

Google Gemini-assisted writing and lexical appropriateness in EFL Paragraph Writing: Evidence of stability across tasks

Pandu Prasodjo^{a,b,1}, Siusana Kweldju^{a,2}, Nur Mukminatien^{a,3}, Francisca Maria Ivone^{a,4}, Hilarius Raditya Priambada Purba^{b,5}



^a Universitas Negeri Malang, Jl. Cakrawala No.5, Sumber Sari, Kec. Lowokwaru, Kota Malang, Jawa Timur 65145, Indonesia

^b Universitas Internasional Batam, Baloi-Sei Ladi, Jl. Gajah Mada, Tiban Indah, Kec. Sekupang, Kota Batam, Kepulauan Riau 29426, Indonesia

¹ pandu@uib.ac.id *, ² siusana.kweldju.fs@um.ac.id, ³ nur.mukminatien.fs@um.ac.id, ⁴ francisca.maria.fs@um.ac.id, ⁵ hilarius@uib.ac.id

*Corresponding author

ARTICLE INFO

Article history

Received 02 January 2026

Revised 09 March 2026;

Accepted 02 April 2026

Available Online 30 April 2026

Keywords

Artificial intelligence

EFL

Google Gemini

Lexical appropriateness

Paragraph Writing

ABSTRACT

This study examined the influence of Google Gemini, a generative artificial-intelligence (AI) tool, on the lexical appropriateness of Indonesian EFL students' paragraph writing. Drawing on theories of lexical richness and appropriateness, the study employed an exploratory quasi-experimental design involving three writing conditions: a baseline task, an AI-assisted revision task, and a delayed post-test without AI support. Fourteen second-year English Language Teaching students (A2-B1 CEFR) participated in the study. Lexical appropriateness was measured using Type-Token Ratio (TTR) and Measure of Textual Lexical Diversity (MTLD), and the data were analyzed using repeated-measures ANOVA to examine differences across tasks. The results revealed no statistically significant differences in lexical diversity across the three conditions, suggesting that short-term interaction with Google Gemini did not lead to measurable improvement in lexical appropriateness. Qualitative observations further indicated that limited vocabulary depth and low metalinguistic awareness constrained students' ability to evaluate and apply AI-generated lexical suggestions. These findings suggest that lexical appropriateness is reflective, and instruction-dependent process, and that AI tools such as Google Gemini function more as a potential scaffolding partner rather than autonomous enhancers of lexical development.



© The Authors 2026. Published by Universitas Ahmad Dahlan.
This is an open access article under the [CC-BY-SA](#) license.



How to Cite: Prasodjo, P., Kweldju, S., Mukminatien, N., Ivone, F. M., Purba, H. R. P. (2026). Google Gemini-assisted writing and lexical appropriateness in EFL Paragraph Writing: Evidence of stability across tasks. *English Language Teaching Educational Journal*, 9(1), 69-82. <https://doi.org/10.12928/eltej.v9i1.15214>

1. Introduction

1.1. Lexical Appropriateness

Lexical appropriateness refers to a writer's ability to select words that are accurate in meaning, contextually suitable, and stylistically consistent with communicative intent. It extends beyond lexical variety or diversity by emphasizing fit: how well a lexical choice aligns with situational context, discourse register, and pragmatic nuance (Dewaele, 2008; Durrant & Durrant, 2022). Within second language writing, this concept represents the intersection between vocabulary knowledge, contextual awareness, and communicative precision, all of which are essential indicators of lexical maturity (Crossley et al., 2014; Laufer & Nation, 1995). Unlike lexical diversity, which measures the range of vocabulary used, or lexical sophistication, which tracks low-frequency word use, lexical appropriateness evaluates whether chosen words convey the intended meaning naturally and effectively

Research in English as a Foreign Language (EFL) writing contexts indicates that learners frequently struggle with appropriateness despite possessing adequate grammatical competence. Indonesian EFL students, for instance, tend to rely on “lexical teddy bears” (Hasselgård, 2019; Kweldju, 2004) which are safe, high-frequency words such as good, nice, or bad that are semantically shallow and repetitive. This overreliance diminishes precision and stylistic variation. The underlying issue stems not only from limited vocabulary breadth but also from insufficient lexical depth, including collocational knowledge, register sensitivity, and connotative understanding (Nurweni & Read, 1999; Rahmat & Coxhead, 2021). Consequently, lexical inappropriateness often manifests in word choices that are grammatically correct yet pragmatically awkward or imprecise.

Several factors contribute to this challenge. Instruction in many EFL classrooms remains dominated by grammatical accuracy (Cahyono & Widiati, 2006; Li & Iwashita, 2019), leaving lexical development underemphasized. Furthermore, limited exposure to diverse textual genres restricts learners’ awareness of register variation (Hyland, 2003; Songbatumis, 2017). Affective variables, such as fear of making mistakes, also prompt learners to choose “safe” vocabulary instead of experimenting with nuanced expressions (Aulia, 2018). The combination of these factors produces writing that is syntactically sound but lexically flat.

Developing lexical appropriateness, therefore, requires metalinguistic and metacognitive engagement. Learners must reflect on word meaning, collocation, and tone, and assess how lexical choices shape communicative clarity (Mahajan, 2020; Nagy, 2007). Previous studies indicate that instruction targeting these dimensions, such as vocabulary-focused writing tasks (Lee, 2003), reflective revisions (Laqaei, 2015), and lexical feedback (Ali, 2024), can improve learners’ lexical awareness and contextual precision. Yet such practices are time-intensive and rarely sustained in classroom settings. This gap has led scholars to explore AI-assisted feedback as a supplementary scaffold for lexical learning (Javaid et al., 2023; Phuong & Anh, 2024).

Although lexical appropriateness is primarily a qualitative construct, it is closely related to patterns of lexical variation observable in learner output. Measures such as Type-Token Ratio (TTR) and Measure of Textual Lexical Diversity (MTLD), while not direct indicators of appropriateness, can provide insight into the range of lexical choices available to learners. A limited lexical repertoire often constrains learners to high-frequency, semantically general vocabulary, which may reduce contextual precision. Conversely, greater lexical diversity increases the potential for selecting more contextually appropriate and nuanced expressions. In this sense, quantitative indices of lexical diversity can serve as indirect proxies for examining constraints and possibilities in lexical appropriateness.

