Edcomtech: Jurnal Kajian Teknologi Pendidikan
Vol 11, No 1, April 2026, pp. 170-184
https://doi.org/10.17977/um039v11i12026p170-184
pISSN: 2548-9879 (print) eISSN: 2599-2139 (online)
http://journal-fip.um.ac.id/index.php/edcomtech

ChatGPT as a Macro-Level Cognitive Scaffold for First-Language Argumentative Writing: A Quasi-Experimental Study in a Madrasah Context

Jumaddil[1*], Abdul Rani[2], Dyah Werdiningsing[3]
[1 2 3]Universitas Islam Malang, Kota Malang, Jawa Timur, Indonesia
**Email corresponding author: Jumaddilid@gmail.com*

Article Info

Article history:
Received 08-04-2026
Revised 25-04-2026
Accepted 27-04-2026
Published 30-04-2026

How to cite: Jumaddil, Rani, A., & Werdiningsih, D. (2026). ChatGPT as a Macro-Level Cognitive Scaffold for L1 Argumentative Writing: A Quasi-Experimental Madrasah Study. *Edcomtech: Jurnal Kajian Teknologi Pendidikan, 11*(1), 170–184. https://doi.org/10.17977/um039v11i12026p170-184

Abstrak

Bukti empiris mengenai penggunaan kecerdasan buatan generatif dalam pembelajaran menulis bahasa pertama (L1) di sekolah menengah Indonesia masih terbatas, dan sedikit penelitian yang mengkaji kompetensi menulis spesifik mana yang paling memperoleh manfaat dari intervensi tersebut. Penelitian ini mengisi kesenjangan tersebut dengan memposisikan ChatGPT bukan sekadar sebagai alat penghasil teks, melainkan sebagai scaffolding dialogis yang mendukung peserta didik dalam zona perkembangan proksimal mereka. Penelitian menggunakan desain kuasi eksperimen nonequivalent pretest-posttest control group dengan melibatkan enam puluh siswa kelas XI di Madrasah Aliyah Negeri (MAN) 2 Parigi, Sulawesi Tengah, yang diajar oleh guru yang sama selama delapan kali pertemuan. Kemampuan menulis argumentatif diukur menggunakan rubrik analitik yang diadaptasi dan divalidasi untuk konteks bahasa Indonesia sebagai L1. Hasil penelitian menunjukkan bahwa pembelajaran berbantuan ChatGPT menghasilkan peningkatan yang besar terhadap kemampuan menulis argumentatif secara keseluruhan (Cohen's d = 1,30). Namun, setelah dilakukan koreksi Holm–Bonferroni untuk perbandingan multipel, pengaruh signifikan hanya ditemukan pada dimensi makro, yaitu organisasi dan isi, sedangkan kosakata, penggunaan bahasa, dan mekanik tidak mempertahankan signifikansi statistik. Penelitian ini menawarkan perspektif dimensional-specificity yang menunjukkan bahwa manfaat pedagogis kecerdasan buatan generatif dalam pembelajaran menulis tidak terdistribusi secara merata pada seluruh komponen menulis, melainkan terkonsentrasi pada tingkat makrostruktur argumentatif. Temuan ini memperluas teori scaffolding Vygotsky ke dalam konteks interaksi manusia–AI serta memperdalam pemahaman mengenai bagaimana kecerdasan buatan generatif mendukung perkembangan menulis L1 pada konteks madrasah dan wilayah pendidikan yang masih kurang terjangkau penelitian.

Kata Kunci: *Manajemen Pengetahuan; Organisasi Belajar; Lembaga Pelatihan.*

Abstract

Empirical evidence on the use of generative AI for first-language (L1) writing instruction in Indonesian secondary schools remains limited, and few studies have examined which specific writing competencies

benefit most from such interventions. This study addresses this gap by positioning ChatGPT not merely as a text-generating tool, but as a dialogic scaffold that supports learners within their zone of proximal development. A quasi-experimental nonequivalent pretest-posttest control group design was conducted with sixty eleventh-grade students at Madrasah Aliyah Negeri (MAN) 2 Parigi, Central Sulawesi, Indonesia, taught by the same teacher across eight instructional meetings. Argumentative writing performance was assessed using an analytical rubric adapted and validated for L1 Indonesian contexts. The findings showed that ChatGPT-assisted instruction produced a large overall effect on argumentative writing achievement (Cohen's d = 1.30). However, after Holm-Bonferroni correction for multiple comparisons, significant effects remained only in the macro-level dimensions of organisation and content, whereas vocabulary, language use, and mechanics did not retain statistical significance. The study contributes a dimensional-specificity perspective, suggesting that the pedagogical benefits of generative AI in writing instruction are not uniformly distributed across writing components but are concentrated at the level of argumentative macrostructure. These findings extend Vygotskian scaffolding theory into human-AI interaction and deepen current understanding of how generative AI supports L1 writing development in under-researched madrasah and geographically disadvantaged educational contexts.

Keywords: *ChatGPT; Argumentative Writing; Indonesian Language; Quasi-experimental; Madrasah.*

INTRODUCTION

Writing argumentative essays is one of the most cognitively demanding competencies in secondary education. It requires integrating critical thinking, logical reasoning, and sophisticated linguistic resources (Ekalia et al., 2025; OECD, 2023). In Indonesia, empirical evidence consistently shows that students' argumentative writing in Bahasa Indonesia remains below expected standards. The Program for International Student Assessment (PISA) 2022 reported a reading literacy score of 359, placing Indonesia 71st of 81 participating countries (Kemendikbudristek, 2023; OECD, 2023). Argumentative text production was among the weakest subcomponents. This problem is particularly pronounced in Islamic senior secondary schools (Madrasah Aliyah Negeri/MAN), especially in geographically disadvantaged regions such as Parigi Moutong Regency, where access to pedagogical innovation is limited (Siregar et al., 2025; Zh et al., 2024a).

