frontiers | Frontiers in Education

TYPE Perspective
PUBLISHED 09 October 2025
DOI 10.3389/feduc.2025.1697554

OPEN ACCESS

EDITED BY
Sergio Ruiz-Viruel,
University of Malaga, Spain

REVIEWED BY
Anitia Lubbe,
North-West University, South Africa

*CORRESPONDENCE
Hayato Tomisu
✉ tomisu@fc.ritsumei.ac.jp

RECEIVED 02 September 2025
ACCEPTED 18 September 2025
PUBLISHED 09 October 2025

CITATION
Tomisu H, Ueda J and Yamanaka T (2025) The cognitive mirror: a framework for AI-powered metacognition and self-regulated learning. *Front. Educ.* 10:1697554. doi: 10.3389/feduc.2025.1697554

The cognitive mirror: a framework for AI-powered metacognition and self-regulated learning

Hayato Tomisu[1,2]*, Junya Ueda[3] and Tsukasa Yamanaka[4]

[1]Ritsumeikan Global Innovation Research Organization, Ritsumeikan University, Osaka, Japan, [2]Graduate School of Data Science, Shiga University, Shiga, Japan, [3]ImpactLab, Shiga, Japan, [4]Department of Biotechnology, College of Life Sciences, Ritsumeikan University, Shiga, Japan

Introduction: The dominant paradigm of generative artificial intelligence (AI) in education positions it as an omniscient oracle, a model that risks hindering genuine learning by fostering cognitive offloading.
Objective: This study proposes a fundamental shift from "AI as Oracle" model to a "Cognitive Mirror" paradigm, which reconceptualizes AI as a teachable novice engineered to reflect the quality of a learner's explanation. The core innovation is the repurposing of AI safety guardrails as didactic mechanisms to deliberately sculpt AI's ignorance, creating a "pedagogically useful deficit." This conceptual shift enables a detailed implementation of the "learning by teaching" principle.
Method: Within this paradigm, a framework driven by a Teaching Quality Index is introduced. This metric assesses the learner's explanation and activates an instructional guidance level to modulate the AI's responses, from feigning confusion to asking clarifying questions.
Results: Grounded in learning science principles, such as the Protégé Effect and Reflective Practice, this approach positions the AI as a metacognitive partner. It may support a shift from knowledge transfer to knowledge construction, and a re-orientation from answer correctness to explanation quality in the contexts we describe.
Conclusion: By re-centering human agency, the "Cognitive Mirror" externalizes the learner's thought processes, making their misconceptions objects of repair. This study discusses the implications on assessment, addresses critical risks, including algorithmic bias, and outlines a research agenda for a symbiotic human-AI coexistence that promotes effortful work at the heart of deep learning.

KEYWORDS

teachable agents, reflective practice, role inversion, instructional guidance level, instructional scaffolding, knowledge scope control, protégé effect, generative AI

1 Introduction

Powerful large language models (LLMs) are transforming education technology. It is not limited to individual educational institutions and teachers (Almuhanna, 2024; Zhai, 2022; Chen et al., 2020). Second language acquisition (SLA) quickly adopted this technology, demonstrating its suitability for language tasks (Tsai et al., 2025; Fathi et al., 2025; Zhang and Huang, 2024; Chiu et al., 2024; Wei, 2023). As of 2025, chat tools provided by OpenAI and Google offer an "education mode" (OpenAI, 2025; Google, 2025). To support efficient learning, some tools offer learning experiences based on the Socratic method, which use Socratic

learning in dialog; whereby, LLM has been proven effective in language learning and varied fields, such as law, medicine, and mathematics (Xie et al., 2025; Adewumi et al., 2025; Favero et al., 2024; Yong et al., 2024).

However, this dominant model presents a fundamental pedagogical problem. It treats artificial intelligence (AI) as an all-knowing oracle. This omniscience, the source of AI's power, ironically becomes a barrier to genuine learning. We term this issue "knowledge scope misalignment." It happens when AI tries to be a helpful conversationalist and responds beyond the learner's curriculum. This misalignment of knowledge scope was especially noted in AI-driven language learning (Li, 2024; Schmidt and Strasser, 2022). It can overwhelm and confuse the learner and derail the focused acquisition of specific skills.