1.2. Paragraph Writing in EFL Context

Paragraph writing constitutes the foundational stage of academic composition in English Language Teaching (ELT). It involves not only organizing ideas into a coherent structure, i.e. topic sentence, supporting details, and concluding statement, but also selecting words and expressions that convey precise meaning and appropriate tone (Rohim, 2019; Wirantaka, 2016). In EFL contexts, paragraph writing bridges sentence-level accuracy and discourse-level coherence, serving as the medium through which learners demonstrate grammatical control, logical reasoning, and lexical precision.

Despite its pedagogical importance, paragraph writing remains one of the most challenging skills for EFL learners. Numerous studies report persistent problems in cohesion, coherence, and lexical selection (Tustiawati & Wiguna, 2024; Wali & Madani, 2020). Indonesian students, in particular, often exhibit limited vocabulary range and overdependence on structural templates, producing writing that is syntactically correct but semantically shallow (Rohim, 2019; Wau, 2022). These issues are compounded by restricted classroom time, insufficient feedback on vocabulary use, and students’ low confidence in experimenting with alternative wordings (Rahman, 2022; Usman, 2018).

Effective paragraph instruction, therefore, requires an integrated approach that balances form, meaning, and lexical choice. Rather than treating paragraph construction as a mechanical process, instruction should engage students in idea development, i.e. planning, drafting, and revising with explicit attention to lexical appropriateness (Bekele et al., 2022; Schaefer, 2023). Teachers play a key role in modelling how vocabulary influences precision and coherence and in designing reflective activities that allow students to assess their lexical decisions. Scaffolded tools such as vocabulary banks, lexical checklists, and guided peer review can gradually build learners’ awareness of how words create nuance and cohesion.

Recent research suggests that technology-assisted feedback can strengthen this process. When integrated purposefully, generative AI tools can function as reflective partners in providing real-time lexical alternatives, collocations, or contextual rephrasings that help students notice gaps in their lexical repertoire (Phuong & Anh, 2024; Shidiq, 2023). However, uncritical dependence on AI may result in surface-level editing rather than genuine lexical learning. Thus, AI must be embedded within a process-based writing framework, where learners are guided to evaluate, adapt, and justify AI-generated suggestions. This integration transforms paragraph writing from a static structural exercise into a dynamic, metalinguistic practice that cultivates both lexical awareness and communicative competence.

1.3. Generative AI Google Gemini

Google Gemini is a generative artificial intelligence (Gen AI) platform developed by Google that integrates large language model (LLM) capabilities with multimodal processing, enabling real-time text generation, rewriting, summarization, and contextual feedback (Silva & Ulbricht, 2024; Phuong & Anh, 2024). As part of the newest generation of LLMs, Gemini is characterized by its ability to produce context-sensitive lexical alternatives, generate paraphrases, and offer interactive explanations.

Research on AI-assisted writing has expanded rapidly over the last three years, reflecting broader shifts in language education toward automated feedback, adaptive learning, and learner-centered technologies (Crompton et al., 2024; Darwin et al., 2024). Teachers generally perceive AI tools as supportive in enhancing student engagement, reducing feedback workload, and supplementing traditional instruction (Moybeka et al., 2023; Nazim, 2024). At the same time, scholars caution that AI-generated feedback may promote surface-level revisions if learners adopt suggestions without critically evaluating their contextual appropriateness (Kasneji et al., 2023; Prasodjo et al., 2025; Zhai et al., 2024). Thus, AI's pedagogical value depends not only on its capabilities but also on how learners interact with the feedback it provides.

Compared with earlier AI writing assistants such as early generative models or grammar-oriented tools like Grammarly, Gemini is designed to function as a dialogic assistant (Gemini Team et al., 2023) that provides multi-turn feedback, clarifications, and explanations. This affords opportunities for deeper engagement with lexical alternatives, collocation patterns, and stylistic variation (Barrot, 2024). Unlike rule-based systems, Gemini can supply multiple rephrasing options or gradations of meaning (Gemma Team et al., 2024), potentially helping EFL learners notice subtle semantic differences. For example, Gemini may offer choices such as investigate, examine, or look into, prompting learners to consider tone, formality, and precision.

Despite these affordances, empirical investigations into Gemini's impact on EFL writing remain scarce. Most published AI-in-ELT research focuses on ChatGPT, general AI literacy, or grammar improvement (Lee, 2025; Maeda, 2024). Very few studies have examined how Gemini supports lexical development, particularly in foundational writing tasks such as paragraph writing where vocabulary determines clarity, nuance, and coherence. This gap is notable because EFL learners often struggle with lexical appropriateness due to restricted vocabulary depth, limited collocational knowledge, and inadequate metalinguistic awareness (Hartono & Prima, 2021; Kweldju, 2004).

At the same time, concerns remain regarding the risk of AI-overreliance, where learners adopt AI-generated vocabulary without understanding usage, register, or nuance (Alharbi, 2023). This phenomenon may hinder rather than enhance lexical development if students treat Gemini as a source of ready-made solutions instead of a reflective partner. Consequently, scholars call for pedagogical frameworks that position AI tools as guided scaffolds within process-based writing instruction, rather than autonomous problem-solving machines (Bedrio et al., 2025; Shidiq, 2023; Strobl et al., 2024).

Google Gemini offers promising but largely unexplored potential for supporting lexical appropriateness in EFL writing. Its contextualized suggestions and dialogic interaction features may help learners notice lexical gaps, reflect on word choices, and revise with greater precision. However, whether these capabilities translate into measurable improvements particularly for lower-intermediate learners with limited vocabulary depth, remains an open empirical question. This study addresses this gap by investigating how Gemini-assisted writing influences EFL learners' lexical appropriateness across multiple writing tasks.

1.4. Statement of Problems

Previous studies on Indonesian EFL learners show persistent difficulties in lexical appropriateness, largely stemming from limited vocabulary depth, restricted exposure to varied English input, and low

metalinguistic awareness (Hartono & Prima, 2021; Limantoro, 2018; Saadi & Saadat, 2015). Many students equate writing proficiency with grammatical accuracy while overlooking the importance of precise, context-sensitive lexical choices. This misconception often results in paragraphs that are structurally correct but semantically vague or imprecise. The issue is particularly salient in lower-intermediate learners, whose vocabulary size frequently falls below the threshold required for nuanced expression and lexical decision-making (Novianti, 2016; Nurweni & Read, 1999).

