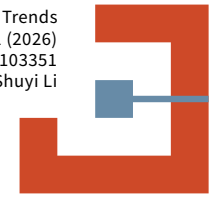
Trends

JALTCALL Trends
ISSN 2982-3749, 2(1), 103351 (2026)
https://doi.org/10.29140/jct.v2n1.103351



Castledown

Building rhetorical scaffolding: Closing the envisioning gulf in LLM-generated feedback for argumentative writing

Shuyi Li
The University of Tokyo, Japan
lishuyi213@g.ecc.u-tokyo.ac.jp

Large Language Models (LLMs) are increasingly used as writing assistants, yet their feedback often defaults to superficial, sentence-level corrections that fail to develop students' rhetorical skills. This paper addresses the cognitive challenges writers face when using LLMs, conceptualized as the "gulf of envisioning" – the gap between a writer's intentions and their ability to formulate effective prompts. It is argued that this gulf is particularly wide for complex, under-specified tasks like argumentative writing. Drawing on a qualitative synthesis of research in Human Computer Interaction (HCI) and Computer-Assisted Language Learning (CALL), "rhetorical scaffolding" is proposed as a framework of structured, dialogic prompting strategies for pedagogical intervention. This paper demonstrates how these strategies can bridge the envisioning gulf by transforming the AI-writer interaction from a corrective monologue into a metacognitive partnership.

Keywords: rhetorical situation, gulf of envisioning, Large Language Models (LLMs), argumentative writing

Introduction

The rapid integration of Large Language Models (LLMs) into educational settings highlights their potential as valuable tools for writing instruction. While their ability to provide timely, scalable feedback addresses challenges such as large class sizes and limited instructor time, their effectiveness is largely perceived as strongest in improving surface-level writing aspects like grammar, vocabulary, and mechanics. Despite the positive reception of AI-generated feedback, there is a clear preference among both students and educators for the nuanced, expert guidance of human teachers or a hybrid approach that combines human and artificial intelligence (Gozali et al., 2024; Wang & Gayed, 2024). When LLMs are used solely as error-correction tools, the revision process risks becoming a superficial exercise in accepting

automated suggestions rather than fostering a deeper, reflective engagement with ideas and argumentation. The central challenge for writing pedagogy is to redefine the student-LLM interaction from a transactional exchange focused on correction to a formative dialogue that enhances, rather than replaces, the writer's critical agency. This is particularly vital for L2 writers, who may feel linguistically vulnerable and thus more likely to uncritically accept AI-generated text, bypassing the "desirable difficulty" of rhetorical formulation (Mah et al., 2025; Solovey, 2024).

The cognitive challenge: Bridging the "Gulf of Envisioning"

From a technology design perspective, interacting with LLMs poses a significant cognitive challenge, as users struggle to translate abstract intentions into effective prompts (i.e., text input from the user to guide the model). This creates a critical disconnect for learners in LLM-assisted writing, particularly during the revision process, when they must evaluate text they did not author themselves. To examine this pedagogical mismatch, this paper applies the theoretical framework of the "gulf of envisioning" (Subramonyam et al., 2024), which refers to the cognitive distance between a writer's goals and their ability to formulate prompts that elicit the desired output from an LLM. This gulf consists of three distinct challenges: the capability gap (uncertainty about the LLM's capabilities), the instruction gap (difficulty articulating intentions into effective prompts), and the intentionality gap (lack of clear criteria for evaluating the LLM's output).

Research purpose and guiding framework

The envisioning gulf is most evident in LLM-assisted writing, where the prompt is not a well-defined task but an underspecified, real-world challenge requiring complex modeling, assumption-making, and rhetorical reasoning. Each component of this gulf corresponds to a specific failure point in the AI-writer dialogue for argumentation.

1. The capability gap: Unsure of an LLM's ability to critique abstract concepts like argumentation, students default to simple requests like grammar checks. This limits feedback to the surface level, leaving deeper flaws unaddressed.
2. The instruction gap: Students spot higher-order issues but lack the vocabulary to prompt for targeted revisions. Vague commands like "make this better" yield generic LLM responses, preventing a sophisticated dialogue.
3. The intentionality gap: Students accept flawed but polished LLM output, missing opportunities to learn critical self-evaluation and improvement.