Prior research has attempted to address this problem through various interventions. Process-genre approaches have yielded moderate improvements (Huang & Zhang, 2020) but offer limited individualized feedback. Digital-based interventions such as interactive e-books and Wordwall gamification have improved engagement and vocabulary (Zh et al., 2024b), though they primarily target lower-order skills. The release of ChatGPT by OpenAI has since attracted substantial scholarly attention. Global studies report significant positive effects of ChatGPT on the quality of academic writing (Mahapatra, 2024; Song & Song, 2023). Quasi-experimental investigations confirm effects on argumentative writing in higher education settings (Darmawansah et al., 2025; Li et al., 2024; Yasmin et al., 2025).

In the Indonesian context, a growing body of empirical work has begun to examine the use of ChatGPT in writing instruction. Apriani et al. (2025) conducted a mixed-methods

experiment with 50 EFL university students in Indonesia. They reported significantly higher posttest scores for the ChatGPT group (M = 81.11) compared to the control group (M = 60.30). Nugroho et al. (2024) documented Indonesian EFL students' generally positive appraisals of ChatGPT for academic writing, while identifying concerns about academic dishonesty. Anam et al. (2025) examined senior high school English teachers' perspectives and found both recognition of ChatGPT's pedagogical potential and concerns about student overdependence. Utami et al. (2023) mapped Indonesian senior high school students' perceptions of AI writing tools and identified EYD-specific limitations in generative models.

Despite these contributions, three important gaps remain. First, existing Indonesian empirical work focuses predominantly on English as a foreign language (EFL); rigorous experimental studies on ChatGPT for first-language (L1) writing in Bahasa Indonesia are scarce. Second, the institutional contexts studied are largely secular universities or general secondary schools; Islamic secondary schools (madrasah), which serve approximately 15% of Indonesian secondary students, have received minimal empirical attention (Siregar et al., 2025; Zh et al., 2025). Third, most studies originate from metropolitan regions, producing limited insight into how generative AI functions in geographically disadvantaged eastern Indonesia.

Addressing these gaps is both timely and consequential. The present study conceptualizes ChatGPT-assisted writing as a pedagogically scaffolded intervention. The AI functions as a dialogic tutor throughout ideation, drafting, and revision, rather than as a text-generating substitute for student authorship. The framework draws on Vygotsky's (1978) sociocultural theory and its contemporary reinterpretation as human-AI collaborative scaffolding within the zone of proximal development (Rad & Mirzaei, 2024). The study aims to (1) examine the effect of ChatGPT-assisted writing instruction on argumentative essay skills, (2) determine the magnitude of this effect, and (3) identify the dimensions of argumentative writing most responsive to the intervention.

METHOD

Research Design

This study employed a quasi-experimental design with a nonequivalent pretest-posttest control group structure (Creswell & Creswell, 2023). Random assignment of individual students was not feasible within the intact Classroom structure of MAN 2 Parigi. Two parallel classes were therefore assigned as experimental and control groups. The null hypothesis (H_o) stated that there would be no significant difference in posttest argumentative writing scores between the two groups. The alternative hypothesis (H_1) predicted superior posttest performance in the experimental group.

Threats to Internal Validity and Mitigation Strategies

Following Shadish et al.'s (2002) framework, we explicitly identified and addressed four principal threats to internal validity. First, to minimize selection bias from non-random assignment, we compared the two classes on three baseline characteristics: prior semester Indonesian language grades ($M_{exp} = 76.4$, $M_{ctrl} = 77.1$, $p = .48$), age distribution (both M = 16.3 years), and gender ratios. No statistically significant differences emerged. A coin toss determined class assignment as experimental or control after baseline equivalence had been confirmed.

Second, to control instructor-related confounds, the same teacher (the first author) taught both classes using carefully scripted lesson plans. The key difference between the

conditions was the use of ChatGPT as a scaffolding tool. Both conditions covered the same content and allocated equivalent instructional time to ideation, drafting, and revision phases.

Third, to reduce Hawthorne effects, the control condition was not framed as a passive comparison. Rather, students in the control group received structured peer-review scaffolding within a process-writing framework. Both groups were informed that they were participating in an innovative instructional approach being piloted by a graduate researcher. A post-intervention manipulation check asked students about the perceived novelty of their instruction; responses did not differ significantly between groups ($p = .73$), indicating comparable engagement with the perceived novelty.

Fourth, to address potential contamination from external exposure to ChatGPT, all participants completed a pre-intervention questionnaire regarding prior use of AI writing tools. Only students without systematic prior exposure were retained. During the four-week intervention, students in both groups signed a commitment to refrain from using AI tools on individual assignments. Compliance was verified through a post-intervention interview. These measures cannot fully exclude outside exposure, but they reduce its likelihood.