Furthermore, the "AI as Oracle" design fosters cognitive offloading. When a perfect answer is always available, learners reallocate effort from internal computation to tool use (Dror and Harnad, 2008). While the broader AI deployment amplifies access and efficiency, it simultaneously induces cognitive dependency that suppresses active recall and problem-solving (Jose et al., 2025; Gerlich, 2025). Consequently, learners are less likely to engage in effortful but essential processes, such as retrieval practice, knowledge reconstruction, and error analysis (Lee et al., 2025; Fan et al., 2024). At the mechanistic level, neural and behavioral evidence suggests the accrual of "cognitive debt" (Kosmyna et al., 2025). However, current integrations privilege efficiency over learners' epistemic agency, normalizing quick answer-taking rather than verification, sense-making, and productive struggle (Chen, 2025; Jose et al., 2025). In such Oracle-style pipelines, learning devolves into a passive, one-way flow of information from AI to humans; a structure several conventional AI tutoring systems instantiate.

In contrast, educational research has long championed "learning by teaching" (Martin and Oebel, 2007). This phenomenon is called the "Protégé Effect." The act of explaining a concept compels the teacher to structure their thoughts. Early research showed that students who acted as tutors gained a deeper understanding of the subjects (Allen, 1967). Furthermore, although involving young children, it is evident that encouraging explanation promotes transfer, and whether the child is being listened to is an important factor (Rittle-Johnson et al., 2008). It leads to greater and persistent understanding (Fiorella and Mayer, 2013, 2014). Research shows that these benefits exist even when the tutee is a computer agent or a virtual avatar (Chase et al., 2009; Okita and Schwartz, 2013a; Okita et al., 2013b). Additionally, several studies have proposed approaches that emphasize learning by teaching through interactive agents (Love et al., 2022; Chhibber and Law, 2019). However, it has been challenging for agents to respond to various subjects dynamically.

Currently, powerful generative AI and the proven pedagogical method of learning by teaching co-exist. Studies have reported the effectiveness of prompt set-ups that position LLMs as "teachable students" (Chen et al., 2025). Furthermore, in the previous "nurturing AI" prototype, the authors proposed this path with a controlled-forgetting design. Hence, the AI agent's competence is reflected in the prompt and database (Tomisu et al., 2025). However, control using prompts alone cannot escape the "AI as Oracle" paradigm. This is because technical issues, such as memory and token limits (Maharana et al., 2024) or jailbreaking by prompt (Anthropic, 2024), can cause the system to drift from the intended design (Choi et al., 2024). This reveals a significant gap. AI agents are built to provide knowledge; however, not to serve as a vessel for learners' knowledge construction.

This study argues that one should not relentlessly pursue a better AI tutor. Instead, one should focus on designing a better AI student. Hence, it introduces the "Cognitive Mirror paradigm," a novel approach that inverts the traditional human-AI relationship. Its core innovation is a Diversion Guardrail Mechanism, which deliberately limits the AI's knowledge. It intentionally limits the AI's accessible knowledge, positioning it as a "teachable" entity. It becomes a clear mirror for self-reflection. This study outlines the conceptual framework and describes an initial prototype implementation. Moreover, it presents insights from an initial classroom deployment and argues for a future where educational AI would be defined by people's ability to control it thoughtfully. The terms used in this paper were defined in detail in Supplementary materials 1.1.

2 From oracle to mirror: AI as a teaching-quality checker

Cognitive Mirror is an AI-usage framework designed to reflect a learner's current state of knowledge. It shifts the AI's role from a content provider to an instrument for assessing teaching quality. Figure 1 presents an overview of the pipeline, which integrates the session loop, the Teaching Quality Index (TQI) conditioned instructional guidance level (M0–M3), a scope guardrail, and long-term profiling.

2.1 The core interaction loop

Interaction with the Cognitive Mirror follows a four-step loop, driven by a lightweight TQI that measures explanation quality and acts as an instructional guidance level for the AI:

- Present: The learner, acting as the teacher, explains a concept to AI, which acts as the learner.
- Query: The system elicits a response from AI based solely on the information provided by the learner in the current session. The AI is intentionally isolated from its pre-existing knowledge base.
- Reflect: The quality of AI's response serves as a direct index of the quality of the learner's explanation. If AI's response is confused or inaccurate, it provides diagnostic feedback regarding shortcomings in the original teaching.
- Refine: The learner observes this reflection, identifies the gaps in their explanation, revises their approach, and attempts to teach the concept again.