To address these intertwined challenges, this paper advances the AI-integrated Self-Directed Writing (SDW) framework, adapted from research on self-directed learning and rhetorical genre studies (Bitzer 1968; Ranade et al., 2024; Wang et al., 2024), to guide writers in constructing rhetorical scaffolding while interacting with language models during the revision process. The framework reconceptualizes the writer not as a passive recipient of automated feedback, but as an active, self-directed agent who strategically integrates the LLM alongside other tools and supports within

a broader learning ecology. As illustrated in Figure 1, this ecology relies on the dynamic interplay of the rhetorical situation (e.g., writer, audience, subject, and purpose). The figure maps how these core components must be explicitly aligned within the situational context to bridge the envisioning gulf. Within this perspective, the study is guided by the following question:

- How can rhetorical scaffolding in LLM prompts bridge the envisioning gulf and enhance the pedagogical value of LLM-generated feedback in argumentative writing?

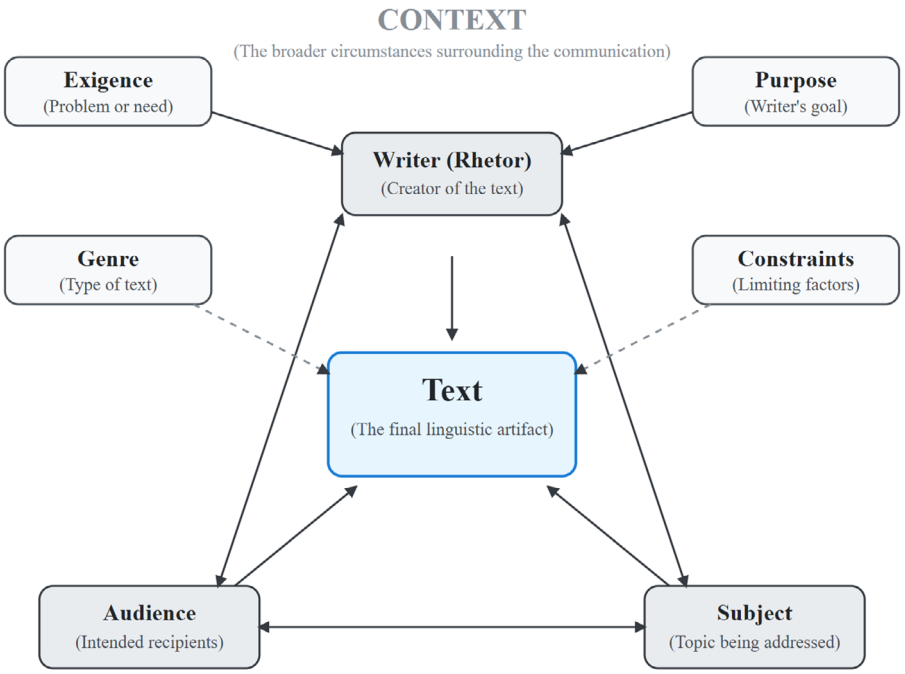


Figure 1. The rhetorical situation framework: Core components and their relationship to situational context

Methods: A/B testing and comparative analysis

The initial phase involved the systematic collection of a corpus of over 200 ($N = 215$) publicly available prompts designed for academic writing tasks. The collection was intentionally focused on prompts intended for the revision stage (e.g., strengthening a thesis statement, improving paragraph cohesion, refining academic tone, checking for logical fallacies). These prompts were sourced from diverse online platforms, including prompt libraries, university repositories, and communities of practice where educators and learners share resources.