Participants and Setting

The participants comprised 60 eleventh-grade (Kelas XI) students at MAN 2 Parigi, Parigi Moutong Regency, Central Sulawesi Province, during the 2025/2026 academic year. Purposive sampling was used with three inclusion criteria: (1) enrollment in the Social Sciences (IPS) program, (2) active participation in Indonesian language classes, and (3) no systematic prior exposure to ChatGPT for academic writing as verified by the pre-intervention questionnaire. Students were assigned to an experimental group (XI IPS 1, $n = 30$; 17 male, 13 female) or a control group (XI IPS 2, $n = 30$; 16 male, 14 female). The age range was 16 to 17 years.

Instrument and Rubric Adaptation

The primary instrument was an argumentative essay writing test administered at pretest and posttest, with two topically equivalent prompts addressing contemporary socio-educational issues. Student essays were scored using an analytical rubric adapted from Jacobs et al. (1981) with five dimensions: content (0 to 30), organization (0 to 20), vocabulary (0 to 20), language use (0 to 25), and mechanics (0 to 5). Because the original rubric was designed for English-as-a-Second-Language contexts, two key adaptations were made for L1 Indonesian writing. The language use dimension was reformulated to evaluate Indonesian grammatical and syntactic features rather than English structures, with specific indicators for subject-predicate agreement, the use of coherent conjunctions (e.g., sehingga, namun, oleh karena itu), and appropriate sentence variation. The mechanics dimension was aligned with the Ejaan Bahasa Indonesia yang Disempurnakan (EYD) conventions, including capitalization of proper nouns, hyphenation of prefixes, punctuation of direct quotations, and spelling of loanwords.

The adapted rubric was validated by three experts (two Indonesian language lecturers at Universitas Islam Malang and one AI-in-education specialist), yielding a content validity index (CVI) of 0.89. Convergent validity was examined through the correlation between the analytical rubric total and an independent holistic rating (1- to 6-point scale) from a fourth rater. The Pearson correlation was $r = .78$ ($p < .001$), indicating acceptable convergent validity. Inter-rater reliability between two independent raters produced Cohen's Kappa = .81, representing substantial agreement (Landis & Koch, 1977). This value is acceptable for analytical rubrics, but below the .90 threshold often desired for high-stakes assessment. To

mitigate residual reliability concerns, all disagreements were resolved through a third-rater adjudication procedure.

Prompt Equivalence Pilot Test

To verify that the two essay prompts were of comparable difficulty, a within-subjects pilot test was conducted with 15 non-sample students from the same school. Each student wrote essays in response to both prompts in a counterbalanced order, with a one-week interval between prompts. Essays were scored using the adapted rubric. A paired samples t-test revealed no significant difference between prompts, $t(14) = -0.09$, $p = .932$ ($M_A = 57.83$, $SD = 4.22$; $M_B = 58.02$, $SD = 6.97$). Three Indonesian language teachers independently rated prompt difficulty on a five-point scale, producing closely comparable means ($M_A = 3.8$, $M_B = 3.9$). These results confirmed the equivalence of the prompts for counterbalanced assignment to the pretest and posttest conditions across the two classes.

Treatment Procedure

The intervention was conducted over eight 90-minute meetings across four weeks. In the experimental group, Meeting 1 comprised the pretest and an introduction to responsible AI use and prompt engineering. Meetings 2 and 3 focused on brainstorming arguments, counter-arguments, and evidence using ChatGPT as an ideation partner. Meetings 4 and 5 involved drafting with ChatGPT functioning as a dialogic feedback provider on structure and reasoning. Meetings 6 and 7 emphasized revising and editing through iterative AI feedback combined with peer review. Meeting 8 concluded with the posttest.

Scaffolding was differentiated based on students' baseline pretest scores. Students in the lower quartile received more structured prompt templates (e.g., "Ask ChatGPT: What are three weaknesses in the logical flow of my paragraph?"). Students in the upper quartile received open-ended prompt options to promote higher-order reasoning. This differentiation operationalized the zone of proximal development by calibrating the scaffolding intensity to each student's current ability level (Rad & Mirzaei, 2024). Throughout all sessions, students were required to author every sentence themselves. ChatGPT's role was strictly limited to providing questions, critiques, and suggestions, in line with the AI-as-tutor principle (Ghafouri et al., 2024). The control group received instruction on the same content for the same duration through structured peer-review scaffolding within a process-writing framework, without AI assistance.

Ethical Considerations and Academic Integrity

Ethical clearance for this study was obtained from the Research Ethics Committee of the Graduate Program, Universitas Islam Malang. Because participants were minors aged 16 to 17, written informed consent was obtained from parents or legal guardians, and written informed assent was obtained from each student. Students were informed of their right to withdraw at any time without academic consequence. Data were anonymized using student identification codes (E01 through E30 for the experimental group; C01 through C30 for the control group).

Two academic integrity safeguards were implemented. First, all pretest and posttest essays were checked for textual similarity using Turnitin; the average similarity index was 8.4% ($SD = 3.1$), well below the 20% threshold. Second, all experimental group essays were analyzed with GPTZero to detect AI-generated content; the mean AI probability score was 11.7% ($SD = 6.8$), indicating predominantly human authorship. Students whose essays exceeded 30% AI

probability (n = 2) were retained in the analysis, but their data were flagged for sensitivity checking. Rerunning the main analyses excluding these two cases did not change the pattern of findings.

