A Comparative Review of Experimental Studies on Learning Outcomes: Teacher Responses Versus Generative AI Responses

Wendi Yu *

School of Foreign Languages, East China University of Science and Technology, Shanghai, 201424, China

* Corresponding Author Email: 23012580@mail.ecust.edu.cn

Abstract. Advances in artificial intelligence (AI) have substantially broadened the scope of students' inquiry, enabling them to engage in questioning through more varied and complex modalities. This paper presents a comparative review of recent experimental and meta-analytic studies examining the effects of teacher responses and generative AI responses on student learning outcomes. Drawing on evidence from writing instruction, general academic performance, and higher-order cognitive tasks, the review synthesizes findings across cognitive, affective, and motivational dimensions. Results indicate that generative AI responses—particularly from large language models—can produce substantial short-term gains in performance and comprehension, especially in structured tasks requiring rapid feedback. However, teacher responses remain more effective for fostering deep conceptual understanding, knowledge transfer, and emotional support. The discussion highlights theoretical explanations grounded in constructivism, trust models, and productive-struggle theory, and argues for hybrid instructional designs that integrate AI feedback with teacher scaffolding. Recommendations for future research emphasize longitudinal designs, discipline-specific analyses, and strategies for improving the accuracy and socio-emotional sensitivity of AI feedback.

Keywords: Teacher Responses, Generative AI Responses, hybrid instructional designs, student learning outcomes.

1. Introduction

1.1. Background and significance

From Socrates' maieutics, which used questioning to lead students into self-reflection and contradiction, to modern constructivism, where students are encouraged to actively generate questions as a way of deepening their understanding, the act of questioning has always played a central role in education. Before the digital era, students relied primarily on teachers and books as their main knowledge sources, and the act of questioning teachers was regarded as a key pathway to learning. With the advent of the internet and, more recently, generative artificial intelligence (AI), this landscape has shifted: learners can now obtain answers instantly from AI systems such as ChatGPT, and the teacher's role as the primary respondent is being redefined. However, it remains unclear how responses from teachers and those generated by AI differ in shaping students' learning outcomes. This question is at the heart of the present review.

Learning outcomes are broadly defined as the measurable changes in students' knowledge, skills, and affective states resulting from educational interventions. At the most basic level, these outcomes include factual recall and conceptual understanding, often assessed through immediate post-tests. More advanced outcomes involve knowledge transfer and application, such as the ability to use acquired knowledge in new contexts. Beyond cognitive aspects, learning outcomes also encompass higher-order thinking skills, including critical thinking, problem-solving, and creativity. Finally, emotional and motivational dimensions—such as confidence, engagement, or reduced anxiety—are also integral, as they shape learners' long-term academic trajectories.

This review takes an explicitly comparative perspective, examining empirical studies that have contrasted teacher responses and AI-generated responses in relation to these outcomes. Teachers

contribute through adaptive feedback, pedagogical judgment, and emotional support, while AI systems offer efficiency, immediacy, and access to extensive informational resources. By systematically reviewing recent experimental studies, this paper seeks to clarify their respective advantages, limitations, and potential complementarities in supporting student learning.

1.2. Research questions

This review aims to systematically analyze and compare existing experimental evidence on the response methods of teachers and AI.

RQ: Based on experimental studies, what are the relative strengths and weaknesses of teacher responses and AI responses in promoting different student learning outcomes?

2. Review of literature: Teacher responses

2.1. Definition of teacher responses

Teacher responses refer to the verbal and non-verbal feedback that instructors provide to questions and learning behaviors. These responses range from cognitive scaffolding and diagnostic explanations to affective encouragement and metacognitive prompts. In contrast to automated responses, teacher feedback is grounded in pedagogical judgment, classroom context, and relational dynamics. Theoretical lenses such as constructivism highlight teachers' role in guiding learners through active knowledge construction, while the Technology Acceptance Model (TAM) and trust-based models of learning support emphasize how perceived credibility and relational trust in teachers influence students' willingness to engage with feedback. The social presence framework further positions teacher responses as a means of fostering interpersonal connectedness, which is critical for motivation and engagement in both face-to-face and online environments.