Paragraph Writing courses aim to strengthen foundational academic writing skills, including coherence, unity, and clarity of expression. Achieving these outcomes requires not only correct sentence construction but also the ability to select vocabulary that accurately conveys meaning within a specific communicative context (Rohim, 2019; Wirantaka, 2016). However, classroom constraints, such as limited feedback time, predominant focus on grammar, and minimal practice in evaluating lexical choices, often leave learners without adequate support for developing lexical appropriateness.

Gen AI such as Google Gemini presents a potential avenue for addressing these challenges by offering instant lexical suggestions, alternative phrasings, and contextualized feedback. These features may help learners notice lexical gaps and consider more precise vocabulary options during the writing process. Yet, it remains unclear whether lower-intermediate EFL students have the lexical readiness and metalinguistic awareness necessary to evaluate, adapt, and internalize Gemini-generated suggestions. Without such abilities, students may adopt AI output superficially, resulting in minimal improvement or potentially inappropriate lexical substitutions.

Despite growing interest in AI-assisted writing, several gaps remain underexplored. First, limited attention has been given to lexical appropriateness as a distinct construct, as most studies prioritize lexical diversity or general writing improvement without examining whether word choices are contextually suitable. Second, empirical evidence on the role of Google Gemini specifically remains scarce, with existing research largely centered on other AI platforms such as ChatGPT. Third, little is known about how lower-intermediate EFL learners engage with AI-generated lexical suggestions and whether such interaction leads to sustained lexical development beyond immediate revision. These gaps highlight the need for a more nuanced investigation into the relationship between AI-assisted feedback and context-sensitive lexical use.

To address this gap, the present study investigates whether exposure to Google Gemini's lexical suggestions affects the lexical appropriateness of EFL learners' paragraph writing across three tasks with varying levels of AI involvement. The study is guided by the following hypotheses:

H₀: There is no significant difference in learners' lexical appropriateness across tasks with different levels of exposure to Google Gemini.

H₁: There is a significant difference in learners' lexical appropriateness across tasks with different levels of exposure to Google Gemini.

By examining both short-term and delayed effects of Gemini-assisted writing, this study aims to offer empirical insight into the conditions under which Gemini feedback may support (or fail to support) lexical development in foundational ELT writing instruction.

2. Method

2.1. Research Design

This study employed an exploratory quasi-experimental design to examine the effects of Google Gemini-assisted writing on EFL learners' lexical appropriateness. A quasi-experimental approach was chosen because the intervention took place in an intact Paragraph Writing class without random assignment (Shadish & Luellen, 2006). The design included three writing tasks with varying levels of AI exposure: Task 1 (Pre-test): No AI involvement, Task 2 (Post-test 1): Short-term guided interaction with Google Gemini, and Task 3 (Post-test 2): Delayed task without AI support.

Given the small sample size of 14 students drawn from an intact Paragraph Writing class through convenience sampling, and the exploratory nature of the study, the analysis focused on identifying patterns of lexical behavior rather than making broad generalizations. This framing aligns with calls in AI-in-education research to document early-stage classroom effects before scaling to larger populations (Barrot, 2024; Crompton et al., 2024).

2.2. Participants

Fourteen second-year English Language Teaching (ELT) students enrolled in a Paragraph Writing course participated in the study. Their English proficiency ranged from A2 to B1 according to CEFR descriptors, making them representative of lower-intermediate learners who frequently struggle with lexical appropriateness.

Participants were selected based on classroom membership (convenience sampling), consistent with exploratory classroom-based research. Their involvement was relevant because paragraph writing is a core component of their curriculum, and the course's learning outcomes directly involve the development of coherent, cohesive, and lexically appropriate written expression.

All participants were informed of the study's purpose, procedures, and voluntary nature. Informed consent was obtained, and anonymity was ensured during analysis and reporting.

2.3. Data Collecting Instruments

Three instruments were used to collect data on students' lexical performance and their interaction with Google Gemini: (1) writing tasks, (2) AI-interaction guidelines, and (3) lexical measurement tools.

a) *Writing tasks*

The primary data source consisted of three paragraph-writing tasks administered across different stages of the study. Each task required students to produce a 100–150-word paragraph on a controlled prompt that remained consistent in topic type and cognitive demand. All prompts were designed as personal-experience descriptive tasks, requiring students to generate and elaborate ideas at a comparable level of cognitive complexity, without involving argumentative or analytical reasoning. The uniformity of these tasks ensured comparability across the pre-test, AI-assisted, and delayed post-test conditions. All paragraphs were written by hand during class time, with no access to digital devices, to minimize reliance on external tools.

b) *Google Gemini guidelines sheet*

For Task 2, students received a structured guideline sheet outlining how Google Gemini could be used during the revision stage. The sheet included specific prompts (e.g., "Suggest alternative words for...", "Explain the difference between...", "Check the tone of this sentence") and emphasized reflective evaluation over direct acceptance. The guidelines ensured standardized exposure to AI feedback and minimized misuse or copy-paste reliance.

The guidelines functioned as an instructional scaffold designed to standardize the nature of AI interaction across participants and to prevent direct copy-pasting of AI-generated segments. This instrument ensured that students' lexical decision-making remained active rather than fully automated.

c) *Lexical measurement tools*

To evaluate lexical variety and diversity, two quantitative analysis tools were employed Text Inspector (textinspector.com) to extract word counts, type-token distribution, and MTLTD values. It provided a standardized platform for calculating lexical richness indices.

2.4. Procedure

The study was conducted over a twelve weeks period as part of regular classroom instruction. The three writing tasks were administered at four-week intervals to allow sufficient time for instructional exposure while minimizing task familiarity effects. Task 1 (pre-test) was conducted in Week 1, Task 2 (AI-assisted) in Week 5, and Task 3 (delayed post-test) in Week 9.