TQI assesses the quality of the learner's explanation and dynamically adjusts AI's response mode to provide targeted scaffolding. This guidance control manifests in four distinct modes:

- M0 (Confused Restatement): A deliberately low-competence mirroring of the learner's explanation, designed to surface ambiguity and elicit re-explanation.
- M1 (Clarifying Probe): Targeted questions that push the learner to sharpen definitions and provide clearer examples.

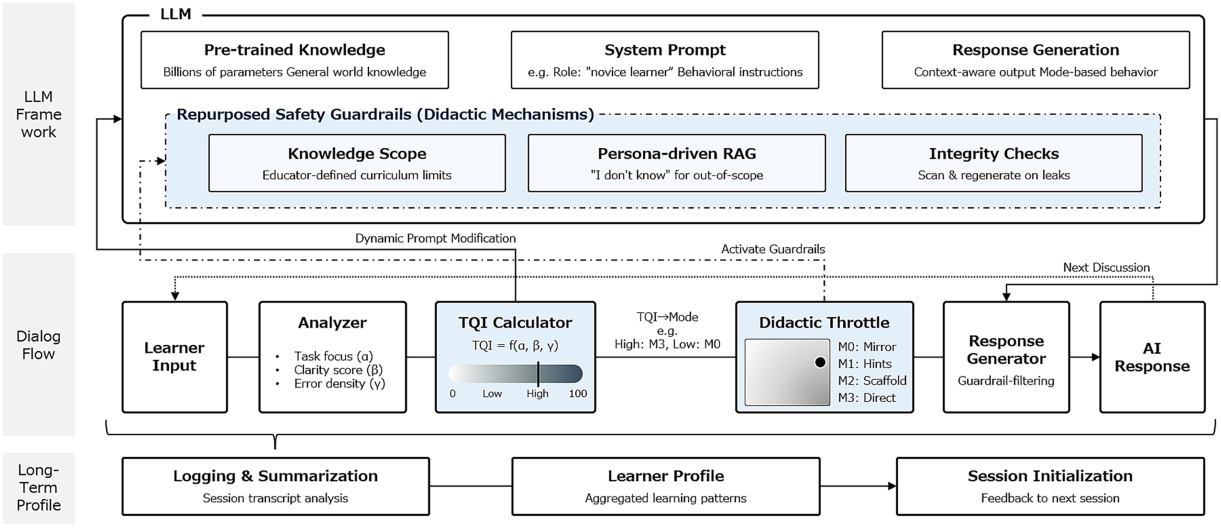

FIGURE 1
Overview of the cognitive mirror pipeline. Each learner's utterance is analyzed to extract task focus and evidence of understanding. The analysis yields a TQI that conditions a single instructional guidance level, selecting one of four response modes: M0 (Mirror), M1 (Hints), M2 (Scaffolded), and M3 (Direct). A scope guardrail placed before response generation prevents knowledge-scope deviation beyond the intended curriculum by reframing out-of-scope requests. Session logs and summaries update a privacy-preserving learner profile, which informs subsequent sessions.

- M2 (Socratic Gap): Prompts that point out missing logical links or unstated assumptions, inviting the learner to self-correct.
- M3 (Accurate Reformulation): A concise, correct paraphrase of the explanation to consolidate understanding once a high TQI is achieved.

Hence, by design, AI's errors become diagnostic. They create a private, low-stakes canvas on which the learner's misconceptions appear and can be repaired. The learner remains in charge of the reasoning workflow, while AI merely reflects and applies pressure, prompting the explanation to become clearer, precise, and complete. For the brief TQI operationalization, Supplementary materials 1.2 and 1.3.

2.2 Theoretical foundations: why the mirror works

The Cognitive Mirror paradigm is not merely a technical novelty; it is deeply rooted in established theories from learning and cognitive science. This section reviews three theoretical pillars – the Protégé Effect, Reflective Practice, and Metacognition –to motivate the pedagogical plausibility of this model.

2.2.1 The protégé effect revisited: the power of teaching

The Protégé Effect, or learning by teaching, is a powerful principle with demonstrated benefits. The Effect is most potent when the tutee is not omniscient, and visibly struggles, asks questions, and improves with guidance. Therefore, in this study's paradigm, AI is an adaptive student rather than a passive repository of knowledge, whereby the TQI modulates its response mode from M0 to M3. These modes keep the learner in charge of the reasoning workflow, while the AI's behavior mirrors and challenges the explanation given.