2.2. Strengths of teacher responses

Teachers can provide cognitive support and diagnostic scaffolding. Teachers provide feedback that is temporally and functionally targeted: effective classroom feedback identifies a gap between current and desired performance, diagnoses sources of error, and suggests next steps or strategies rather than merely supplying the correct answer. Foundational syntheses and meta-reviews characterize the mechanisms and potency of such feedback [1-3]. Empirical and theoretical work highlights that feedback is most powerful when it (a) makes goals explicit ("feed-up"), (b) points to discrepancies ("feed-back"), and (c) offers feed-forward guidance that helps students adopt strategies for future problems [4]. Meanwhile, tutoring research shows that contingent questioning and step-wise scaffolding—hallmarks of skilled teachers—produce larger learning gains than non-contingent instruction [5, 6].

Teachers can provide supporting durable learning. Teachers can deliberately design practice that promotes retrieval, spacing and desirable difficulties—techniques shown in cognitive psychology to support durable learning [7, 8]. In classroom practice, experienced teachers orchestrate sequencing and feedback to encourage productive struggle and retrieval practice.

Teachers can provide effective and social functions. Teacher responses do more than transmit information; they shape the social and emotional climate that enables risk-taking and persistence. Meta-analytic work links affective quality in teacher-student relationships to greater engagement and achievement [9]. First-grade and early-years research further shows that classrooms with higher emotional support produce better trajectories for at-risk children [10]. Reviews of learner-centred teaching likewise find positive effects of teacher responsiveness on motivation and willingness to re-attempt hard tasks [11]. A recent systematic review of online peer feedback practices in higher education highlighted that cognitive feedback, structured presentation modes, and integration of multiple feedback sources are critical for enhancing students' task performance and engagement [12].

Teachers can provide metacognitive and self-regulated learning support. Teachers cue metacognition—prompting students to reflect on strategy use, plan, monitor, and evaluate their work. Classic guidance for "good feedback practice" emphasises prompts that foster self-assessment and regulation. When teachers explicitly teach and model metacognitive strategies, students become better at monitoring comprehension and selecting subsequent strategies. Recent research in higher education highlights that implementing structured assessment and feedback drivers—including policy-aligned guidance, rubric-based evaluations, and formative feedback—effectively enhances students' self-regulation, engagement, and academic performance [13].

Taken together, the literature suggests teacher responses combine diagnostic cognition (tailored correction), strategy scaffolding (procedural cues), and socio-emotional support; these three functions interact to produce learning gains that go beyond immediate task completion.

2.3. Limitations of teacher responses

Teacher responses have coverage and timeliness constraints. A practical limit is simple bandwidth: teachers cannot provide detailed, individualized, timely responses to every student across all tasks in large classes or under resource constraints. Observational and policy studies show that class size and time allocation shape how often and how well teachers interact one-to-one with students [14]. Bloom's classic 2-sigma observation underscores the potency of one-to-one tutoring but also the cost and scalability problem for classrooms.

Teacher responses may have variability, bias, and bounded knowledge. Teacher responses are shaped by beliefs, expectations, and local context. Econometric and field studies document systematic variation and some evidence of biased assessments tied to students' backgrounds [15]. Teachers' subject-matter knowledge limits also constrain the depth and accuracy of their responses in highly specialised topics, especially outside their core expertise.

Teacher responses may lack consistency and reliability. Different teachers (or the same teacher on different days) may reply to similar student questions in inconsistent ways; this heterogeneity affects the predictability of instruction and can introduce inequities in student opportunity. Recent evidence from tertiary EFL contexts indicates that student engagement with teacher feedback is generally consistent across affective, behavioral, and cognitive dimensions, but variability in feedback delivery across instructors can still influence perceived fairness and trust in the learning process [16].