All paragraph-writing tasks were completed under controlled conditions. Students were required to write their initial drafts by hand during class time, with no access to digital devices, to ensure consistency and minimize reliance on external tools. The primary distinction between the three tasks lay in the availability of AI support during the revision stage. In Task 1, all paragraphs were initially written by hand during class time, with no access to AI support. In Task 2, after completing their handwritten drafts, students engaged in a guided AI-assisted revision phase lasting approximately 20–30 minutes after completing their initial handwritten drafts. The interaction with Google Gemini was structured using a guideline sheet that provided example prompts to standardize usage. Typical prompts included: "Suggest more appropriate words for this sentence," "Explain the difference between [word A] and [word B]," and "Check whether this expression is natural in this context." They used Gemini to explore alternative lexical choices and explanations, but were required to revise their paragraphs independently without copying AI-

generated sentences. Students were instructed to evaluate Gemini's suggestions critically rather than adopt them directly. Specifically, they were asked to compare alternatives, consider contextual fit, and decide whether the suggested expressions aligned with their intended meaning. Final revisions were made independently, and direct copying of AI-generated sentences was prohibited. This process was designed to promote active lexical decision-making rather than passive reliance on AI output. In Task 3, students again completed the task independently without AI support, allowing for the examination of any retained effects of prior AI interaction.

2.5. Data Analysis

Lexical appropriateness was estimated quantitatively using Type-Token Ratio (TTR) and Measure of Textual Lexical Diversity (MTLD). While these indices do not directly capture contextual appropriateness, they provide indirect indicators of learners' lexical performance. Greater lexical diversity is often associated with a wider range of lexical choices, which increase the potential for contextually appropriate and precise word selection (Crossley et al., 2014; Laufer & Nation, 1995). Conversely, limited diversity may reflect reliance on high-frequency, semantically general vocabulary, which has been linked to reduce lexical precision in EFL writing. TTR reflects lexical variety, while MTLD provides a more stable measure of lexical diversity unaffected by text length variations (Jarvis & Hashimoto, 2021). Therefore, TTR and MTLD were used as indirect measures or indicators to examine patterns of lexical variation, while acknowledging that contextual nuance and appropriateness, extend beyond what automated metrics can capture.

All paragraphs were uploaded to Text Inspector for lexical calculation. A minimum threshold of 50 words was applied for MTLD analyses to ensure reliability. Statistical analysis was performed using SPSS version 26. A repeated-measures ANOVA tested differences across the three tasks, with normality and sphericity assumptions checked prior to interpretation. Effect sizes (η^2) were reported to contextualize patterns within this small-sample exploratory design.

It is important to note that TTR and MTLD were not used as direct measures of lexical appropriateness, but rather as indicators of lexical variation that may constrain or enable contextually appropriate word choice. In this study, lexical appropriateness is interpreted as a higher-order construct, while TTR and MTLD provide observable patterns that reflect learners' lexical resources and potential for context-sensitive expression.

Accordingly, this study does not equate lexical diversity with lexical appropriateness. Instead, TTR and MTLD are treated as indirect indicators of learners' lexical resources, which may influence their ability to produce contextually appropriate expressions. Lexical appropriateness is thus conceptualized as a higher-order, context-dependent construct that extends beyond what quantitative indices alone can capture.

2.6. Ethical Considerations

The study adhered to institutional ethical standards for classroom-based research. Students were informed about the study's objectives, procedures, and their rights—including voluntary participation and withdrawal without penalty. Written consent was obtained. All writing samples were anonymized, stored securely, and used solely for research purposes.

3. Findings and Discussion

3.1. Findings

This section presents the results of the lexical measures Type-Token Ratio (TTR) and Measure of Textual Lexical Diversity (MTLD) across the three writing tasks: (1) baseline without AI, (2) AI-assisted revision using Google Gemini, and (3) delayed post-test without AI. The goal was to determine whether short-term or extended exposure to Gemini contributed to changes in lexical variety or diversity.

a) *Type-Token Ratio*

The mean TTR values across the three tasks were highly similar. As shown in [Table 1](#), students' average lexical variety was 0.656 in Task 1, increased slightly to 0.657 in Task 2, and decreased to 0.651 in Task 3. Although individual scores varied, overall changes were minimal.

Table 1. Descriptive Statistics for TTR

	N	Minimum	Maximum	Mean	Std. Deviation
TTR1	14	.48	.94	.6564	.11386
TTR2	14	.50	.75	.6571	.07290
TTR3	14	.52	.90	.6507	.09068
Valid N (listwise)	14				

The initial task showed the widest variability ($SD = 0.114$), while variability decreased in the later tasks. The lowest performer showed modest improvement (from .48 to .52), suggesting that some learners may have benefited from exposure to Gemini's suggestions. However, the highest performer slightly declined from .94 to .90, consistent with the possibility that AI feedback may have encouraged more standardized, less risk-taking lexical choices.

Table 2. Repeated ANOVA Measure

Source	SS	df	MS	F	p-value	η_p^2
Task	0.008	2	0.004	0.418	0.663	0.031
Error	0.247	26	0.009			

A repeated-measures ANOVA (Table 2) confirmed that differences across tasks were not statistically significant, $F(2, 26) = 0.418$, $p = .663$, $\eta_p^2 = .031$. Normality (Shapiro–Wilk tests) and sphericity (Mauchly's test) assumptions were met. These findings indicate that Google Gemini exposure had no measurable effect on the lexical variety of students' paragraph writing.

b) Measure of Textual Lexical Diversity

MTLD scores showed greater variability than TTR, reflecting differences in students' vocabulary depth and writing fluency. Table 3 summarizes the results for texts that met the minimum 50-word threshold.

Table 3. Descriptive Statistics for MTLD

Task	N	Mean	SD	Min	Median	Max	Range
MTLD1	12	101.34	41.07	62.29	104.45	203.28	140.99
MTLD2	14	87.04	37.81	44.63	76.21	180.40	135.77
MTLD3	13	89.40	31.99	32.05	91.72	140.87	108.82

A decline was observed from Task 1 to Task 2, followed by a slight recovery in Task 3. High SD values suggest substantial between-learner variability, which is typical for A2–B1 learners. The reduction in MTLD variability from Task 1 to Task 3 may indicate that learners converged on more familiar or "safer" vocabulary when interacting with AI, mirroring patterns seen in TTR.