2.2.2 Schön's reflective practice: dialog as a learning model

Schön's model of Reflective Practice provides an excellent theoretical framework for understanding the dialog loop within this study's system (Schön, 1983, 1987). Schön identified two key processes—"reflection-in-action" and "reflection-on-action." The real-time dialog between the learner and the Cognitive Mirror is a direct application of reflection-in-action. When AI feigns confusion (M0) or asks a clarifying question (M1), the learner must "think on their feet," re-evaluating their explanation and adapting their teaching strategy. This immediate cycle of trial, error, and correction is critical for enhancing practical understanding. Conversely, reviewing the TQI dashboard after a session corresponds to reflection-on-action, where the learner can analyze their teaching performance using objective data and identify improvement strategies (Hsia et al., 2024; Farrell, 2022; Vysotskaya et al., 2020).

2.2.3 Fostering metacognitive monitoring and control

Both the Protégé Effect and Reflective Practice are intertwined with metacognition; that is, the capacity to think about one's thinking. The Cognitive Mirror is designed to encourage these metacognitive processes (Corwin et al., 2023), while AI quantitative feedback improves the accuracy of metacognitive monitoring (Hardianingsih and Haryanto, 2025). The act of explaining a concept to an external agent requires the learner to externalize their internal thought processes. This makes the process of monitoring one's understanding explicit and objective. The AI's feedback, especially the quantitative evaluation provided by the TQI, serves as critical external data to improve the accuracy of this metacognitive monitoring. Learners often overestimate their understanding level. Hence, objective AI feedback helps calibrate this self-assessment bias. The adaptive modes (M0-M3) serve as a checklist for metacognitive control, prompting the learner to repeatedly cycle through planning, monitoring, and

evaluating their teaching strategies in response to the AI's feedback. This is supported by the analysis of metacognitive interventions, which has demonstrated the effectiveness of metacognitive self-control cycles (Eberhart et al., 2025).

2.3 A critique of the omniscient AI through a theoretical lens

Integrating these frameworks clarifies why the conventional "AI-as-Tutor/Teacher" setup is insufficient for fostering deep learning.

- Attenuated Protégé Effect: When the system positions AI as an omniscient oracle, it diminished teaching accountability, weakening the Protégé Effect. Furthermore, a plain answer-providing LLM removes key desirable difficulties.
- Suppressed Reflective Practice: If the AI routinely supplies the correct solution path, learners lose the drive to try → observe → adjust, which is the core of reflection-in-action. Even "Socratic" exchanges can preserve oracle dynamics if AI implicitly "knows the answer."
- Inhibited Metacognitive Monitoring: Providing solutions before learners articulate their explanations hampers self-evaluation and self-correction, degrading calibration.

Hence, the Cognitive Mirror reconfigures the interaction to make AI an educational partner rather than a delivery channel – it induces responsibility, creates space for in-the-moment adaptation, and facilitates generate-then-judge cycles with objective feedback. Table 1 illustrates the comparison of AI paradigm frameworks in education.

TABLE 1 Comparison of AI paradigms in education across goals, roles, interaction mode, and metrics, emphasizing the shift from answer correctness to explanation quality indexed by the TQI.

Feature	Tutor paradigm (conventional AI tutor)	Cognitive mirror paradigm (our proposal)
Pedagogical Goal	Knowledge transfer, problem-solving	Knowledge construction, metacognitive development
Role of AI	Expert, teacher, information provider	Novice, tutee, reflective mirror
Role of Learner	Student, problem-solver, information consumer	Teacher, explainer, knowledge constructor
Primary Interaction	Q&A (Learner asks, AI answers)	Teaching Dialogue (Learner explains, AI probes)
Primary Metric	Correctness of the final answer	Quality of the explanation (TQI)
Theoretical Basis	Behaviorism, Information Processing Theory	Protégé Effect, Reflective Practice, Metacognition

2.4 Repurposing guardrails as instructional guidance controls

Rather than relying on conventional role-based prompting, this study repurposes what has been viewed merely as "safety guardrails" for generative AI (Lexman et al., 2025) into a didactic mechanism that deliberately cultivates productive ignorance. It reframes the system as a checker that faithfully reflects the quality of human instruction, opening new pathways for AI education coexistence. Guardrails can be reimagined as didactic mechanisms that sculpt AI's ignorance:

