00455-4.
[149] Li Y, Shan Z, Raković M, Guan Q, Gašević D, Chen G. When AI explains in natural language: Unveiling the impact of generative AI explanations on educators' grading and feedback practices. Educ inf technol 2025;30(17):24931–64. https://doi.org/10.1007/s10639-025-13741-z.
[150] Çağlar-Özhan Ş, Tekeli P, Arkün-Kocadere S. Comparison of AI-Generated and Instructor Feedback: No Significant Difference in Perceived Feedback Quality and Neither on Performance. Journal of Computer Assisted Learning 2025;41(5): e70134. https://doi.org/10.1111/jcal.70134.
[151] Noroozi O, Haddadian G, Gao X, Schunn C, Alqassab M, Banihashem SK. The Value of GenAI for Peer Feedback Provision: Student Perceptions and Impacts. International Journal of Educational Technology in Higher Education 2025:22.
[152] Timmers CF, Braber-van Den Broek J, Van Den Berg SM. Motivational beliefs, student effort, and feedback behaviour in computer-based formative assessment. Computers & Education 2013;60(1):25–31. https://doi.org/10.1016/j.compedu.2012.07.007.

Further reading

[153] Biggs JB, Collis KF. Evaluating the quality of learning: the solo taxonomy (structure of the observed learning outcome). Academic press; 1982.
[154] Nguyen HH. ChatGPT-assisted language learning: effects on Vietnamese english majors' Writing skills and motivation. JALT CALL Journal 2025;21(2).
[155] Pozdniakov S, Brazil J, Banihashem SK, Noroozi O, Gašević D, Sadiq S, Khosravi H. AI assistance in peer feedback provision: pedagogically sound, but minimally adopted. Comput Educ 2026:105591. https://doi.org/10.1016/j.compedu.2026.105591.