Time devoted to social-emotional support is essential, but in contexts of heavy curricular demand teachers sometimes face trade-offs between depth of cognitive scaffolding and breadth of coverage. Empirical work highlights that these trade-offs vary by grade and subject.

In summary, teacher responses remain central because they integrate diagnosis, scaffolding, metacognitive support and socio-emotional care—attributes that are tightly bound to human judgement and relationship-building. But these strengths are tempered by scalability, time, bias, and variability concerns that motivate interest in complementary or augmentative tools.

3. Review of literature: Generative AI responses

3.1. Definition of generative AI responses

Generative AI responses are textual or multimodal feedback produced by models such as ChatGPT or other large language models (LLMs) trained on vast data corpora. They generate output based on statistical pattern matching rather than lived or contextualized understanding. Strengths include high information density, immediacy of response, consistency across queries, and the capacity to scale to many learners. Weaknesses arise from a lack of affective sensitivity, inability to detect unspoken confusion, and risks of factual error or "hallucination" [17]. From a trust model lens, the credibility of AI responses depends on transparency, reliability, and students' previous experience with AI.

3.2. Advantages of generative AI responses

Empirical studies show several positive effects of generative AI responses on learning outcomes. A recent meta-analysis by Wang & Fan found that ChatGPT has a large positive impact (g = 0.867) on the learning performance of students across many studies between November 2022 and February 2025; also moderate positive effects for higher-order thinking (g = 0.457) and learning perceptions (g = 0.456) [18,19]. This suggests that when used appropriately, AI responses can meaningfully improve cognitive outcomes, especially in performance and problem-solving contexts.

In writing instruction contexts, automated writing evaluation (AWE) systems, which are a subtype of generative AI feedback, have been applied to improve students' academic writing skills. Fleckenstein, Liebenow, and Meyer conducted a multi-level meta-analysis indicating that AWE systems lead to statistically significant improvements in writing performance, particularly in grammar, coherence, and structure, though effect sizes vary by task and feedback specificity [20].

Additionally, Meyer et al. studied LLM-based feedback in classroom settings and found improvements not only in cognitive outcomes (task performance) but also in affective-motivational outcomes (students reported higher confidence and lower anxiety when receiving AI feedback under scaffolded conditions) [21].

Generative AI responses show access, immediacy, and scale. Large language models are available on demand and can respond instantly to many users simultaneously; this accessibility addresses the bandwidth shortfall described above. Work in real classrooms shows that well-designed LLM-based tutors or chat interfaces can deliver frequent practice and immediate corrective feedback at scale [22,23].

LLMs can be prompted or engineered to adapt language complexity, scaffold steps, and reframe explanations. Several experimental implementations report that GPT-based homework tutors or chat assistants increase engagement and, for users who adopt them, can raise short-term performance [24]. Recent systematic syntheses and meta-analyses indicate an overall positive average effect of ChatGPT/LLM interventions on immediate academic performance and some affective outcomes, albeit with high heterogeneity across contexts and designs.

Generative AI responses need lower social cost for questioning. Qualitative and experimental work suggests students sometimes feel more willing to ask "embarrassing" or low-stakes questions in an anonymous AI interface; that reduced anxiety can increase help-seeking and iterative practice.

AI systems can deliver standardized hints, keep logs of student attempts, and produce analytics that help instructors triage support. When integrated into curricula with teacher oversight, this can extend the teacher's reach and make formative data visible at scale.

3.3. Risks and challenges of AI responses

A central empirical finding is that LLMs sometimes generate incorrect or fabricated statements with high fluency (so-called "hallucinations"). Survey and empirical work documents the prevalence, taxonomy, and mitigation efforts for hallucinations; this remains a primary reliability concern in educational deployments.

Generative AI may lay more emphasis on surface fluency rather than deep diagnostic understanding. LLMs produce plausible explanations but do not possess human comprehension of a learner's affective state or the causal chains that produced an error. Several experiments find that while LLM assistance can raise short-term task performance for users who adopt it, unguided access can reduce long-term learning gains if the system supplies answers that short-circuit productive struggle [25].