- Educator-defined curriculum scope: teachers upload the precise slice of curriculum to mirror, which becomes the AI's universe during the session. Approaches for aligning AI responses to an educator-defined curriculum have been proposed as part of effective guardrails in educational AI tools (Clark et al., 2025).
- Persona-driven retrieval-augmented generation: the prompt injects only in-scope materials and enforces a student persona: "if you are asked something you were not taught, say you do not know."
- Knowledge-integrity checks: responses are scanned for out-of-scope concepts; leaks prompt regeneration under stricter constraints.

Together, they prevent the model's pretraining knowledge from "helping." By extending this method into a dynamic and adaptive style, the system can provide greater educational value, allowing real-time modulation of AI's accessible knowledge scope according to evolving teaching objectives and learner needs. The AI using these guardrail settings cannot outshine the human; it can only reflect and challenge the explanation received. Hence, this is guardrails-for-scaffolding; instead of merely being guardrails-for-safety.

3 Holding up the mirror: an illustrative classroom demonstration

This activity served as anecdotal evidence of feasibility rather than a controlled evaluation; thus, we treat it as illustrative. The Cognitive Mirror paradigm was implemented as a working prototype developed by the authors. To illustrate its real-world feasibility and pedagogical potential, a public demonstration lesson was conducted on July 17, 2025, at Ritsumeikan Moriyama Junior and Senior High School. It was a classroom activity conducted as part of ordinary instruction rather than a formal research study. The examples of dialog were indicated in Supplementary materials 1.2.

The class involved 36 third-year students in an advanced course. The learning objective was the use of English relative adverbs (where, when, and why). For this session, the AI was configured without prior access to this grammar point, and a simple classroom rule was set: students were not allowed to use external AI tools and had to rely on their understanding. After brief whole-class activities, pairs engaged with the Cognitive Mirror application to "teach" the AI. Initial attempts produced a distorted reflection, given their incomplete explanations; the AI failed on quiz items. This immediate, model-mediated feedback prompted the students to refine their explanations, analyze errors, negotiate wording, add counterexamples, and clarify exceptions.

The observations during the lesson suggested a shift from answer-seeking to explanation-building. Informal student reflection during the lesson indicated that several students became aware of ambiguities in their understanding while trying to produce a "clear reflection." No personally identifiable student data were reported in this study, and all examples are de-identified and illustrative.

4 Discussion

The Cognitive Mirror paradigm aims to reframe the role of generative AI in education, shifting its purpose from an omniscient tutor to a metacognitive partner who reflects the quality of a learner's explanation. By re-centering human agency, the learner takes the role of an explainer, and the educator is a curriculum architect. This paradigm offers a path toward a symbiotic coexistence that protects the effortful, reflective work at the heart of genuine learning.

4.1 Educational and pedagogical implications

This paradigm carries significant implications for pedagogy and assessment. It redefines the educator's role from a "transmitter of knowledge" to a "facilitator of learning." As AI assumes fine-grained instruction in students' explanatory skills, a teacher role reallocation occurs: responsibilities shift so that educators focus on distinctly human work—cultivating motivation, kindling intellectual curiosity, and guiding deep, dialogical learning. The same mechanism points to scalable, authentic assessment. By evaluating the act of explaining in one's words, TQI can support high-quality, formative feedback at the cohort scale and may reduce reliance on proxy items. Since it focuses on the process of reconstruction rather than final answers, it may better accommodate diverse ways of knowing and may help move assessment toward greater equity, provided it remains low-stakes and instructional in intent.

4.2 Challenges and ethical considerations

Despite its potential, the implementation of the Cognitive Mirror must be approached with careful consideration of its inherent challenges. A central technical risk lies in tuning the guidance level to feel natural without leaking the AI's pre-existing knowledge. Additionally, there is a pedagogical risk of cheating, where learners focus on maximizing their TQI score rather than achieving genuine understanding, which necessitates careful design to reward semantic depth. Our classroom illustration lacks randomization, controls, and systematic measurement; therefore, it does not support causal claims. We outline an evaluation plan for future work, including (i) controlled classroom studies with pre/post measures of explanation quality and learning outcomes; (ii) calibration of TQI against expert ratings and assessment of inter-rater reliability; and (iii) bias and robustness checks across topics and discourse styles. The most complex ethical challenge is ensuring algorithmic fairness. AI models amplify biases from their training data if the TQI is trained to recognize a "standard" explanation style. It risks unfairly penalizing learners from backgrounds that favor different rhetorical styles, such as narrative or holistic approaches. Hence, the paradigm is designed to promote autonomy by withholding information to elicit cognitive effort from the learner and make scaffolding unnecessary.