Following the data collection, an A/B testing design was employed in which sample texts ($N = 15$) were submitted to three contemporary LLMs (Claude 3.5 Sonnet, Gemini 2.5 Pro, and ChatGPT-4o) under two prompt conditions: generic prompts (e.g., "Improve this text") and rhetorically-informed prompts, which intentionally embedded components of the contextualized rhetorical situation. The resulting LLM revisions were rated on a scale from 1 to 5 (in 0.5-point increments) by two expert raters (applied linguists with 5+ years of experience). Intra-rater reliability was calculated

using Cohen's Kappa, yielding a strong agreement coefficient of 0.82. Discrepancies were resolved through discussion to reach a consensus score. This rubric (see Table 1) was used to evaluate the impact of rhetorical prompting on feedback quality:

Table 1. Adapted CAF framework for evaluating LLM-generated feedback

Dimension	Description	Example of high scoring feedback
Complexity	Assesses the **depth** and **intricacy** of the feedback, including the variety of strategies suggested.	Feedback that offers multiple approaches to refining a claim.
Accuracy	Evaluates the **correctness** and **appropriateness** of the feedback provided.	Feedback that accurately identifies and corrects grammatical errors.
Fluency	Measures the **clarity** and **coherence** of the feedback.	Feedback that is clearly phrased and logically structured.
Relevance	Assesses how well the feedback addresses the **specific needs** and **context** of the text.	Feedback that is tailored to the rhetorical goals of the writer.
Engagement	Evaluates how effectively the feedback **motivates** and **involves** the writer in the revision process.	Feedback that encourages the writer to explore different rhetorical strategies.

Results: The impact of rhetorical grounding on feedback quality

As illustrated in Figure 2, the vast majority of prompts (83.3%) were categorized as baseline prompts, containing no explicit rhetorical guidance, whereas only 16.7% were identified as rhetorically-informed prompts.

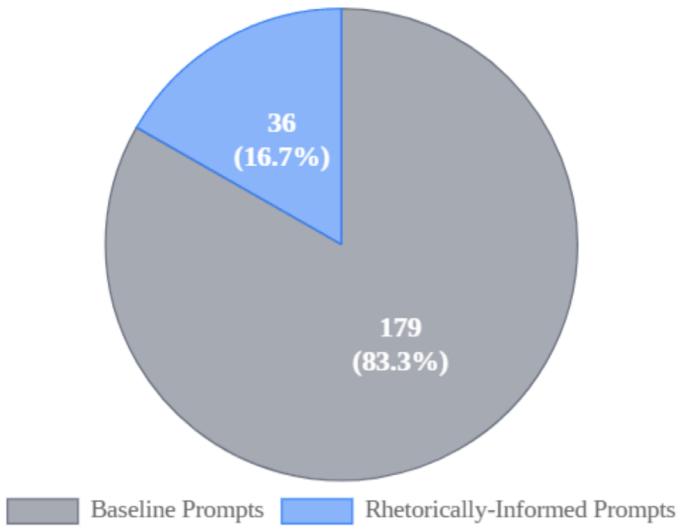


Figure 2. Distribution of prompt types

Furthermore, the comparative analysis of the sample data revealed clear differences between the two prompt types, particularly in their functions and prompting methods (see Table 2). Baseline prompts predominantly addressed lower-order concerns, with 88.3% ($n = 158$) focusing on grammar/syntax and 7.3% ($n = 13$) on summarization.

These prompts overwhelmingly relied on the Zero-shot methodology, where the model is given a task without any examples, which accounted for 91.5% (n = 165) of all baseline prompts, a distribution detailed in Figure 3. Conversely, rhetorically-informed prompts primarily addressed higher-order concerns, most often structure/organization (36.1%, n = 13) and content/ideation (33.3%, n = 12). As illustrated in Figure 4, they employed more complex methods, with Few-shot (providing examples of the desired output) and Chain-of-Thought (CoT) (explicitly requesting step-by-step reasoning) prompting used in 16 prompts each, together accounting for 89% of the set.