Table 4. Repeated-measures ANOVA of MTLD

Source	SS	Df	MS	F	p-value	η_p^2
Task	340.45	2	170.22	0.374	0.693	0.036
Error	9096.82	20	454.84			

A repeated-measures ANOVA ($n = 11$ complete cases) found no significant differences across tasks, $F(2, 20) = 0.374$, $p = .693$, $\eta_p^2 = .036$. All assumptions were satisfied (Shapiro–Wilk normality; Mauchly's test of sphericity). These results (Table 4) suggest that neither short-term nor delayed exposure to Google Gemini produced meaningful changes in lexical diversity.

Across both indices, students' lexical performance remained statistically stable despite interacting with Google Gemini. Minor individual improvements or declines did not form a consistent pattern. The stability of scores suggests that Gemini's lexical suggestions did not lead to measurable lexical gains,

learners may have adopted AI recommendations superficially without integrating them into their linguistic repertoire, and lower-intermediate learners may require explicit lexical training before they can benefit from AI-assisted vocabulary support.

To complement the quantitative findings, an illustrative case from one participant (CT) is presented to demonstrate how lexical appropriateness evolved across tasks. In Task 1, the student produced lexically imprecise expressions such as *"I'm really hard in grammar"* and *"hard to understand,"* reflecting reliance on semantically vague and non-target-like forms. Gemini feedback explicitly suggested more appropriate alternatives, including *"I'm struggling with grammar"* and *"difficult to understand,"* highlighting issues in both lexical choice and collocational accuracy.

In Task 2, after interacting with Gemini, the student's writing showed noticeable surface-level improvement. The revised paragraph included more target-like expressions such as *"I have a hard time with grammar"* and *"I find it challenging to grasp,"* indicating partial uptake of AI-generated suggestions. However, the improvements remained largely at the level of lexical substitution rather than deeper contextual adaptation. The student continued to rely on familiar patterns and did not consistently demonstrate nuanced lexical variation or register awareness.

In Task 3, when AI support was removed, the student maintained some improved forms but did not exhibit further lexical development. While expressions were generally more grammatical and acceptable, they remained within a limited lexical range and did not reflect expanded lexical repertoire or increased contextual sensitivity. This pattern suggests that although Gemini feedback may facilitate short-term lexical correction, it does not necessarily lead to sustained improvement in lexical appropriateness without deeper metalinguistic engagement.

These findings align with prior observations that lexical appropriateness is a cognitively demanding skill requiring rich vocabulary knowledge, contextual awareness, and guided metalinguistic reflection, factors not automatically strengthened through brief AI exposure.

3.2. Discussion

This study investigated whether guided interaction with Google Gemini could influence the lexical appropriateness of EFL learners' paragraph writing. Using TTR and MTLTD as proxies for lexical variety and diversity, the study compared learners' performance across a pre-test, a short-term AI-assisted task, and a delayed task without AI support. The findings showed no statistically significant differences across the three tasks for either measure, suggesting that short-term exposure to Gemini did not produce measurable improvements in lexical performance.

a) *Lack of measurable improvement and lexical appropriateness theory*

The absence of significant gains aligns with theoretical claims that lexical appropriateness is a deep-level lexical skill, requiring not only vocabulary breadth but also depth of knowledge, contextual sensitivity, and metalinguistic awareness (Durrant & Durrant, 2022; Laufer & Nation, 1995). Lower-intermediate learners in this study who typically possess limited vocabulary resources (Novianti, 2016; Nurweni & Read, 1999) may not yet have the lexical foundation necessary to critically assess and apply AI-generated alternatives.

Instead of evaluating nuance, register, or collocation, learners may have relied on their habitual "lexical teddy bears" (Hasselgård, 2019; Kweldju, 2004), resulting in paragraphs that were grammatically correct but lexically conservative. This resonates with research showing that learners often prioritize grammatical correctness over lexical precision (Komara & Tiarsiwi, 2021; Saadi & Saadat, 2015).

b) *Learner readiness and Gemini-generated feedback*

While Gemini provided context-sensitive vocabulary suggestions, the effectiveness of such feedback depends heavily on learners' lexical readiness and metalinguistic awareness. Research in AI-assisted learning emphasizes that students must interpret, evaluate, and adapt AI suggestions, not adopt them mechanically, to benefit from the tool (Alharbi, 2023; Vita Losi et al., 2024).

For A2–B1 learners, however, this cognitive processing may exceed their current proficiency level. Many may not fully understand the difference between suggested alternatives such as *investigate*, *examine*, or *look into*. When learners cannot judge nuance, they are likely to either ignore suggestions or accept them mechanically, which limits meaningful lexical development. This explains the descriptive findings: minor fluctuations in scores, but no consistent upward trend.

To further illustrate how learners engaged with AI-generated feedback, examples from student texts reveal that lexical adjustments were often limited to surface-level substitutions rather than deeper contextual refinement. For instance, in Task 1, a student produced the expression “*I’m really hard in grammar,*” which reflects inappropriate lexical choice and collocation. Following Gemini feedback, this was revised in Task 2 to “*I have a hard time with grammar*” or “*I find grammar challenging,*” indicating improved grammaticality and collocational accuracy. However, such revisions largely involved replacing isolated lexical items rather than restructuring meaning or expanding lexical repertoire.

In some cases, AI suggestions introduced more sophisticated alternatives that learners did not fully adopt or adapt appropriately. For example, suggestions such as “*alleviate stress*” or “*contextually appropriate expressions*” were either ignored or replaced with simpler, familiar forms. This pattern suggests that while AI can provide lexically richer options, learners’ limited vocabulary depth and metalinguistic awareness constrain their ability to evaluate and integrate these alternatives effectively. As a result, lexical development remains incremental and largely surface-oriented.

c) *Google Gemini as a scaffolding tool rather than an automatic enhancer*

The findings reflect growing consensus that AI tools, including LLMs like Gemini and ChatGPT, do not automatically enhance writing skills (Crompton et al., 2024; Kasneci et al., 2023). Their effectiveness depends on instructional design, learner training, and the integration of reflective practice (Strobl et al., 2024).

In this study, Gemini functioned as a passive suggestion provider. Without structured lexical-awareness activities such as evaluating collocations, analyzing tone shifts, or comparing semantic gradients, students may have interacted with Gemini superficially. This limited interaction likely contributed to the stability of TTR and MTLD scores.