4.3 A call for a new research agenda

The concepts presented in this study are a starting point for a new research domain. It proposes a research roadmap to explore and develop the Cognitive Mirror paradigm:

1. Validation and refinement: Rigorously test TQI and throttle behaviors across tasks, domains, languages, and model versions, and compare learning gains against feedback-as-answers baselines with controlled studies.
2. Longitudinal and adaptive models: Develop a "Forgetting Mirror" that uses time-decayed profiles to prompt re-teaching at optimal intervals, supporting durable consolidation and mastery tracking.
3. Ethical personalization and fairness: Build frameworks for cultural awareness evaluation and feedback styles without stereotyping and combine dataset diversification with ongoing bias diagnostics and user-controllable personas.

In conclusion, the Cognitive Mirror is a platform for inquiry; that is, a way to branch model power from answer provision to explanation-centric learning. This study invites communities in AI, learning sciences, and ethics to examine and challenge the findings. This marks the beginning of an intellectual journey to cultivate AI as a partner in augmenting learning capacity.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

HT: Visualization, Project administration, Investigation, Writing – review & editing, Conceptualization, Writing – original draft. JU: Investigation, Resources, Software, Writing – review & editing. TY: Supervision, Writing – review & editing, Conceptualization, Funding acquisition.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. The research reported in this publication was supported by JST RISTEX Japan Grant Number JPMJRS24K3, the Sasakawa Scientific Research Grant from the Japan Science Society, the Institute of Social

Systems at Ritsumeikan University, and the Ritsumeikan Global Innovation Research Organization.

Acknowledgments

The authors thank Takumi Ueda for their comments on a draft of this article. Yuki Nakai developed the testing tool and Yuma Yamauchi cooperated in pilot testing this tool.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that Gen AI was used in the creation of this manuscript. The initial draft of this article was written in Japanese. To translate portions of the manuscript into English, the author used OpenAI's ChatGPT, Anthropic's Claude and Google's Gemini. After using these services, the author reviewed and edited the content as needed and took full responsibility for the content of the published article.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2025.1697554/full#supplementary-material

References

Adewumi, T., Liwicki, F. S., Liwicki, M., Gardelli, V., Alkhaled, L., and Mokayed, H. (2025). Findings of mega: maths explanation with LLMs using the Socratic method for active learning. *arXiv*. doi: 10.48550/arXiv.2507.12079

Allen, A. (1967). Children as teachers. *Child. Educ.* 43, 345–350. doi: 10.1080/00094056.1967.10728069

Almuhanna, M. A. (2024). Teachers' perspectives of integrating AI-powered technologies in K-12 education for creating customized learning materials and resources. *Educ. Inf. Technol.* 30, 10343–10371. doi: 10.1007/s10639-024-13257-y

Anthropic. (2024). Many-shot jailbreaking. Available online at: https://www.anthropic.com/research/many-shot-jailbreaking (Accessed August 8, 2025).

Chase, C. C., Chin, D. B., Oppezzo, M. A., and Schwartz, D. L. (2009). Teachable agents and the protégé effect: increasing the effort towards learning. *J. Sci. Educ. Technol.* 18, 334–352. doi: 10.1007/s10956-009-9180-4

Chen, B. (2025). Beyond tools: generative AI as epistemic infrastructure in education. *arXiv*. doi: 10.48550/arXiv.2504.06928

Chen, L., Chen, P., and Lin, Z. (2020). Artificial intelligence in education: a review. *IEEE Access* 8, 75264–75278. doi: 10.1109/ACCESS.2020.2988510

Chen, A., Wei, Y., Le, H., and Zhang, Y. (2025). Learning by teaching with ChatGPT: the effect of teachable ChatGPT agent on programming education. *Br. J. Educ. Technol.* doi: 10.1111/bjet.70001