Table 2. Categorization of prompt types by function and prompting methodology

Prompt type	Methodology	Grammar/ syntax (LOC)	Summarization/ paraphrasing (LOC)	Content/ ideation (HOC)	Structure/ organization (HOC)	Subtotal by methodology
Baseline	**Total**	**158 (88.3%)**	**13 (7.3%)**	**5 (2.8%)**	**3 (1.7%)**	**179**
	Zero-shot	151 (91.5%)	9 (5.5%)	3 (1.8%)	2 (1.2%)	**165**
	Few-shot	7 (50.0%)	4 (28.6%)	2 (14.3%)	1 (7.1%)	**14**
	Chain-of-thought	0 (0%)	0 (0%)	0 (0%)	0 (0%)	**0**
Rhetorically-informed	**Total**	**5 (13.9%)**	**6 (16.7%)**	**12 (33.3%)**	**13 (36.1%)**	**36**
	Zero-shot	4 (100%)	0 (0%)	0 (0%)	0 (0%)	**4**
	Few-shot	1 (6.3%)	5 (31.3%)	6 (37.5%)	4 (25.0%)	**16**
	Chain-of-thought	0 (0%)	1 (6.3%)	6 (37.5%)	9 (56.3%)	**16**
Grand total		**163 (75.8%)**	**19 (8.8%)**	**17 (7.9%)**	**16 (7.4%)**	**215**

Note. HOC refers to Higher-Order Concerns (global issues of argumentation and structure), while LOC refers to Lower-Order Concerns (surface-, sentence-level issues that relate to correctness and polish).