Models inherit training-data biases and can reproduce or amplify stereotypes. Position papers and empirical audits warn that unmoderated LLM output can reproduce harmful patterns; mitigation requires both technical and governance strategies.

Field experiments indicate a paradox: easy access to answer-giving systems can depress engagement (fewer attempts, lower persistence) even as adopters sometimes benefit in the short run. Bastani and colleagues show that without "guardrails" designed to preserve learning (e.g., forcing

stepwise reasoning, hint-only modes, or teacher-designed scaffolds), reliance on generative AI can harm acquisition of underlying skills when AI access is later removed.

The effectiveness of an AI helper depends heavily on interface design, prompting strategy, and integration with teacher practice. Studies that embed LLMs into curriculum with explicit pedagogical constraints (prompt engineering to provide hints rather than answers; teacher oversight) report far fewer negative effects and sometimes net learning benefits.

Generative LLMs offer scalable, rapid, and—when carefully engineered—pedagogically useful responses. However, the literature consistently stresses conditionality: benefits accrue under deliberate design (guardrails, hint policies, teacher integration). Left unconstrained, LLMs' fluency, occasional factual errors, and potential to short-circuit productive struggle present tangible risks to durable learning.

4. Discussion

In summary, both sources bring unique strengths. Generative AI excels at efficiency and scalability — offering instant, detailed information whenever asked —which often translates into quick learning of factual content. Human teachers excel at social-emotional and adaptive support-understanding a student's misunderstandings and nurturing engagement. A useful way to see this is as a complementarity matrix: teachers provide encouragement, context, and deep scaffolding, whereas AI provides speed, breadth of facts, and anonymity for shy students. In practice, this implies that an optimal learning ecosystem will leverage both. For example, instructional design might have students first use AI to gather information or draft ideas, then engage in a guided teacher-led discussion to explore nuances. Teachers can adopt an "AI-augmented" role, fact-checking AI answers, curating high-quality AI content, and using it as a springboard for deeper dialogue.

These findings have practical implications. Educators should teach students AI literacy and critical questioning: simply having ChatGPT answers available is not enough without critical appraisal. Likewise, instructors should develop "AI pedagogy" - for instance, using AI to handle routine queries or generate examples, freeing them to focus on higher-level coaching. Crucially, the goal is human-machine synergy. Indeed, one experimental feedback study concludes that hybrid models combining AI's efficiency and teachers' expertise are likely the most effective.

This review also highlights research gaps. Most existing experiments are short-term and in controlled settings; longitudinal studies are needed to see how benefits persist. Studies should diversify beyond university settings (to K-12, vocational, and special education contexts) and across disciplines. We also need to explore beyond text: voice-based AI tutors or multimodal AI might offer richer interaction. Future work should investigate student differences in depth and systematically test collaborative models to establish best practices.

5. Conclusion

In sum, recent empirical evidence underscores that generative AI responses hold substantial promise for enhancing learning outcomes, especially in performance, retention, and certain higher-order thinking tasks, when used appropriately. However, teacher responses remain essential for deep conceptual learning, emotional support, and sustained motivation. The strongest outcomes emerge in contexts where AI feedback is integrated into instructional designs that preserve teacher mediation and encourage reflection.

Future research should deepen our understanding of how long-term learning outcomes evolve under AI vs teacher feedback, especially in K-12 settings; explore more varied disciplines; and examine how multimodal and emotionally enriched AI feedback might close the gap in affective and metacognitive outcomes.

References

[1] J. Hattie and H. Timperley, The Power of Feedback. Rev. Educ. Res. 77, 81-112 (2007).

[2] V.J. Shute, Focus on Formative Feedback. Rev. Educ. Res. 78, 153-189 (2008).

[3] P. Black and D. Wiliam, Assessment and Classroom Learning. Assess. Educ. 5, 7-74 (1998).