Data Analysis

Assumption checks included normality (Shapiro-Wilk test and Q-Q plots), homogeneity of variance (Levene's test), and homogeneity of regression slopes for ANCOVA. Hypothesis testing proceeded through four complementary analyses: (1) independent samples t-test comparing posttest means, (2) analysis of covariance (ANCOVA) with pretest as covariate, (3) normalized gain score (Hake, 1998), and (4) Cohen's d effect size with Hedges' g small-sample correction (Cohen, 1988). Dimensional analyses on the five rubric components were performed using separate t-tests with Holm-Bonferroni step-down correction to control the familywise error rate at $\alpha = .05$. All analyses were conducted in JASP 0.18 and cross-verified in SPSS 27.

RESULTS

Assumption Testing

The Shapiro-Wilk test indicated normal distribution across all four datasets: experimental pretest ($W = .972$, $p = .589$), experimental posttest ($W = .982$, $p = .874$), control pretest ($W = .981$, $p = .860$), and control posttest ($W = .983$, $p = .888$). Q-Q plots supporting these conclusions are presented in Figure 3. Levene's test demonstrated homogeneity of variance for both pretest ($F = 0.44$, $p = .512$) and posttest ($F = 0.49$, $p = .486$). The homogeneity of regression slopes assumption for ANCOVA was satisfied (Group x Pretest interaction: $F = 0.04$, $p = .843$). These results justified the use of parametric statistical tests.

Baseline Equivalence

An independent-samples t-test on pretest scores yielded $t(58) = -1.86$, $p = .069$. The difference did not reach statistical significance, confirming the comparability of the groups prior to the intervention. To further control for the minor numeric baseline discrepancy ($M_{exp} = 61.07$; $M_{ctrl} = 63.97$), ANCOVA with pretest as covariate was subsequently employed.

Descriptive Statistics

Descriptive statistics for both groups are presented in Table 1. Figure 1 displays the corresponding score distributions through boxplots with individual data points.

Table 1. Descriptive Statistics of Argumentative Writing Total Scores

Group	N	Pre M	Pre SD	Post M	Post SD	Gain M	Range Post
Experimental	30	61.07	6.30	81.67	5.23	20.60	71 to 93
Control	30	63.97	5.80	74.23	6.13	10.27	62 to 87

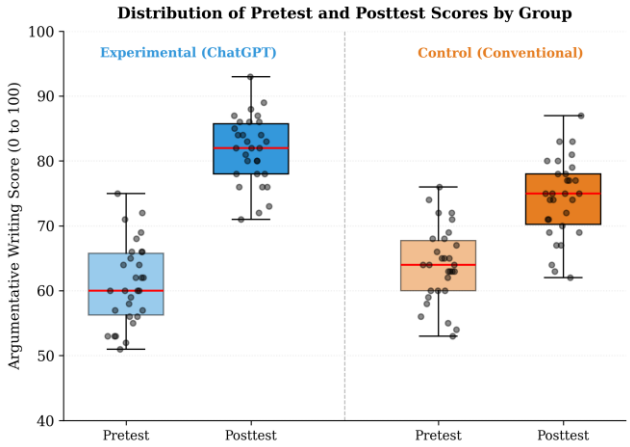

Figure 1. Distribution of Pretest and Posttest Scores by Group, Showing Individual Data Points and Quartile Distributions

Table 1 and Figure 1 together show that the groups started from comparable baselines, then diverged substantially at posttest. The experimental group gained 20.60 points, more than twice the control group's 10.27-point gain. Both posttest distributions were approximately symmetrical and free from ceiling or floor effects.

Independent Samples t-Test

Posttest scores differed significantly between groups, $t(58) = 5.05$, $p < .001$. The mean difference of 7.44 points favored the experimental group (95% CI [4.49, 10.39])—the confidence interval excluded zero, reinforcing the finding's robustness.

Analysis of Covariance (ANCOVA)

After controlling for pretest scores as a covariate, ANCOVA confirmed the treatment effect, $F(1, 57) = 62.78$, $p < .001$, partial $\eta^2 = .52$. The pretest covariate was also a significant predictor, $F(1, 57) = 41.56$, $p < .001$. The adjusted posttest means were M_adj = 82.55 for the experimental group and M_adj = 73.35 for the control group, yielding an adjusted difference of 9.20 points. Results are summarized in Table 2.

Table 2. ANCOVA Results for Posttest Argumentative Writing Scores

Source	Sum of Squares	df	F	p	Partial η^2
Group (Treatment)	1200.02	1	62.78	< .001	.524
Pretest (Covariate)	794.49	1	41.56	< .001	.422
Residual	1089.55	57			

The treatment factor accounted for 52% of the variance in posttest scores, a very large effect (Cohen, 1988).

Effect Size Analysis

Cohen's d for posttest scores was 1.30 (95% CI [0.75, 1.86]), exceeding the threshold for a large effect. Hedges' g, which corrects for small-sample bias, was 1.29. The lower bound of the 95% confidence interval (0.75) remains in the medium-to-large range, indicating that even the most conservative estimate is educationally meaningful.

Normalized Gain (N-Gain) Analysis

The experimental group achieved a mean N-gain of 0.53, classified as moderate, while the control group achieved 0.28, classified as low. The distribution of individual N-gain categories is reported in Table 3.

Table 3. Distribution of Individual N-Gain Categories

Group	Mean N-Gain	High (>0.70)	Moderate (0.30 to 0.70)	Low (<0.30)	Classification
Experimental	0.53	2 (6.7%)	28 (93.3%)	0 (0.0%)	Moderate
Control	0.28	0 (0.0%)	15 (50.0%)	15 (50.0%)	Low

In the experimental group, all students achieved at least moderate learning gains; two students (6.7%) reached the high category. In the control group, half of the students achieved only low gains, and none reached the high category. The intervention was therefore more consistent in producing meaningful gains, not only higher on average.