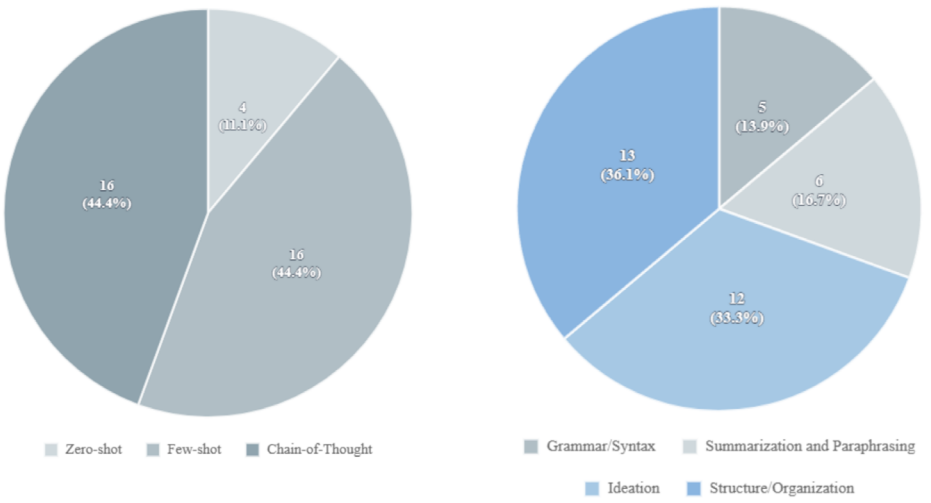


Figure 3. Baseline prompts by methodology & functional purpose

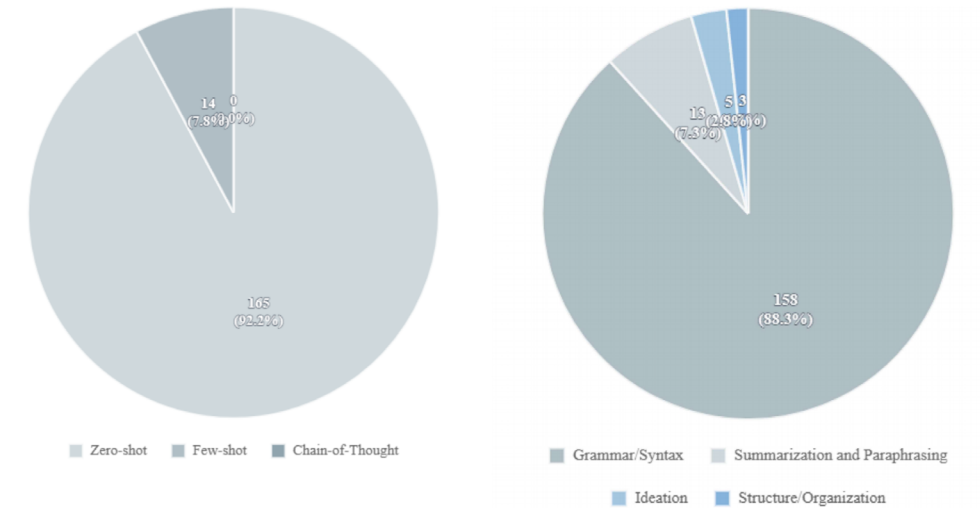


Figure 4. Rhetorically-informed prompts by methodology & functional purpose

The data in Figure 5 demonstrates that scaffolding prompts significantly improved feedback quality compared to the baseline condition. A paired-samples *t*-test confirmed that this difference was statistically significant ($t = 4.82$, $p < .001$). They were particularly effective in two areas: increasing the rhetorical relevance of the feedback (+3.4) and enhancing its pedagogical engagement (+2.9). Generic prompts consistently resulted in superficial, low-context feedback, which was limited to surface-level changes, such as substituting words with synonyms, correcting minor grammatical errors, or rephrasing sentences in ways that did not improve clarity or argumentative strength. In contrast, rhetorically-informed prompts elicited feedback that was consistently more substantive, strategic, and pedagogically useful. Guided by the specified context, the mode provided suggestions that addressed higher-order concerns of argumentation, structure, and clarity. It often explained the reasoning behind its suggestions, thereby offering a teachable moment rather than just a quick fix.

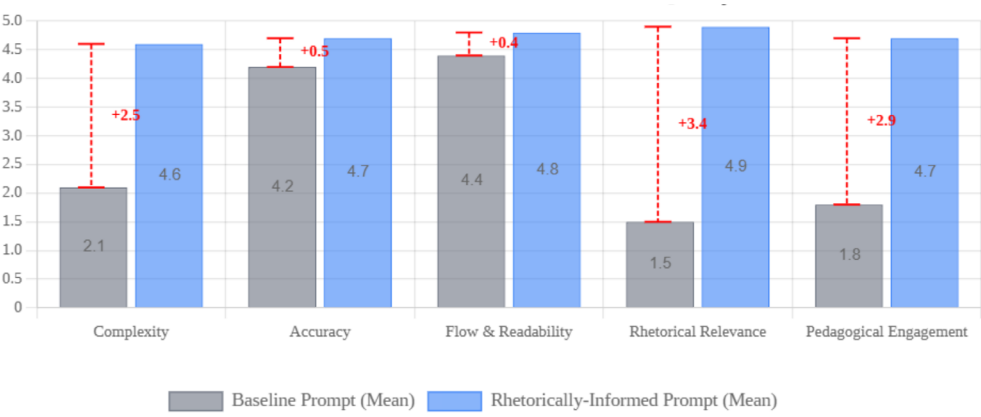


Figure 5. Mean scores of LLM-generated feedback quality

Discussion: Bridging the gulf with rhetorical scaffolding

As shown in Table 1, the results highlight a clear association between the rhetorical construction of a prompt, the use of advanced prompting methodologies, and the generation of higher-order feedback. These strategies focus on two primary dimensions of the gulf: instruction (clarifying the writer's communicative aim) and intentionality (emphasizing the writer's agency in shaping their argument). In some cases, the strategies also extend to capability, enhancing the range of rhetorical moves available to the writer.

Table 3. Mapping rhetorical prompt components to the gulf of envisioning

Gulf component	Description (the user's problem)	Rhetorical prompt component	Illustrative prompt snippet
Capability gap	"What can the LLM do for my specific writing task?"	Genre, Context, Subject	"Review the following **argumentative essay** on the topic of climate change policy. The **context** is a submission for an undergraduate political science journal."
Instruction gap	"How do I ask the LLM to give me the feedback I need?"	Audience, Role, Purpose	"Act as a **peer reviewer**. My **purpose** is to strengthen my main claim. Focus your feedback on the connection between my evidence and my thesis for an **audience** of fellow students."
Intentionality gap	"What should I expect as a good response? How do I judge the output?"	Constraints, exigence + iterative loop	"My **main problem** (exigence) is that my argument feels weak. **Constraint**: Do not rewrite my sentences. First, identify any logical fallacies in my reasoning. Then, explain why they are fallacies. Finally, suggest a way I could revise to fix them."