Chhibber, N., and Law, E. (2019). Using conversational agents to support learning by teaching. *arXiv*. doi: 10.48550/arXiv.1909.13443

Chiu, T. K. F., Moorhouse, B. L., Chai, C. S., and Ismailov, M. (2024). Teacher support and student motivation to learn with artificial intelligence (AI) based chatbot. *Interact. Learn. Environ.* 32, 3240–3256. doi: 10.1080/10494820.2023.2172044

Choi, J., Hong, Y., Kim, M., and Kim, B. (2024). Examining identity drift in conversations of LLM agents. *arXiv*. doi: 10.48550/arXiv.2412.00804

Clark, H. B., Benton, L., Searle, E., Dowland, M., Gregory, M., Gayne, W., et al. (2025). Building effective safety guardrails in AI education tools. In International Conference on Artificial Intelligence in Education, 129–136.

Corwin, T., Kosa, M., Nasri, M., Holmgård, C., and Harteveld, C. (2023). The teaching efficacy of the protégé effect in gamified education. In 2023 IEEE Conference on Games, 1–8.

Dror, I. E., and Harnad, S. (2008). "Offloading cognition onto cognitive technology" in Cognition distributed: how cognitive technology extends our minds. eds. I. E. Dror and S. Harnad (Amsterdam: John Benjamins Publishing Company), 1–23.

Eberhart, J., Ingendahl, F., and Bryce, D. (2025). Are metacognition interventions in young children effective? Evidence from a series of meta-analyses. *Metacogn. Learn.* 20:7. doi: 10.1007/s11409-024-09405-x

Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., et al. (2024). Beware of metacognitive laziness: effects of generative artificial intelligence on learning motivation, processes, and performance. *Br. J. Educ. Technol.* 56, 489–530. doi: 10.1111/bjet.13544

Farrell, T. S. C. (2022). Reflective practice in language teaching. Cambridge: Cambridge University Press.

Fathi, J., Rahimi, M., and Teo, T. (2025). Applying intelligent personal assistants to develop fluency and comprehensibility and reduce accentedness in EFL learners: an empirical study of Google assistant. *Lang. Teach. Res.*:13621688251317786. doi: 10.1177/13621688251317786

Favero, L., Pérez-Ortiz, J. A., Käser, T., and Oliver, N. (2024). Enhancing critical thinking in education by means of a Socratic chatbot. In International workshop on AI in education and educational research, 17–32.

Fiorella, L., and Mayer, R. E. (2013). The relative benefits of learning by teaching and teaching expectancy. *Contemp. Educ. Psychol.* 38, 281–288. doi: 10.1016/j.cedpsych.2013.06.001

Fiorella, L., and Mayer, R. E. (2014). Role of expectations and explanations in learning by teaching. *Contemp. Educ. Psychol.* 39, 75–85. doi: 10.1016/j.cedpsych.2014.01.001

Gerlich, M. (2025). AI tools in society: impacts on cognitive offloading and the future of critical thinking. *Societies* 15:6. doi: 10.3390/soc15010006

Google. (2025). Guided learning in Gemini: from answers to understanding. Available online at: https://blog.google/outreach-initiatives/education/guided-learning/ (Accessed August 15, 2025).

Hardianingsih, R., and Haryanto, H. (2025). The effect of learning motivation and roboguru integration on metacognition in independent learning. *Int. J. Community Engagement Payungi* 5, 174–181. doi: 10.58879/ijcep.v5i1.86

Hsia, L. H., Hwang, G. J., and Hwang, J. P. (2024). Ai-facilitated reflective practice in physical education: an auto-assessment and feedback approach. *Interact. Learn. Environ.* 32, 5267–5286. doi: 10.1080/10494820.2023.2212712

Jose, B., Cherian, J., Verghis, A. M., Varghise, S. M., S, M., and Joseph, S. (2025). The cognitive paradox of AI in education: between enhancement and erosion. *Front. Psychol.* 16:1550621. doi: 10.3389/fpsyg.2025.1550621

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X. H., Beresnitzky, A. V., et al. (2025). Your brain on ChatGPT: accumulation of cognitive debt when using an AI assistant for essay writing task. *arXiv*. doi: 10.48550/arXiv.2506.08872

Lee, H. P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., et al. (2025). The impact of generative AI on critical thinking: self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In Proceedings of the 2025 CHI conference on human factors in computing systems. 1–22.