Dimensional Analysis with Holm-Bonferroni Correction

Separate independent samples t-tests were computed for each of the five rubric dimensions. Because five tests were conducted on the same dataset, the Holm-Bonferroni step-down procedure was applied to control the familywise error rate at $\alpha = .05$. Results are presented in Table 4 and visualized in Figure 2.

Table 4. Dimensional Analysis with Holm-Bonferroni Correction

Dimension	Max	Exp Pre	Exp Post	Ctrl Pre	Ctrl Post	t	p (raw)	p (Holm)	d
Organization	20	11.30	17.10	12.27	14.27	4.31	.0001	.0005*	1.11
Content	30	17.90	24.77	18.43	22.10	2.88	.006	.022*	0.74
Vocabulary	20	12.87	16.03	13.53	14.90	2.15	.036	.107	0.56
Language Use	25	15.53	19.50	16.00	18.63	1.27	.208	.415	0.33
Mechanics	5	3.47	4.27	3.73	4.33	−0.34	.739	.739	−0.09

Note. * = significant after Holm-Bonferroni correction at familywise $\alpha = .05$.

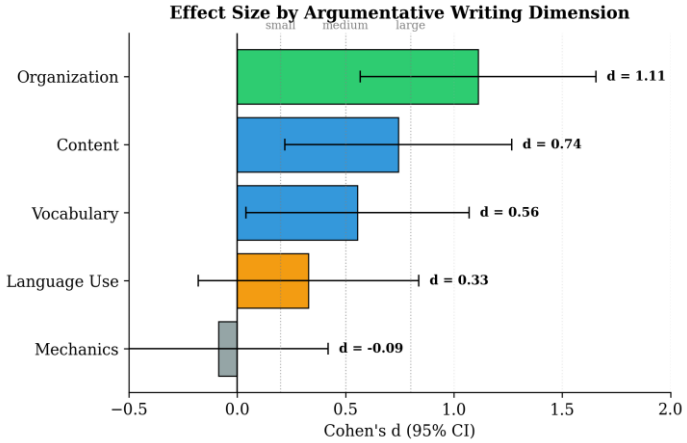

Figure 2. Cohen's d Effect Sizes per Rubric Dimension With 95% Confidence Intervals. Dotted Vertical Lines Mark Small (d = 0.2), Medium (d = 0.5), and Large (d = 0.8) Thresholds (Cohen, 1988)

The dimensional analysis reveals a clear hierarchical pattern. Before correction, four dimensions (organization, content, vocabulary, language use) showed meaningful effects. After Holm-Bonferroni correction, only organization (p_holm = .0005) and content (p_holm = .022) remained statistically significant. The effects on vocabulary (p_holm = .107), language use (p_holm = .415), and mechanics (p_holm = .739) did not survive correction. The correction substantially narrows the claims this study can make: the intervention's replicable benefit lies specifically in the macro-level structural and substantive features of argumentation. Claims about effects on lexical richness or grammatical form should be treated as exploratory findings warranting replication.

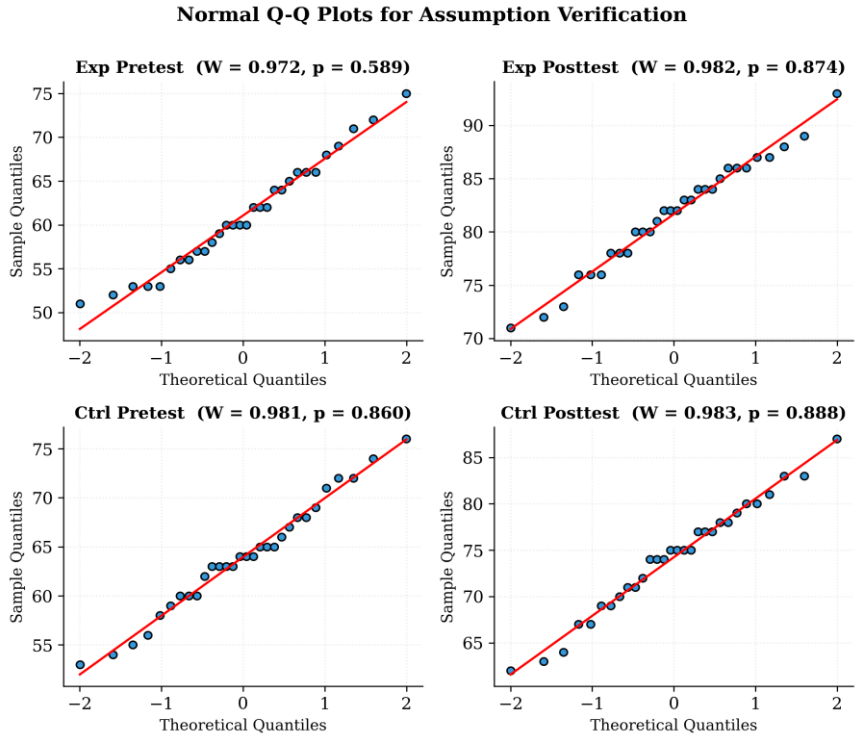

Figure 3. Normal Q–Q Plots for the Four Distributions Used in Hypothesis Testing, Supporting the Normality Assumption Reported in the Assumption Testing Section

Summary of Findings

Four converging analyses (t-test, ANCOVA, Cohen's d, and N-gain) indicate that ChatGPT-assisted writing instruction produced a substantial improvement in overall argumentative writing. Dimensional analysis with Holm-Bonferroni correction, however, localizes this effect to the macro-level dimensions of organization and content. The effect did not extend robustly to vocabulary, language use, or mechanics. Notably, the mechanics dimension showed a slight negative direction (d = −0.09), meriting interpretive attention in the discussion.