However, the goal is not to craft the perfect prompt to secure a polished product from the LLM, but to use dialogue as a tool for developing the writer's critical thinking and rhetorical awareness. Structured scaffolding transforms feedback from a one-way correction into an exchange where the LLM acts as a dialogue partner, modeling advanced argumentation strategies. This is particularly valuable in CALL contexts, where efficiency often comes at the expense of depth, and where L2 writers benefit from guided engagement that integrates format, content, and audience awareness (Zhang & Wang, 2025).

Building on this foundation, five rhetorical scaffolding strategies simulate diverse rhetorical perspectives to provide targeted, strategic support for complex writing tasks. *Claim Refinement* positions the writer in a rhetorical decision space, prompting simultaneous attention to stance, tone, and audience adaptation. *Evidence Interrogation* applies a Socratic approach to examine the warrants supporting an argument, an aspect where LLM feedback often remains superficial. To illustrate this, Excerpt 1 demonstrates how rhetorical scaffolding transforms the interaction. *Counterargument Simulation* introduces dialogic tension, encouraging writers to engage with oppositional reasoning and strengthen argumentative resilience.

Audience Adaptation surfaces and reframes implicit audience assumptions, making reader considerations both explicit and measurable. Finally, *Structural Coherence Check* shifts attention from sentence-level refinement to global organization, addressing a persistent shortcoming of LLM feedback without explicit prompting.

Excerpt 1. Rhetorically Scaffolded Interaction (Evidence Interrogation)

> **Prompt:** "Act as a skeptic of remote work. I am arguing that working from home increases productivity. Review my second paragraph and identify a logical gap in my cause-and-effect reasoning."
>
> **LLM Output:** "You claim that 'fewer office interruptions' leads directly to 'more work done.' However, you ignore the potential for new interruptions at home (e.g., childcare, chores). You need to qualify your claim: does it increase productivity for all workers, or only those with dedicated home offices?" (Critique: Identifies a specific logical gap and prompts the writer to qualify their claim.)

Taken together, these strategies encourage a shift from reliance on LLM-generated corrections toward a collaborative, pedagogy-driven partnership. Anchored in Self-Directed Learning theory, the scaffolding framework uses a sequenced set of prompts to promote both rhetorical and metacognitive engagement. Writers are prompted to define their communicative goals and evaluate their own work before receiving feedback. This transforms the interaction into an active exchange about writing strategy, with prompts functioning as cognitive tools to clarify intent. Ultimately, this approach reflects the Human-in-the-Loop (HITL) principle, strengthening writer agency and intentionally integrating technology to support learning objectives in LLM-based writing (see Figure 6).

