Language Ecology and Application
WISDOM ACADEMIC https://doi.org/10.63808/lea.v2i1.485 Volume 2 , Issue 1

Research Article

Prompting Depth as a Behavioural Proxy for Cognitive Engagement: A Learning Analytics Study of Student–Chatbot Dialogues in EFL Writing

Sihao Zhang[1,*]
[1]Educational Technology, Faculty of Educational Sciences and Technology, Universiti Teknologi Malaysia, Johor Bahru 81310, Malaysia
[*]**Correspondence:** Sihao Zhang, sihaoz223@gmail.com

© The Author(s) 2026.
This article is published by Wisdom Academic Press Ltd. under a Creative Commons Attribution 4.0 International License. The license permits use, sharing, adaptation, distribution, and reproduction in any medium or format, provided that appropriate credit is given to the original author(s) and the source, a link to the license is supplied, and any modifications are indicated. Unless otherwise stated, all images and third-party materials included in this article are also covered by the same license. For any material not covered by this license, if the intended use is neither permitted by applicable law nor allowed under the license terms, direct permission must be obtained from the copyright holder. To view a copy of this license, please visit http://creativecommons.org/licenses/by/4.0/.

Abstract

Cognitive engagement shapes learning but resists direct observation, and the spread of generative-AI chatbots in EFL writing has made post hoc self-reports even harder to trust. This study asks whether the depth of a learner's prompts (their specificity, elaboration, metacognitive framing, higher-order quality, and iterative refinement) can serve as a behavioural trace of that engagement. Using the RECIPE4U dataset (1,913 prompts from 130 EFL learners), two coders scored three prompt-level dimensions with substantial to near-perfect agreement (κ = .74–.84); the remaining dimensions were derived from intent annotations and session structure. The five dimensions did not behave as one construct. At the student level, only elaboration correlated with essay revision (r = .34) and, weakly, with satisfaction (r = .20); metacognitive framing and higher-order request added little, and iterative refinement related negatively to revision (r = −.18). An equal-weight composite showed no association with either criterion (r = .07). Each one-point increase in prompting depth was associated with more than twice the odds of subsequent revision (OR = 2.37; course-adjusted OR = 2.09), an association that persisted in a random-intercept model despite marked between-student variance (ICC = .41). Prompting depth is best read dimension by dimension: elaboration, not depth in the aggregate, tracks revision-oriented engagement.

Keywords

Prompting depth, Cognitive engagement, Learning analytics, Generative AI chatbots, EFL writing

1. Introduction

Generative-AI chatbots now sit inside many EFL writing classrooms, yet learners do not draw on them equally: some interrogate, revise, and reason with the tool, while others issue

surface requests and copy the reply back into their drafts. This uneven uptake is a question of access as much as ability—who converts AI assistance into thinking, and who does not—and it speaks directly to concerns about fairness and cognitive diversity in digital language use. Cognitive engagement can be defined as the level of thinking that a learner puts into a task, and it is treated as an important aspect of meaningful learning, yet it is hard to observe because it happens internally. It is usually measured through post-writing questionnaires that report what the learner thinks about the performance rather than the writing process itself [1].

Conversations between students and chatbots offer a largely neglected source of data for this problem. Every learner input carries clues to how the writer thinks, plans, and monitors the task, yet such behavioural data are rarely analysed for their cognitive significance. Prior work indicates that prompt type and structure vary with a learner's competency and shape the responses received [2]. Prompts are therefore not mere directives but reflect a learner's cognitive state. On this basis the paper develops prompting depth—the degree of specificity, elaboration, metacognitive focus, higher-order orientation, and iterative refinement of a prompt—as an indicator of engagement that can be read from the visible transcript.

Drawing on RECIPE4U, a publicly available dataset of real exchanges between 130 EFL learners and ChatGPT in university writing classes, with 1,913 student prompts and records of essay editing [3], the paper addresses three questions: how prompting depth can be defined and measured from dialogue records (RQ1); which aspects of prompting depth function as consistent behavioural correlates of revision-oriented engagement (RQ2); and whether prompting depth predicts learners' revision of their writing (RQ3). Engagement with AI-supported feedback is itself unequal, with only some learners showing cognitive engagement through dialogue and revision [4]. By using prompting depth as a trace and testing each dimension against observed revision, the study identifies which aspects of student prompting genuinely track cognitive engagement in EFL writing.