DISCUSSION

Principal Findings

ChatGPT-assisted writing instruction improved students' overall argumentative writing, with a large Cohen's d of 1.30 and a very large partial η^2 of .52. These effects align with recent experimental evidence from higher education (Li et al., 2024; Mahapatra, 2024; Song & Song,

2023) and from Indonesian EFL university contexts (Apriani et al., 2025). The more informative finding, however, emerges from the dimensional analysis with Holm-Bonferroni correction. Only organization and content showed effects that survived correction for multiple comparisons. This specificity, rather than the aggregate effect, constitutes the main empirical contribution of this study. It indicates that ChatGPT's reliable pedagogical value lies in scaffolding argumentative macrostructure, not in refining linguistic form.

Comparison with Indonesian L1 and EFL Studies

The present study extends recent Indonesian empirical work in several ways. Apriani et al. (2025), working with Indonesian EFL university students, reported an experimental posttest mean of 81.11 against a control mean of 60.30, closely paralleling the present means of 81.67 and 74.23. The smaller gap in the present study likely reflects the stronger pedagogical scaffolding in the control condition, which included structured peer review rather than passive lecture. Nugroho et al. (2024) documented Indonesian students' positive appraisals of ChatGPT for writing but expressed concerns about academic dishonesty. The present study addresses these concerns operationally through Turnitin screening and GPTZero detection, offering a methodological model for future studies. Anam et al. (2025) reported Indonesian teachers' concerns about student overdependence on AI. The scaffolded design used here, in which ChatGPT may question but may not write, directly addresses this concern at the classroom level.

The present findings diverge from those of Bašić et al. (2023), who found no significant advantage of ChatGPT-assisted writing. Their design allowed students to use ChatGPT to generate prose directly, thereby reducing their cognitive engagement with the task. The present study's restriction of ChatGPT to tutor-like functions appears critical to the observed gains. The findings also contrast with those of Niloy et al. (2023), who reported adverse effects of ChatGPT on creative writing. Argumentative writing, which rewards logical structure rather than original voice, may be more amenable to AI scaffolding than creative genres.

Mechanisms: Why Organization and Content Benefited Most

The concentration of treatment effects in organization ($d = 1.11$) and content ($d = 0.74$) can be explained through three mechanisms. First, the scaffolded prompts operationalized Vygotsky's (1978) zone of proximal development at the micro-interactional level. Lower-quartile students used structured prompts that externalized the metacognitive moves of experienced writers, such as "What is my main claim?" "What evidence supports it?" "What counter-argument might a critic raise?" Upper-quartile students used open prompts that pushed them toward more sophisticated dialectical reasoning. This differentiation allowed the same tool to function within each student's ZPD rather than providing uniform scaffolding (Rad & Mirzaei, 2024). Second, the AI's capacity to quickly retrieve and reorganize information reduced extrinsic cognitive load, freeing working memory for the germane processing required to construct arguments. Third, the iterative feedback cycle created dialogic encounters between students and their own claims, consistent with the dialogic function of feedback documented by Banihashem et al. (2024).

Mechanism: The Modest Effect on Language Use

The effect on language use ($d = 0.33$) failed to reach significance after Holm-Bonferroni correction. Language use in the adapted rubric evaluates grammatical accuracy, sentence structure, and syntactic variety. These features depend on internalized linguistic competence

rather than externally scaffolded cognition. Because the intervention deliberately restricted ChatGPT from generating prose, students could not directly incorporate AI-produced grammatical patterns. Students received feedback on grammar and then had to apply it independently, a transfer demand that eight meetings may not have been sufficient to fulfill. Woo et al. (2024) similarly observed that the cognitive load of independent application may outweigh the benefit of AI-scaffolded grammar feedback within short-duration interventions.

Mechanism: The Negligible to Negative Effect on Mechanics

The mechanics dimension showed a slight negative effect ($d = -0.09$) that was not statistically significant but merits interpretation. Three explanations are plausible. First, students in the experimental group may have redirected cognitive resources toward argumentative ideation and away from surface-level editing, resulting in a micro-level attentional trade-off consistent with bounded cognitive load theory. Second, ChatGPT's Indonesian-language performance exhibits documented inconsistencies regarding EYD-specific conventions, including prefix hyphenation and loanword spelling (Utami et al., 2023). Students who adopted ChatGPT suggestions on mechanics may have introduced non-EYD-conformant corrections into their essays. Third, the observed negative direction is small in magnitude, and its confidence interval includes zero, indicating that the true effect could plausibly be null rather than detrimental. Regardless of the preferred interpretation, this finding warns against relying on ChatGPT for L1 Indonesian orthographic instruction.

Theoretical Contributions

This study makes three modest theoretical contributions. First, it provides empirical grounding for extending Vygotsky's (1978) zone of proximal development to human-AI interaction contexts, building on and sharpening the conceptual work of Rad & Mirzaei (2024). The differentiated scaffolding protocol offers a concrete operationalization of how AI might be calibrated to individual learners' ZPD. Second, and most importantly, the Holm-corrected dimensional analysis challenges the assumption that the benefits of generative AI are uniformly distributed across writing competencies. Effects concentrated at the macro level (organization, content) and dissipated toward the micro level (mechanics). This dimensional specificity has not been reported previously in the Indonesian L1 literature. Third, the study offers preliminary evidence that the macro-level scaffolding pattern holds across contexts (higher education, secondary EFL, and now L1 secondary education), suggesting it may be a robust feature of scaffolded ChatGPT instruction rather than a context-specific artifact.