Lexman, R. R., Krishna, A., and Sam, M. P. (2025). AI guardrails in business and education: bridging minds and markets. *Dev. Learn. Organ.* doi: 10.1108/DLO-01-2025-0001

Li, Y. (2024). Usability of ChatGPT in second language acquisition: capabilities, effectiveness, applications, challenges, and solutions. *Stud. Appl. Linguist. TESOL* 24, 24–37. doi: 10.52214/salt.v24i1.12864

Love, R., Law, E., Cohen, P. R., and Kulić, D. (2022). Natural language communication with a teachable agent. *arXiv*. doi: 10.48550/arXiv.2203.09016

Maharana, A., Lee, D. H., Tulyakov, S., Bansal, M., Barbieri, F., and Fang, Y. (2024). Evaluating very long-term conversational memory of LLM agents. *arXiv*. doi: 10.48550/arXiv.2402.17753

Martin, J.-P., and Oebel, G. (2007). "Lernen durch Lehren: Paradigmenwechsel in der Didaktik?" in Deutschunterricht in Japan, vol. 12, 4–21.

Okita, S. Y., and Schwartz, D. L. (2013a). Learning by teaching human pupils and teachable agents: the importance of recursive feedback. *J. Learn. Sci.* 22, 375–412. doi: 10.1080/10508406.2013.807263

Okita, S. Y., Turkay, S., Kim, M., and Murai, Y. (2013b). Learning by teaching with virtual peers and the effects of technological design choices on learning. *Comput. Educ.* 63, 176–196. doi: 10.1016/j.compedu.2012.12.005

OpenAI. (2025). Introducing study mode. Available online at: https://openai.com/index/chatgpt-study-mode/ (Accessed August 15, 2025).

Rittle-Johnson, B., Saylor, M., and Swygert, K. E. (2008). Learning from explaining: does it matter if mom is listening? *J. Exp. Child Psychol.* 100, 215–224. doi: 10.1016/j.jecp.2007.10.002

Schmidt, T., and Strasser, T. (2022). Artificial intelligence in foreign language learning and teaching: a CALL for intelligent practice. *Anglistik* 33, 165–184. doi: 10.33675/ANGL/2022/1/142022

Schön, D. A. (1983). The reflective practitioner: how professionals think in action. New York: Basic Books.

Schön, D. A. (1987). Educating the reflective practitioner: toward a new design for teaching and learning in the professions. San Francisco: Jossey-Bass.

Tomisu, H., Ueda, T., Ueda, J., and Yamanaka, T. (2025). Nurturing AI language companions: a role-reversal approach to language learning. Proceedings of 2025 9th International Conference on Artificial Intelligence and Virtual Reality, in press.

Tsai, C. C., Tsai, J. C., Lin, H. M., Li, Y. Q., and Tseng, S. P. (2025). Applying a large language model to second language acquisition. *Sensors Mater.* 37, 2743–2755. doi: 10.18494/SAM5172

Vysotskaya, P., Zabelina, S., Kuleshova, J., and Pinchuk, I. (2020). Using the capabilities of artificial intelligence in the development of reflection skills. *E3S Web Conf.* 210:22035. doi: 10.1051/e3sconf/202021022035

Wei, L. (2023). Artificial intelligence in language instruction: impact on English learning achievement, L2 motivation, and self-regulated learning. *Front. Psychol.* 14:1261955. doi: 10.3389/fpsyg.2023.1261955

Xie, X., Yang, X., and Cui, R. (2025). A large language model-based system for Socratic inquiry: fostering deep learning and memory consolidation. In 2025 14th International Conference on Educational and Information Technology, 52–57.

Yong, C. L., Furqan, M. S., Lee, J. W. K., Makmur, A., Mariappan, R., Ngoh, C. L. Y., et al. (2024). The use of large language models tuned with Socratic methods on the impact of medical students' learning: a randomised controlled trial. JMIR. Available online at: https://preprints.jmir.org/preprint/57995 (Accessed August 2, 2025).

Zhai, X. (2022). ChatGPT user experience: implications for education. *SSRN*. doi: 10.2139/ssrn.4312418

Zhang, Z., and Huang, X. (2024). The impact of chatbots based on large language models on second language vocabulary acquisition. *Heliyon* 10:e25370. doi: 10.1016/j.heliyon.2024.e25370