2. Theoretical Framework

Cognitive engagement refers to the extent to which a learner uses cognitive processes to work on a task, and it is theorised as the route by which rote activity becomes true learning. The ICAP model is one such theory: overt engagement is distinguished into four modes—interactive, constructive, active, and passive—with the claim that constructive and interactive activity produce more learning [5]. Constructive processing, in which learners generate ideas that go beyond the given information, is held to be the core of deep processing. This matters here because elaborative and generative elements in student prompts are treated as belonging to the same spectrum; the ordering of the modes, however, is taken as a hypothesis to be tested

rather than a premise, and the analysis below bears it out only in part.

Inferring inner states from observable measures has long been the work of learning analytics, which has used traces such as logins, clicks, and time on task as proxies for motivation, regulation, or engagement. The approach rests on the premise that behaviour is indicative of cognition, and its central problem is validity: a trace earns its status as a proxy only if it correlates with an external measure of the construct [6]. Engagement models built on digital traces explain meaningful variance when several theory-based variables are combined, but each behavioural variable alone gives an incomplete and unreliable signal, since surface behaviour can diverge from deeper processing [7]. The present study follows this logic: prompting depth is treated not as a single variable but as several, each of which must earn proxy status in its own right.

Prompting can also be read as a literacy practice rather than a mechanical act of issuing instructions, in which students reflect on how they approach tasks and steer tools toward them. Work on prompt literacy treats prompts as a competence that can be described, taught, and assessed, which makes them cognitive artifacts rather than random inputs [8]. Read this way, a record of a student's prompts becomes a trace of the reasoning behind them.

These threads meet in the present construct. Prompting depth is operationalised on five dimensions with separate anchors: specificity and higher-order request to ICAP's active–constructive and constructive–interactive distinctions, elaboration to depth of processing, metacognitive framing to self-regulated learning, and iterative refinement to interactive engagement. Because the dimensions come from different traditions, there is no requirement that they behave alike, and repeated contact in particular may signal difficulty as readily as depth. An association-and-validity design is therefore used.

Prior work bears on this question from three directions but leaves the gap the study addresses. Studies of prompt patterns show that how learners phrase requests varies with AI literacy and shapes what they receive [2], and work on prompt literacy treats such phrasing as a teachable competence [8]; both describe prompts rather than test them as indicators of engagement. Research on engagement with AI feedback finds that only some learners engage cognitively and revise [4], and intervention studies report gains in participation from chatbot use [9][10][11], yet these gauge engagement through self-report or aggregate outcomes rather than through the prompt record. Behavioural-trace research establishes the logic of validating a trace against an external criterion and warns that any single trace is unreliable [6][7], but it has not been applied, dimension by dimension, to student–chatbot prompts. The present study sits at this intersection: it reads prompts as traces, separates their depth into dimensions, and tests each against observed revision.

3. Methodology

3.1. Dataset and context

The study uses the RECIPE4U dataset, a publicly available record of real exchanges between EFL learners and ChatGPT in writing courses at a Korean university [3]. The full dataset covers 212 students across three courses; the present analysis uses the subset of 130 learners for whom prompt-level coding was completed, comprising 508 within-week sessions and 1,913 student prompts. On the RECIPE platform, ChatGPT (GPT-3.5-turbo) was configured as an English-writing tutor and asked to guide students step by step: it first invited each student to summarise what they had learned in class, then supported revision through dialogue. Students could edit their essays during the interaction and rated the preceding ChatGPT response on a five-point satisfaction scale. Because the data come from actual coursework rather than an experiment, they carry ecological validity for observing how learners prompt a chatbot while writing.