Contextual Significance: The Madrasah Setting

The study's location at a madrasah (Islamic senior secondary school) in Parigi Moutong Regency adds two forms of contextual significance that warrant careful qualification. Institutionally, madrasahs have been underrepresented in the Indonesian educational technology literature, yet they serve a substantial portion of Indonesian secondary students (Siregar et al., 2025; Zh et al., 2024a). Zh et al. (2024b) and Zh et al. (2025) have recently demonstrated the feasibility of digital media interventions for madrasah, including Wordwall-based gamification and comparisons of learning media. The present study extends this line of work to generative AI applications in a non-religious-content subject. We explicitly caution against overinterpreting the Islamic character of the intervention: no specific Islamic values were integrated into the pedagogical design, and no qualitative evidence was collected to

examine whether the madrasah context moderated the effect. The study's institutional contribution is therefore descriptive rather than culturally specific.

Geographically, the successful implementation in Parigi Moutong, a region with the resource constraints of eastern Indonesia, suggests that meaningful generative AI integration is achievable in such settings when accompanied by appropriate pedagogical scaffolding and teacher mediation. The intervention required only the school's existing computer laboratory and standard broadband. This finding supports the view that the principal barrier to AI integration in under-resourced schools may be pedagogical rather than infrastructural.

Practical Implications

Three practical implications follow from the present findings. First, Indonesian language teachers should integrate ChatGPT as a scaffolded tutor during the ideation and revision phases of the writing cycle, where its benefits are concentrated. Second, teachers should set explicit behavioral norms that prevent ChatGPT from generating text while permitting it to ask questions and offer critiques, thereby preserving student authorship. Third, because the effect on mechanics did not materialize, explicit EYD instruction remains necessary alongside AI integration; ChatGPT cannot currently substitute for this instructional role. These implications are local to the intervention context and should not yet be generalized to national policy recommendations without multi-site replication.

Limitations

Seven limitations warrant explicit acknowledgment. First, the sample comprised 60 students from a single madrasah in a single region, which constrained external validity. Second, the intervention spanned eight meetings across four weeks; the longitudinal sustainability of the observed gains is unknown. Third, although the same teacher taught both conditions, Hawthorne effects were partially mitigated through control-group novelty framing; residual demand characteristics cannot be fully excluded in a non-blinded Classroom study. Fourth, exposure to ChatGPT outside the Classroom, while reduced by the pre-screening questionnaire and compliance agreement, could not be fully prevented in students' private lives; self-reported compliance may be imperfect. Fifth, the inter-rater reliability of $\kappa = .81$, though substantial, is below the .90 threshold often preferred for high-stakes assessment; a larger rater pool and additional training might yield higher agreement. Sixth, affective and motivational variables such as writing self-efficacy, engagement, and AI-related anxiety were not measured. Seventh, this study did not collect qualitative data on students' or teachers' experiences, which would have illuminated the mechanisms proposed in the discussion.

Directions for Future Research

Future research should address these limitations along five dimensions. First, multi-site replications across Indonesia (spanning Java, Sumatra, and the eastern archipelago) would test the geographical generalizability of the present pattern. Second, longitudinal studies tracking students over one or more academic years would examine whether ChatGPT-induced gains persist, decay, or transfer to other writing genres. Third, mixed-methods designs incorporating motivational and self-regulatory variables would illuminate the affective pathways of AI-assisted learning. Fourth, comparative studies of ChatGPT against specialized educational AI tools (Gemini, Copilot, or locally fine-tuned Indonesian language models) would clarify the relative affordances of different architectures. Fifth, given the null effect on

mechanics, targeted studies examining hybrid integrations of ChatGPT with EYD-conformant orthographic tools would test whether layered AI architectures can produce more uniformly distributed writing gains.

CONCLUSION

This quasi-experimental study examined the effect of ChatGPT-assisted writing instruction on argumentative essay skills at MAN 2 Parigi, Central Sulawesi. The intervention produced a large overall effect (Cohen's d = 1.30) and a very large ANCOVA-adjusted effect (partial η^2 = .52). After Holm-Bonferroni correction for multiple comparisons, however, the effect was specifically localized to organization and content. Vocabulary, language use, and mechanics did not survive correction. The dimensional specificity of the effect constitutes the main empirical contribution: ChatGPT, when scaffolded as an AI tutor rather than a text generator, operates as a macro-level cognitive scaffold rather than a general-purpose writing tutor. The study also demonstrates the feasibility of implementing generative AI in a madrasah in a geographically disadvantaged region.

Given the constraints of a single-site sample, the findings should be treated as locally significant and as hypotheses for larger-scale replication. Seven limitations warrant explicit acknowledgment, including the single-site sample, the short duration of the intervention, residual demand characteristics, imperfect control over external exposure, inter-rater reliability below the .90 threshold, the absence of affective variables, and the lack of qualitative data. Future work should examine longitudinal sustainability, motivational mediators, multi-site replication across Indonesia, comparative architectures of generative AI tools, and hybrid AI integrations that address the present intervention's limitations at the orthographic level.