The findings reinforce that AI should be positioned as a guided scaffolding partner, not a standalone solution for lexical development. Without explicit instruction, students may focus on surface-level edits (e.g., grammar, punctuation) rather than deeper lexical issues.

d) *Stability across tasks and the role of Paragraph Writing instruction*

Results also highlight an instructional issue, e.g. paragraph writing in many EFL contexts tends to emphasize structure and grammar, often marginalizing vocabulary development (Cahyono & Widiati, 2006; Rohim, 2019). If students have internalized paragraph formulae but not lexical decision-making processes, AI support may strengthen organization but not lexical appropriateness.

The stability of learners’ lexical profiles across all tasks suggests that their lexical repertoire remained largely unchanged over time; Gemini’s suggestions did not trigger deep lexical processing; and paragraph writing tasks of 100–150 words may not provide sufficient space for meaningful lexical experimentation among lower-level learners.

e) *Implication for ai integration in EFL writing*

Despite the insignificant results, the study offers several pedagogical implications, those are 1) explicit lexical-awareness instruction is essential. Gemini should be paired with activities that guide learners to examine language features, e.g. meaning, register, collocation, and contextual fit to produce contextually appropriate and lexically precise expression; 2) Gemini interaction must be taught, not assumed. Students need training in how to prompt Gemini effectively and how to evaluate suggestions critically; 3) Gemini fits within process-based writing models (longer learning period). Gemini can support students to do brainstorming, revising, and lexical noticing, but must be aligned within the phases of writing; and 4) Teacher mediation remains central. Gemini feedback must be interpreted through teacher modeling, reflection guides, and scaffolded revision task.

Beyond explaining the absence of significant quantitative improvement, these findings highlight important implications for AI literacy in language education. The results suggest that effective use of AI tools such as Google Gemini requires not only access but also the ability to critically interpret, evaluate, and adapt AI-generated suggestions. AI literacy, in this context, involves understanding the limitations of automated feedback, recognizing context-sensitive language use, and making informed lexical decisions rather than relying on AI outputs uncritically.

From a pedagogical perspective, this underscores the need to design AI-mediated writing tasks that go beyond revision support. Instruction should incorporate guided activities that explicitly train learners

to compare lexical alternatives, analyze collocational patterns, and justify their lexical choices. For example, tasks may include evaluating multiple AI-generated options, identifying contextually inappropriate substitutions, or reflecting on why certain lexical choices are more suitable than others. Such approaches position AI not as a corrective tool, but as a catalyst for metalinguistic awareness and reflective learning within process-based writing instruction.

However, limitations must be acknowledged. First, the small sample size and quasi-experimental nature restrict generalizability. Second, TTR and MTLTD measure lexical richness but do not fully capture contextual appropriateness, which requires qualitative judgment. Third, the study relied on short paragraphs; longer texts may reveal more substantial lexical variation. These findings contribute to a more nuanced understanding of AI-assisted writing, emphasizing that technological support must be accompanied by pedagogical design and learner training to produce meaningful lexical development.

4. Conclusion

This study examined whether Google Gemini-assisted writing could influence the lexical appropriateness of Indonesian EFL learners' paragraph writing. Across three writing tasks, i.e. baseline, AI-assisted, and delayed post-test, no statistically significant differences were found in learners' lexical performance. These findings indicate that short-term interaction with Gemini does not automatically lead to measurable improvements in lexical appropriateness. The results highlight that lexical appropriateness is a complex, cognitively demanding skill that depends on vocabulary depth, contextual awareness, and metalinguistic engagement. For lower-intermediate learners, AI-generated suggestions may exceed their current lexical readiness, limiting their ability to evaluate and internalize the alternative word choices. Rather than functioning as autonomous solution, Google Gemini should be positioned as pedagogical support tool within guided, process-based writing instruction. Its potential lies in prompting lexical awareness and reflection, provided that learners receive explicit training and teacher mediation.

This study is limited by its small sample size, short writing tasks, and reliance on quantitative indices that do not fully capture contextual appropriateness. Future research should incorporate larger samples, qualitative analyses of lexical decision-making, longer writing tasks to better understand how AI tools can support lexical development over time.

Acknowledgment

The authors express their profound gratitude to all individuals who played a role in the research undertaking, including those involved in the collection and examination of data, as well as those who provided constructive feedback.

Declarations

- Author contribution** : Pandu Prasodjo conceptualized the study, designed the research, collected the data, conducted the analysis, and drafted the manuscript. Siusana Kweldju, Nur Mukminatien, Francisca Maria Ivone, and Hilarius Raditya Priambada Purba contributed to the refinement of the research design, provided theoretical and methodological guidance, and critically revised and finalized the manuscript. All authors reviewed and approved the final version of the article.
- Funding statement** : This research was funded by the Lembaga Penelitian dan Pengabdian kepada Masyarakat (LPPM), Universitas Internasional Batam, under contract number 025/LPPM/KP-UIB/VI/2025. The funder had no role in the study design, data collection, analysis, interpretation, or decision to submit the article for publication.
- Conflict of interest** : The authors declare that there are no conflicts of interest, e.g. financial, academic, or personal related to the conduct or reporting of this research.
- Ethical declaration** : We as authors acknowledge that this work has been written based on ethical research that conforms with the regulations of our university and

that we have obtained the permission from the relevant institute when collecting data.

We support ELTEJ in maintaining high standards of personal conduct, practicing honesty in all our professional practices and endeavors

Additional information : No additional data, materials, or supplementary documents are associated with this manuscript. All procedures followed ethical guidelines for research involving human participants, including informed consent and confidentiality of data.