The analysis is a secondary use of the de-identified, publicly released RECIPE4U dataset [3], collected under the original study's consent and containing no direct identifiers; use here is confined to the dataset's stated research purpose, and no further ethical approval was required.

3.2. Variables

Prompting depth is the independent variable, measured on five dimensions (below) and aggregated to the student level for the validity analysis. Two behavioural criteria in the dataset serve as benchmarks for this proxy: whether a student revised the essay after a response, a binary marker of constructive uptake; and the student's satisfaction with the chatbot's response, on a 1–5 scale. With no cognitive-engagement scale in the data, these behavioural criteria approximate engagement rather than measuring it directly, and satisfaction is treated only as an auxiliary criterion, since a learner can be satisfied without being engaged. Course is retained as a control. The three courses are Intermediate Reading and Writing (IRW; referred to as Intermediate Writing in Han et al., 2024 [3]), Advanced Writing (AW), and Scientific Writing (SW), which differ in required proficiency and assignment type. Course is a three-level nominal variable, entered as two dummy indicators with IRW as the reference category; in the analysed subset it comprises 74 students / 1,143 prompts (SW), 30 / 469 (AW), and 26 / 301 (IRW).

3.3. The prompting-depth coding scheme

The five dimensions are defined against an established construct and scored on an ordinal three-point scale, summarised in Table 1. Specificity captures how precisely a prompt targets a place or problem in the text, from generic requests ("Evaluate my essay please") to sentence-

level targeting. Elaboration captures whether a request is a bare command or is developed with reasoning. Metacognitive framing captures explicit metacognition about one's own writing. Higher-order request captures the cognitive level of what is asked, from surface actions to evaluation, negotiation, and generation, and was scored semi-automatically from the dataset's intent annotations: requests for information, revision, confirmation, or question were scored 1; evaluation, negotiation, and generation were scored 2; and non-request turns, acknowledgements, surface language-use requests, and translation requests were scored 0 (no higher-order request). As shown in Table 1, the first three dimensions are human-coded per prompt; higher-order request is derived from intent annotations; and iterative refinement is computed at the session level from the sequence of turns. Prompt- and session-level scores are then combined per participant. Revision is analysed as constructive uptake in the writing process, not as writing quality.

Table 1

Coding scheme for the five dimensions of prompting depth

Dimension	Theoretical anchor	Level	Score 0 / 1 / 2 (brief criterion)
Specificity	ICAP (active–constructive)	Prompt	Generic request / targets a paragraph / targets a specific sentence with quoted text
Elaboration	Depth of processing	Prompt	Bare command / brief justification / developed reasoning
Metacognitive framing	Self-regulated learning	Prompt	None / implicit goal / explicit self-monitoring
Higher-order request	ICAP (constructive–interactive)	Prompt	No higher-order request (surface correction, translation, acknowledgement, or non-request turn) / information, revision, confirmation, question / evaluation, negotiation, generation
Iterative refinement	Interactive engagement	Session	Single turn / limited follow-up (2–3 turns) / sustained build-on (≥4 turns)

Because ICAP was developed to classify modes of engagement with instruction rather than to score prompts, its four modes are used here as heuristic anchors for where a dimension is expected to fall, not as claims that a given level instantiates a particular mode. Two anchors span adjacent modes by design. Specificity is anchored at the active–constructive boundary: moving from a generic request ("evaluate my essay") to one that targets a paragraph requires the learner to attend to and select a locus in the text, which is active engagement in ICAP terms, while quoting and reworking a specific sentence begins to generate content beyond the given, which is constructive. Higher-order request spans constructive–interactive because evaluation and generation are constructive acts, whereas negotiation, in which the learner contests or builds on the tutor's prior turn, is interactive. These anchors are hypotheses about where depth should register rather than measurements of ICAP mode. The analysis tests them, and the results do not bear all of them out: higher-order request and iterative refinement diverge from what a strict ICAP mapping would predict, which is part of the study's contribution rather than an assumption built into the coding.