[4] D.J. Nicol and D. Macfarlane-Dick, Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. Stud. High. Educ. 31, 199-218 (2006).

[5] B.S. Bloom, The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. Educ. Res. 13, 4-16 (1984).

[6] K. VanLehn, The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educ. Psychol. 46, 197-221 (2011).

[7] J.D. Karpicke and J.R. Blunt, Retrieval practice produces more learning than elaborative studying with concept mapping. Science 331, 772-775 (2011).

[8] R.A. Bjork, J. Dunlosky, and N. Kornell, Self-regulated learning: Beliefs, techniques, and illusions. Annu. Rev. Psychol. 64, 417-444 (2013).

[9] D.L. Roorda, H.M.Y. Koomen, J.L. Spilt, and F.J. Oort, The influence of affective teacher-student relationships on students' school engagement and achievement: A meta-analytic approach. Rev. Educ. Res. 81, 493-529 (2011).

[10] B.K. Hamre and R.C. Pianta, Can instruction and emotional support in the first-grade classroom make a difference for children at risk of school failure? Child Dev. 76, 949-967 (2005).

[11] J. Cornelius-White, Learner-centered teacher-student relationships are effective: a meta-analysis. Rev. Educ. Res. 77, 113-143 (2007).

[12] X. Gao, O. Noroozi, J. Gulikers, H.J. Biemans, and S.K. Banihashem, A systematic review of the key components of online peer feedback practices in higher education. Educ. Res. Rev. 42, 100588 (2024).

[13] K. Gomis, M. Saini, M. Arif, and C. Pathirage, Enhancing the assessment and the feedback in higher education. Qual. Assur. Educ. 32, 165-179 (2024).

[14] P. Blatchford, P. Bassett, and P. Brown, Examining the effect of class size on classroom engagement and teacher-pupil interaction. Learn. Instr. 21, 715-730 (2011).

[15] S. Burgess and E. Greaves, Test scores, subjective assessment and stereotyping of ethnic minorities. Inst. Fiscal Stud. Rep. (2009).

[16] X. Cheng, Y. Liu, and C. Wang, Understanding student engagement with teacher and peer feedback in L2 writing. System 119, 103176 (2023).

[17] Z. Ji, N. Lee, R. Frieske, et al., Survey of hallucination in natural language generation. ACM Comput. Surv. (2023).

[18] J. Wang and W. Fan, The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis. Humanit. Soc. Sci. Commun. 12, 621 (2025).

[19] X. Cheng, Y. Liu, and C. Wang, Understanding student engagement with teacher and peer feedback in L2 writing. System 119, 103176 (2023).

[20] J. Fleckenstein, L. Liebenow and J. Meyer, Automated feedback and writing: A multi-level meta-analysis of effects on students' performance. Front. Artif. Intell. 6, 1162454 (2023).

[21] J. Meyer, et al., Using LLMs to bring evidence-based feedback into the classroom: impacts on cognitive and affective outcomes. Comput. Educ. 196, 104765 (2024).

[22] A. Nie, Y. Chandak, M. Suzara, M. Ali, J. Woodrow, M. Peng, M. Sahami, E. Brunskill, and C. Piech, The GPT Surprise: Offering large language model chat in a massive coding class reduced engagement but increased adopters' exam performances. arXiv:2407.09975 (2024)

[23] A. Vanzo, S. Pal Chowdhury, and M. Sachan, GPT-4 as a Homework Tutor Can Improve Student Engagement and Learning Outcomes. arXiv:2409.15981 (2024)

[24] R. Deng, M. Jiang, X. Yu, Y. Lu, and S. Liu, Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. Comput. Educ. 227, 105224 (2025).

[25] H. Bastani, O. Bastani, A. Sungu, H. Ge, Ö. Kabakcı, and R. Mariman, Generative AI without guardrails can harm learning: Evidence from high-school mathematics. Proc. Natl. Acad. Sci. U.S.A. 122, e2422633122 (2025).