REFERENCES

- Alharbi, W. (2023). AI in the foreign language classroom: A pedagogical overview of automated writing assistance tools. *Education Research International*, 2023. <https://doi.org/10.1155/2023/4253331>
- Ali, B. F. (2024). The effectiveness of two feedback strategies for teaching vocabulary among EFL students. *Journal of Language Studies*, 8(10), 184–202. <https://doi.org/10.25130/Lang.8.10.10>
- Aulia, V. (2018). Reducing students' pronunciation anxiety by utilizing WhatsApp group in learning English consonant sounds. *ELT Worldwide: Journal of English Language Teaching*, 5(2), 165–178. <https://doi.org/10.26858/eltww.v5i2.7378>
- Barrot, J. S. (2024). Leveraging Google Gemini as a research writing tool in higher education. *Technology, Knowledge and Learning*, August. <https://doi.org/10.1007/s10758-024-09774-x>
- Bedrio, J., Prasodjo, P., Raditya, H., & Purba, P. (2025). Free writing and ChatGPT to enhance 11th-grade idea development in letter writing. *Journal of English Language and Education*, 10, 2025. <https://doi.org/10.31004/jele.v10i4.1154>
- Bekele, G., Olana, T., & Ali, S. (2022). Effect of critical thinking-infused paragraph writing instruction on university first-year students' argumentative paragraphs writing performance. *East African Journal of Education Studies*, 5(1), 170–181. <https://doi.org/10.37284/eajes.5.1.624>
- Cahyono, B. Y., & Widiati, U. (2006). The teaching of EFL reading in the Indonesian context: The state of the art. *The Teaching of EFL Reading in the Indonesian Context: The State of the Art*, 17(1), 36–58. <https://doi.org/10.15639/teflinjournal.v17i1/37-60>
- Crompton, H., Edmett, A., Ichaporia, N., & Burke, D. (2024). AI and English language teaching: affordances and challenges. *British Journal of Educational Technology*, March, 1–27. <https://doi.org/10.1111/bjet.13460>
- Crossley, S. A., Salsbury, T., & Mcnamara, D. S. (2014). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics*, amt056. <https://doi.org/10.1093/applin/amt056>
- Da Silva, G. S., & Ulbricht, V. R. (2024). Learning with conversational AI: ChatGPT and Bard/Gemini in education. In *Artificial Intelligence for Supporting Human Cognition and Exploratory Learning in Digital Age* (pp. 101–117). https://doi.org/10.1007/978-3-031-66462-5_6
- Darwin, Rusdin, D., Mukminatien, N., Suryati, N., Laksmi, E. D., & Marzuki. (2024). Critical thinking in the AI era: An exploration of EFL students' perceptions, benefits, and limitations. *Cogent Education*, 11(1). <https://doi.org/10.1080/2331186X.2023.2290342>
- Dewaele, J.-M. (2008). "Appropriateness" in foreign language acquisition and use: Some theoretical, methodological and ethical considerations. *IRAL - International Review of Applied Linguistics in Language Teaching*, 46(3), 245–265. <https://doi.org/10.1515/iral.2008.011>
- Durrant, P., & Durrant, A. (2022). Appropriateness as an aspect of lexical richness: What do quantitative measures tell us about children's writing? *Assessing Writing*, 51(November 2021), 100596. <https://doi.org/10.1016/j.asw.2021.100596>

- Gemini Team, Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T., Lazaridou, A., ... Vinyals, O. (2023). Gemini: A family of highly capable multimodal models. 1–90. <http://arxiv.org/abs/2312.11805>
- Gemma Team, Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., Tafti, P., Hussenot, L., Sessa, P. G., Chowdhery, A., Roberts, A., Barua, A., Botev, A., Castro-Ros, A., Slone, A., ... Kenealy, K. (2024). Gemma: Open models based on Gemini research and technology. <http://arxiv.org/abs/2403.08295>
- Hartono, D. A., & Prima, S. A. B. (2021). The correlation between Indonesian university students' receptive vocabulary knowledge and their reading comprehension level. *Indonesian Journal of Applied Linguistics*, 11(1), 21–29. <https://doi.org/10.17509/ijal.v11i1.34590>
- Hasselgård, H. (2019). Phraseological teddy bears: Frequent lexical bundles in academic writing by Norwegian learners and native speakers of English. *Corpus Linguistics, Context and Culture*, 1994, 339–362. <https://doi.org/10.1515/9783110489071-013>
- Hyland, K. (2003). On second language writing. In *Cambridge University Press* (Vol. 36, Issue 4). <https://doi.org/10.2307/3588251>
- Javaid, M., Haleem, A., Singh, R. P., Khan, S., & Khan, I. H. (2023). Unlocking the opportunities through ChatGPT tool towards ameliorating the education system. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 3(2), 100115. <https://doi.org/10.1016/j.tbench.2023.100115>
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103(March). <https://doi.org/10.1016/j.lindif.2023.102274>
- Komara, C., & Tiarsiwi, F. (2021). Exploring Indonesian EFL learners' perception of English learning grammar. *Journal of English Language Teaching and Linguistics*, 6(2), 459. <https://doi.org/10.21462/jeltl.v6i2.564>
- Kweldju, S. (2004). Lexically-based language teaching an innovative step for ELT in Indonesia. In *The Tapestry of English Language Teaching and Learning in Indonesia*.
- Laqaei, N. (2015). The impact of reflective writing on writing achievement, vocabulary achievement and critical thinking of intermediate EFL learners. *Journal of Studies in Education*, 5(3), 174. <https://doi.org/10.5296/jse.v5i3.7964>
- Lauffer, B., & Nation, P. (1995). Vocabulary size and use: lexical richness in L2 written production. *Applied Linguistics*, 16(3), 307–322. <https://doi.org/10.1093/applin/16.3.307>
- Lee, J. H. (2025). ChatGPT-assisted writing: learners' perceptions of AI feedback and influencing factors. *Korean Journal of English Language and Linguistics*, 25(April), 493–515. <https://doi.org/10.15738/kjell.25..202504.493>
- Lee, S. H. (2003). ESL learners' vocabulary use in writing and the effects of explicit vocabulary instruction. *System*, 31(4), 537–561. <https://doi.org/10.1016/j.system.2003.02.004>
- Li, H., & Iwashita, N. (2019). The role of recasts and negotiated prompts in an FL learning context in China with non-English major university students. *Language Teaching Research*, 25(2), 209–233. <https://doi.org/10.1177/1362168819839727>
- Limantoro, S. W. (2018). Developing word-card games to improve English writing. *PUPIL: International Journal of Teaching, Education and Learning*, 2(3), 38–54. <https://doi.org/10.20319/pijtel.2018.23.3854>
- Maeda, H. (2024). Field-testing multiple-choice questions with AI examinees: English Grammar Items. *Educational and Psychological Measurement*. <https://doi.org/10.1177/00131644241281053>