3.4. Inter-coder reliability

The three dimensions requiring interpretive judgement were manually coded. All 1,913 prompts were scored with the coding manual. A randomly selected 20.0% (383 prompts) were independently coded by two coders with EFL-writing experience for the reliability assessment, while the remaining 1,530 prompts were single-coded with the same procedure. Inter-rater reliability used Cohen's κ (Table 2); agreement ranged from 0.74 to 0.84, from substantial to near-perfect. Disagreements, 9.3% of the coded decisions, were resolved by averaging the two coders' scores.

Table 2

Inter-coder reliability for the human-coded dimensions (n = 383)

Dimension	Cohen's κ	Agreement level
Specificity	0.84	Almost perfect
Elaboration	0.75	Substantial
Metacognitive framing	0.83	Almost perfect

3.5. Analytic approach

Because multiple prompts come from each learner, dimension scores are aggregated to the student level before testing their validity: the prompt-level dimensions are averaged per student,

and the session-level dimension (iterative refinement) is averaged across a student's sessions. A session is defined as all chronologically ordered student–ChatGPT turns sharing the same student identifier, course, week, and within-week session number, with the turn index ordering prompts within a session. Student-level correlations then test how each dimension aligns with the revision and satisfaction criteria, followed by a linear regression of the revision rate on all five dimensions. Prediction is examined with logistic regression at the prompt level, with course as a control and with a random intercept for student to account for the nested structure. Each dimension is examined separately, since theory predicts that they behave differently.

4. Results

4.1. The profile of prompting depth

Table 3 shows that the five dimensions have distinct distributions. Specificity and metacognitive framing sit low: about seven in ten prompts are at level 0 for specificity (M = 0.37), and metacognitive framing is the rarest (M = 0.12). Elaboration is more spread (M = 0.53), as many prompts carry at least some reasoning, and iterative refinement is the most present at the session level (M = 1.25), since many dialogues run to four turns or more. Higher-order request, scored across all prompts with non-request turns at level 0, sits low (M = 0.56), with evaluation, negotiation, and generation making up a minority of requests. Where depth appears, it surfaces more through elaboration and interaction than through precise targeting or regulation.

Table 3

Distribution of prompting depth across the five coded dimensions

Dimension	Level 0 (%)	Level 1 (%)	Level 2 (%)	Mean
Specificity	69.1	22.8	6.5	0.37
Elaboration	60.3	23.3	13.5	0.53
Metacognitive framing	89.7	7.5	1.8	0.12
Higher-order request	59.0	25.7	15.3	0.56
Iterative refinement	17.9	39.6	42.5	1.25

Note. Specificity, elaboration, metacognitive framing, and higher-order request are scored at the prompt level; iterative refinement is scored at the session level. Higher-order request scores all 1,913 prompts, with non-request turns at level 0; the full correlation matrix and regression appear in Tables S4–S5, and a sensitivity analysis treating non-request turns as missing is reported in Table S7.

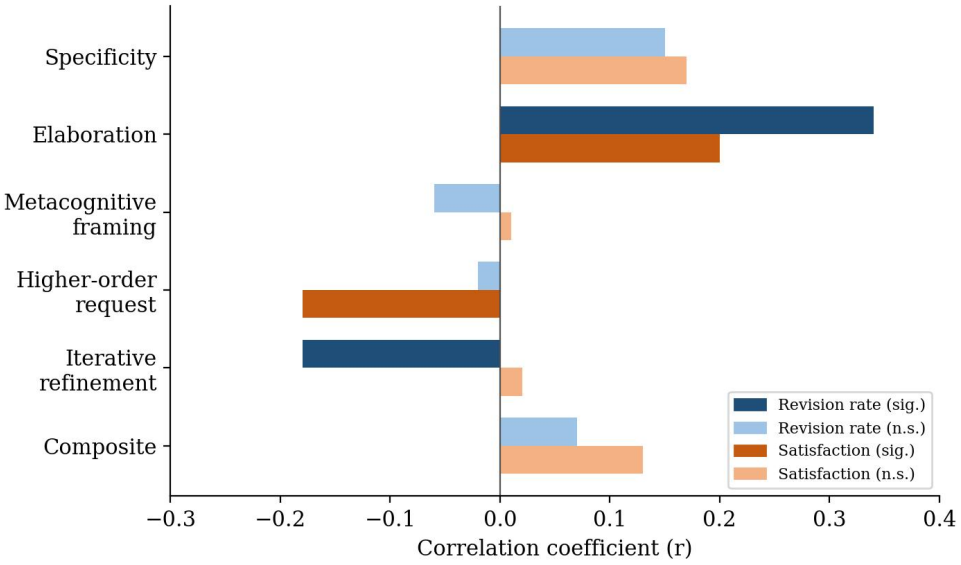
4.2. The validity of prompting depth as a proxy