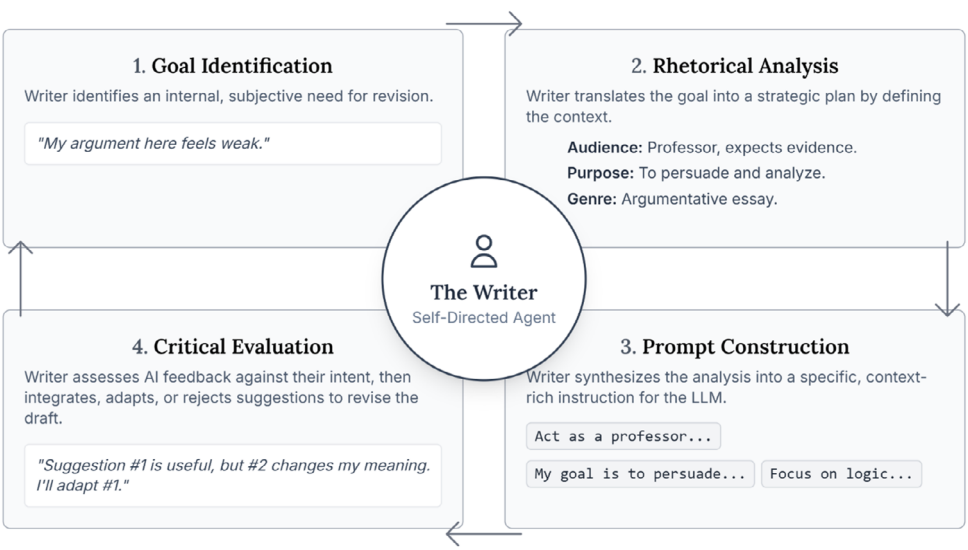


Figure 6. Self-reflection and iteration loop in LLM-based writing

Conclusion

This study proposes rhetorical scaffolding as a practical, theoretically grounded solution. By structuring LLM-writer dialogue around key rhetorical principles, it addresses the capability, instruction, and intentionality gaps, shifting the LLM's role from passive corrector to active collaborator. In doing so, it enables writers to engage LLMs not as shortcuts for text production but as tools for deepening rhetorical thinking.

JALTCALL Trends
Vol. 2 No. 1
2025 JALTCALL Conference Papers Special Issue

References

Bitzer, L. F. (1968). The rhetorical situation. *Philosophy & Rhetoric, 1*(1), 1–14. https://www.jstor.org/stable/40236733

Gozali, I., Wijaya, A. R. T., Lie, A., Cahyono, B. Y., & Suryati, N. (2024). Leveraging the potential of ChatGPT as an automated writing evaluation (AWE) tool: Students' feedback literacy development and AWE tools integration framework. *The JALT CALL Journal, 20*(1), 1–22. https://doi.org/10.29140/jaltcall.v20n1.1200

Mah, C., Tan, M., Phalen, L., Sparks, A., & Demszky, D. (2025). *From sentence-corrections to deeper dialogue: Qualitative insights from LLM and teacher feedback on student writing* (EdWorkingPaper No. 25-1193). Annenberg Institute at Brown University. https://doi.org/10.26300/p397-2p46

Ranade, N., Saravia, M., & Johri, A. (2024). Using rhetorical strategies to design prompts: A human-in-the-loop approach to make AI useful. *AI & Society, 40,* 711–732. https://doi.org/10.1007/s00146-024-01905-3

Solovey, O. Z. (2024). Comparing peer, ChatGPT, and teacher corrective feedback in EFL writing: Students' perceptions and preferences. *Technology in Language Teaching & Learning, 6*(3), 1–23. https://doi.org/10.29140/tltl.v6n3.1482

Subramonyam, H., Pea, R., Pondoc, C. L., Agrawala, M., & Seifert, C. (2024). Bridging the gulf of envisioning: Cognitive challenges in prompt based interactions with LLMs. *In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24).* ACM. https://doi.org/10.1145/3613904.3642754

Wang, K. D., Burkholder, E., Wieman, C., Salehi, S., & Haber, N. (2024). Examining the potential and pitfalls of ChatGPT in science and engineering problem-solving. *Frontiers in Education, 8,* Article 1330486. https://doi.org/10.3389/feduc.2023.1330486

Wang, Q., & Gayed, J. M. (2024). Effectiveness of large language models in automated evaluation of argumentative essays: Finetuning vs. zero-shot prompting. *Computer Assisted Language Learning, 39*(1–2), 209–237. https://doi.org/10.1080/09588221.2024.2371395

Zhang, J., & Wang, J. (2025). Student perceptions of hybrid feedback: Using gen-AI to enhance engagement with EAP writing feedback. *The JALT CALL Journal, 21*(2), Article 2175. https://doi.org/10.29140/jaltcall.v21n2.2175

Author biodata

Shuyi Li is a graduate student at the University of Tokyo researching the intersection of large language models (LLMs) and language education. Her current work centers

on LLM-based writing, with the goal of enhancing learner engagement, supporting diverse needs, and improving educational outcomes.

JALTCALL
Trends
Vol. 2 No. 1
2025 JALTCALL
Conference Papers
Special Issue