- Mahajan, P. (2020). Emergin need to rethink English language pedagogies in Indian technical education: Creative pedagogy a better approach for vocabulary instruction. *Humanities & Social Sciences Reviews*, 8(1), 829–836. <https://doi.org/10.18510/hssr.2020.8199>
- Moybeka, A. M. S., Syariatn, N., Tatipang, D. P., Mushthoza, D. A., Dewi, N. P. J. L., & Tineh, S. (2023). Artificial Intelligence and English classroom: The implications of AI toward EFL students' motivation. *Edumaspul: Jurnal Pendidikan*, 7(2), 2444–2454. <https://doi.org/10.33487/edumaspul.v7i2.6669>
- Nagy, W. (2007). Metalinguistic Awareness and the vocabulary-comprehension connection. In *Vocabulary Acquisition* (pp. 52–77). <https://doi.org/10.4324/9781410614933-15>
- Nazim, M. (2024). Exploring efl teachers' insights regarding artificial intelligence driven tools in student-centered writing instructions. *International Journal of English Linguistics*, 14(3), 90. <https://doi.org/10.5539/ijel.v14n3p90>
- Novianti, R. (2016). A study of Indonesian university students' vocabulary mastery with vocabulary level test. *Global Journal of Foreign Language Teaching*, 06(4), 187–195. <https://doi.org/10.18844/gjflt.v6i4.1669>
- Nurweni, A., & Read, J. (1999). The English vocabulary knowledge of indonesian university students. *English for Specific Purposes*, 18(2), 161–175. [https://doi.org/10.1016/S0889-4906\(98\)00005-2](https://doi.org/10.1016/S0889-4906(98)00005-2)
- Phuong, N., & Anh, D. (2024). Gemini Google: a potential tool for English learning. *Journal of Thu Dau Mot University*, 1, 386–396. <https://doi.org/10.37550/tdmu.ejs/2024.03.586>
- Prasodjo, P., Kweldju, S., Mukminatien, N., & Ivone, F. M. (2025). Unleashing Gemini: Enhancing EFL lexical collocation and writing style through lexically-based language teaching model. *International Journal of Religion*, 6(1), 50–67. <https://doi.org/10.61707/hb2kzm10>
- Rahman, M. M. (2022). The essentials of paragraph writing to develop writing skill. *Global Journal of Human-Social Science*, 22(5), 71–75. <https://doi.org/10.34257/gjhssgvol22is5pg71>
- Rahmat, Y. N., & Coxhead, A. (2021). Investigating vocabulary coverage and load in an Indonesian EFL textbook series. *Indonesian Journal of Applied Linguistics*, 10(3), 804–814. <https://doi.org/10.17509/ijal.v10i3.31768>
- Rohim, A. (2019). An analysis of students writing skill in paragraph writing. *Globish: An English-Indonesian Journal for English, Education, and Culture*, 8(1). <https://doi.org/10.31000/globish.v7i2.1126>
- Saadi, Z. K., & Saadat, M. (2015). EFL learners' writing accuracy: Effects of direct and metalinguistic electronic feedback. *Theory and Practice in Language Studies*, 5(10), 2053–2063. <https://doi.org/10.17507/tpls.0510.11>
- Schaefer, S. (2023). Generating and developing ideas. In *BT - Generating and Developing Ideas* (pp. 51–85). Oxford University Press. <https://doi.org/10.1093/oso/9780198893509.003.0003>
- Shadish, W. R., & Luellen, J. K. (2006). Quasi-Experimentasl design. In J. L. Green, G. Carmilli, & P. B. Elmore (Eds.), *Handbook of Complementary Methods in Education Research* (3rd Editio). Taylor & Francis Group.
- Shidiq, M. (2023, May). The use of artificial intelligence-based Chat-GPT and its challenges for the world of education; from the viewpoint of the development of creative writing skills. In *Proceeding of international conference on education, society and humanity* (Vol. 1, No. 1, pp. 353-357).
- Songbatumis, A. M. (2017). Challenges in teaching English faced by English teachers at MTsN Taliwang, Indonesia. *Journal of Foreign Language Teaching and Learning*, 2(2), 54–67. <https://doi.org/10.18196/ftl.2223>
- Strobl, C., Menke-Bazhutkina, I., Abel, N., & Michel, M. (2024). Adopting ChatGPT as a writing buddy in the advanced L2 writing class. *Technology in Language Teaching & Learning*, 6(1), 1–19. <https://doi.org/10.29140/tlftl.v6n1.1168>
- Tustiawati, I. A. M., & Wiguna, I. B. S. A. (2024). University students' perspectives on difficulties in paragraph writing and their strategies to overcome the difficulties. *Proceedings : Linguistics*,

- Literature, Culture, and Arts International Seminar (LITERATES)*, 300–309.
<https://jurnal.uniraya.ac.id/index.php/faguru/article/download/496/419>
- Usman, S. (2018). The effect of achievement motivation on students' learning outcomes in writing English descriptive paragraph through individual and group work method. *Proceedings of the First Indonesian Communication Forum of Teacher Training and Education Faculty Leaders International Conference on Education 2017 (ICE 2017)*, 174(Ice 2017), 15–18.
<https://doi.org/10.2991/ice-17.2018.5>
- Vita Losi, R., Putra, E., Ali, N., & Silvana Dewi, A. (2024). Investigating artificial intelligence (AI) as a vocabulary learning tool: Students' perception to use ChatGPT. *North Sumatra Jl. Jend. Gatot Subroto KM, 3*, 561–566.
- Wali, O., & Madani, A. Q. (2020). The importance of paragraph writing: An introduction. *International Journal of Latest Research in Humanities and Social Science (IJLRHSS)*, 3(07), 44–50.
- Wau, C. S. W. (2022). Students' difficulties in writing definition paragraph at the third semester students of English Language Education study program of STKIP Nias Selatan. *FAGURU: Jurnal Ilmiah Mahasiswa Keguruan*, 1(1).
<https://jurnal.uniraya.ac.id/index.php/faguru/article/download/496/419>
- Wirantaka, A. (2016). Paragraph writing of academic texts in an EFL context. *Journal of Foreign Language Teaching and Learning*, 1(2), 34–45.
- Zhai, C., Wibowo, S., & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments*, 11(1).
<https://doi.org/10.1186/s40561-024-00316-7>