Assessed at the student level against the two behavioural criteria, the dimensions diverged rather than pointing the same way. Only elaboration correlated with the essay-revision rate ($r = .34$, $p < .001$) and, more weakly, with satisfaction ($r = .20$, $p = .02$). Specificity related positively to revision but below conventional significance ($r = .15$, $p = .09$). Metacognitive framing was unrelated to either criterion ($|r| < .07$). Higher-order request was unrelated to revision ($r = -.02$) but showed a weak negative association with satisfaction ($r = -.18$, $p = .04$). Iterative refinement correlated negatively with revision ($r = -.18$, $p = .04$). Entering all five dimensions in a regression on the revision rate reproduced the pattern: elaboration was the only independent predictor ($b = 0.20$, $p = .002$), with $R^2 = .13$. An equal-weight composite of the five dimensions related to neither criterion ($r = .07$, $p = .46$). The composite is therefore uninformative, and results are reported dimension by dimension, since an informative dimension is combined with uninformative or negatively related ones.

Figure 1 visualises these dimension-level correlations; the full correlation matrix, including intercorrelations among dimensions, is given in Table S4.

Figure 1

Dimension-level correlations between prompting depth and the two behavioural criteria

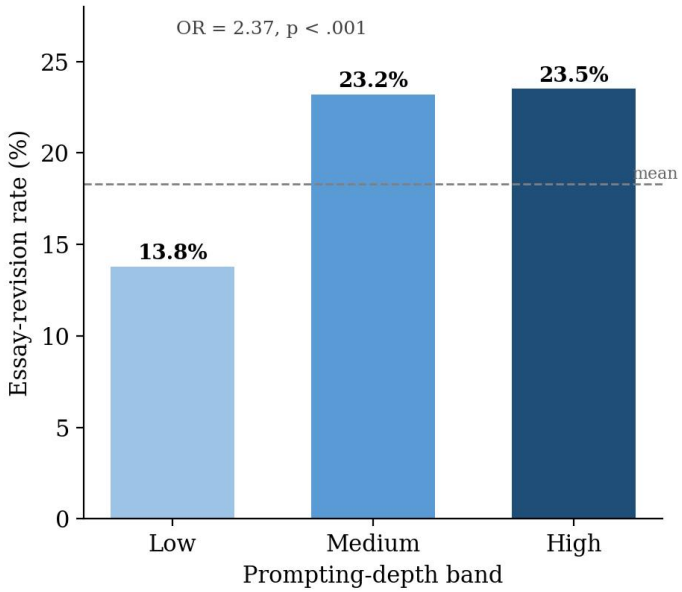


4.3. Predicting essay revision

Deeper prompting predicted subsequent revision. In a prompt-level logistic model, each one-point increase in prompting depth more than doubled the odds of a following revision (OR = 2.37, $p < .001$), and the association held after controlling for course (OR = 2.09, $p < .001$). Because prompts are nested within students, a random-intercept logistic model was also fitted; between-student variance was substantial (ICC = .41), yet the depth effect was essentially unchanged (OR = 2.30, 95% CI [1.49, 3.55], $p < .001$), indicating that clustering does not account for the association. Among individual dimensions, elaboration again carried the effect (OR = 1.54). Banding prompts by depth showed a gradient: the revision rate rose from 13.8% in the shallowest prompts to about 23% in the deepest. The gap is modest but consistent across bands.

Figure 2

Essay-revision rate across levels of prompting depth (OR = 2.37, p < .001)



5. Discussion

The study asked whether the depth of learner prompts can be read as a proxy for cognitive engagement in EFL writing. Rather than acting as one homogeneous indicator, prompting depth was informative only along specific lines. Elaboration—how far a student reasons about a request beyond issuing it—was the only dimension correlated with both behavioural criteria and the only independent predictor of revision after the other dimensions were held constant (b = 0.20, $p = .002$). This fits the theoretical expectation that constructive processing, in which learners build new content from existing information, marks deeper engagement, and it extends

work treating prompting as a literacy that makes learner cognition explicit [8] by identifying which element of that literacy best tracks engagement. One reason may be that explaining one's reasoning makes the writing problem explicit, and therefore more tractable.

The dimensions that failed as proxies are as informative. Higher-order request and metacognitive framing did not predict revision, which suggests that an apparently higher-order request does not by itself mark greater engagement in the act of writing. Higher-order request even showed a weak negative link to satisfaction ($r = -.18$), so an ostensibly demanding request did not leave learners more satisfied. More striking is the inverse relation between iterative refinement and revision ($r = -.18$, $p = .04$): sustained interaction should not be taken as an automatic sign of engagement. Repeated turns may signal difficulty with the task, or the offloading of cognitive effort to the chatbot rather than its internalisation, a reading that connects to accounts of uneven engagement with AI-supported feedback, where cognitive engagement appears only in some learners [4]. This tempers optimistic reports of chatbot-driven gains in participation [9], since such gains may concentrate among particular learners or prompting styles. The gradient in revision across depth bands, from 13.8% to about 23%, indicates that depth carries some predictive value, though a modest one.

A second reading deserves weight. Prompting depth may partly index a learner's familiarity with AI systems—prompt literacy—rather than cognitive engagement as such: students who know how to phrase elaborated requests may also be those who revise, so elaboration could track skill in using the tool as much as depth of thought. The present design cannot separate these, and the association it reports is not evidence that prompting depth measures engagement directly. This uneven command of AI dialogue is itself consequential: if only some learners convert chatbot access into elaborated, revision-oriented use, the tool may widen rather than narrow differences in who benefits from digital writing support.

For teaching, the pattern suggests that directing students to clarify their thinking when they work with a chatbot, rather than to make requests look more complex or to prolong the exchange, is the more promising route to revision. This complements evidence that chatbot responses can support writing development [11] and that dialogic tasks raise participation [10]. For learning analytics, the data argue for disaggregated, theory-informed measures over aggregate indices that can bury an informative signal.

The study has limits. The behavioural markers measure engagement indirectly, and satisfaction is a poor proxy for the construct; revision reflects constructive uptake, not gains in writing quality. The data come from a single platform, a single Korean EFL context, and a cross-sectional design, so no causal claims follow. Future work could pair the coding scheme with a validated engagement measure, extend it beyond one language and platform, and automate

elaboration detection.

6. Conclusion

Prompting depth was not supported as a unitary proxy for cognitive engagement; rather, elaboration emerged as the most consistent behavioural indicator of revision-oriented engagement. Across 1,913 prompts, the dimension that captured whether a learner reasoned about a request, rather than the sheer specificity, metacognitive phrasing, or persistence of that request, was the one that tracked essay revision (OR = 2.37), an association that held under a random-intercept model (OR = 2.30) despite substantial between-student variance. The equal-weight composite carried no signal (r = .07), which cautions against collapsing "depth" into a single number.

The contribution is to show that a disaggregated, theory-based reading of prompts can make a previously invisible construct analysable, and to give EFL writing instruction a specific focus: the reasons learners give to the chatbot. What the study does not settle is whether elaboration indexes engagement itself or its by-products. Revision marks constructive uptake, not writing quality, and satisfaction proved a poor stand-in for either. Whether elaboration keeps its edge against a validated engagement measure, across languages and chatbots, and once its detection is automated, remains open.

Acknowledgments

The author thanks the creators of the RECIPE4U dataset (Han et al., 2024) for making the data available for research use.

Author contributions

Sihao Zhang individually undertook all roles including conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, original draft writing, review & editing, visualization, supervision and project administration for this study, and has read and agreed to the published version of the manuscript.

Funding

Not applicable.

Data availability

This study is a secondary analysis of the RECIPE4U dataset (Han et al., 2024), available for research purposes at https://zeunie.github.io/RECIPE4U/. The data underlying this study are available from the corresponding author on reasonable request.

WISDOM ACADEMIC https://doi.org/10.63808/lea.v2i1.485 Volume 2 , Issue 1

Declarations

Conflict of interest

The author declares that there is no conflict of interest.

Ethics approval

This study is a secondary analysis of the de-identified, publicly released RECIPE4U dataset (Han et al., 2024). No new data involving human participants were collected, and no further ethical approval was required.

Consent to participate

Not applicable.

Consent for publication

The author has given consent for publication.

Additional information

Received: 23 July 2026

Accepted: 10 August 2026

Published online: 25 November 2026

Publisher's Note

Wisdom Academic Press Ltd. remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

[1] Bergdahl, N., Bond, M., Sjöberg, J., Dougherty, M., & Oxley, E. (2024). Unpacking student engagement in higher education learning analytics: A systematic review. *International Journal of Educational Technology in Higher Education, 21*(1), 63. https://doi.org/10.1186/s41239-024-00493-y

[2] Kim, J., Yu, S., Lee, S. S., & Detrick, R. (2026). Students' prompt patterns and its effects in AI-assisted academic writing: Focusing on students' level of AI literacy. *Journal of Research on Technology in Education, 58*(3), 638–655. https://doi.org/10.1080/15391523.2025.2456043

[3] Han, J., Yoo, H., Myung, J., Kim, M., Lee, T. Y., Ahn, S. Y., & Oh, A. (2024). RECIPE4U: Student–ChatGPT interaction dataset in EFL writing education. *In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (LREC-COLING 2024) (pp. 13666–13676). https://doi.org/10.63317/294wito6d2sp

[4] Rong, M., Yao, Y., Li, Q., & Chen, X. (2025). Exploring student engagement with artificial intelligence-guided chatbot feedback in EFL writing: Interactions and revisions. *Computer Assisted Language Learning*, 1-30. https://doi.org/10.1080/09588221.2025.2539979

[5] Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

[6] Jovanović, J., Saqr, M., Joksimović, S., & Gašević, D. (2021). Students matter the most in learning analytics: The effects of internal and instructional conditions in predicting academic success. *Computers & Education*, 172, 104251. https://doi.org/10.1016/j.compedu.2021.104251

[7] Winter, M., Mordel, J., Mendzheritskaya, J., Biedermann, D., Ciordas-Hertel, G. P., Hahnel, C., … Horz, H. (2024). Behavioral trace data in an online learning environment as indicators of learning engagement in university students. *Frontiers in Psychology*, 15, 1396881. https://doi.org/10.3389/fpsyg.2024.1396881

[8] Tour, E., & Zadorozhnyy, A. (2025). Conceptualizing and operationalizing prompt literacy for English language learners. *Journal of Adolescent & Adult Literacy, 69*(3), e70020. https://doi.org/10.1002/jaal.70020

[9] Anjarani, S., Furqon, M., & Rakhmawati, F. (2024). ChatGPT to foster students' engagement in writing class: An intervention study. *Voices of English Language Education Society, 8*(3), 630–640. https://doi.org/10.29408/veles.v8i3.27687

[10] Guo, K., Zhong, Y., Li, D., & Chu, S. K. W. (2024). Investigating students' engagement in chatbot-supported classroom debates. *Interactive Learning Environments, 32*(9), 4917–4933. https://doi.org/10.1080/10494820.2023.2207181

[11] Polakova, P., & Ivenz, P. (2024). The impact of ChatGPT feedback on the development of EFL students' writing skills. *Cogent Education, 11*(1), 2410101. https://doi.org/10.1080/2331186X.2024.2410101