REFERENCES

Anam, K., Rowiyah, S., Wicaksono, B. H., Setyaningrum, R. W., & Lestiono, R. (2025). ChatGPT for developing critical thinking in writing: Perspectives of senior high school English teachers in Indonesia. *English Review: Journal of English Education, 13*(2), 457-468. https://doi.org/10.25134/erjee.v13i2.11828

Annamalai, N., Uthayakumaran, A., Bervell, B., & Kumar, R. (2025). Examining the use of ChatGPT in argumentative writing: A mixed methods study. *Journal of Advanced Academics. Advance online publication*. https://doi.org/10.1177/1932202X251333859

Apriani, E., Daulay, S. H., Aprilia, F., Marzuki, A. G., Warsah, I., Supardan, D., & Muthmainnah. (2025). A mixed-method study on the effectiveness of using ChatGPT in academic writing and students' perceived experiences. *Journal of Language and Education, 11*(1), 26-45. https://doi.org/10.17323/jle.2025.17913

Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: Peer generated or AI generated feedback?. *International Journal of Educational Technology in Higher Education, 21*(1), 23. https://doi.org/10.1186/s41239-024-00455-4

Bašić, Ž., Banovac, A., Kružić, I., & Jerković, I. (2023). ChatGPT-3.5 as writing assistance in students' essays. *Humanities and Social Sciences Communications, 10*(1), 750. https://doi.org/10.1057/s41599-023-02269-7

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.).* New Jersey: Lawrence Erlbaum Associates.

Creswell, J. W., & Creswell, J. D. (2023). *Research design: Qualitative, quantitative, and mixed*

methods approaches (6th ed.). New York: SAGE Publications.

Darmawansah, D., Rachman, D., Febiyani, F., & Hwang, G.-J. (2025). ChatGPT-supported collaborative argumentation: Integrating collaboration script and argument mapping to enhance EFL students' argumentation skills. *Education and Information Technologies, 30*, 3803-3827. https://doi.org/10.1007/s10639-024-12986-4

Ekalia, Y. J., Jemadi, F., & Susanto, I. (2025). Critical thinking skills and argumentative writing ability: Is there any correlation?. *DIAJAR: Jurnal Pendidikan dan Pembelajaran, 4*(3), 471-482. https://doi.org/10.54259/diajar.v4i3.5108

Ghafouri, M., Hassaskhah, J., & Mahdavi-Zafarghandi, A. (2024). From virtual assistant to writing mentor: Exploring the impact of a ChatGPT-based writing instruction protocol on EFL teachers' self-efficacy and learners' writing skill. *Language Teaching Research. Advance online publication*. https://doi.org/10.1177/13621688241239764

Hake, R. R. (1998). Interactive engagement versus traditional methods: A six thousand student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66*(1), 64-74. https://doi.org/10.1119/1.18809

Huang, Y., & Zhang, L. J. (2020). Does a process genre approach help improve students' argumentative writing in English as a foreign language? Findings from an intervention study. *Reading & Writing Quarterly, 36*(4), 339-364. https://doi.org/10.1080/10573569.2019.1649223

Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach.* Boston: Newbury House.

Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi. (2023). *Hasil PISA 2022 Indonesia: Capaian literasi, numerasi, dan sains*. Pusat Asesmen Pendidikan.

Khampusaen, D. (2025). The impact of ChatGPT on academic writing skills and knowledge: An investigation of its use in argumentative essays. *LEARN Journal: Language Education and Acquisition Research Network, 18*(1), 963-988. https://doi.org/10.70730/PGCQ9242

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174. https://doi.org/10.2307/2529310

Li, H., Wang, Y., Luo, S., & Huang, C. (2024). The influence of GenAI on the effectiveness of argumentative writing in higher education: Evidence from a quasi-experimental study in China. *Journal of Asian Public Policy.* Advance online publication. https://doi.org/10.1080/17516234.2024.2363128

Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments, 11*(1), 9. https://doi.org/10.1186/s40561-024-00295-9

Niloy, A. C., Akter, S., Sultana, N., Sultana, J., & Rahman, S. I. U. (2023). Is ChatGPT a menace for creative writing ability? An experiment. *Journal of Computer Assisted Learning, 40*(2), 919-930. https://doi.org/10.1111/jcal.12929

Nugroho, A., Andriyanti, E., Widodo, P., & Mutiaraningrum, I. (2024). Students' appraisals post-ChatGPT use: Students' narrative after using ChatGPT for writing. *Innovations in Education and Teaching International, 62*(2), 499-511. https://doi.org/10.1080/14703297.2024.2319184

OECD. (2023). PISA 2022 results (Volume I): *The state of learning and equity in education*. OECD Publishing. https://doi.org/10.1787/53f23881-en

Rad, H. S., & Mirzaei, A. (2024). Developing feedback literacy, scaffolded writing, and resilience through intervention on feedback processes for L2 writing students. *Studies in Educational Evaluation, 81*, 101354. https://doi.org/10.1016/j.stueduc.2024.101354

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Siregar, H. S., Nurhamzah, N., Munir, M., & Fikri, M. (2025). Enhancing Islamic education through technology integration: A study of teaching practices in Indonesia. *Jurnal Ilmiah Peuradeun, 13*(2), 959-986. https://doi.org/10.26811/peuradeun.v13i2.1875

Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. F*rontiers in Psychology, 14*, 1260843. https://doi.org/10.3389/fpsyg.2023.1260843

Utami, S. P. T., Andayani, Winarni, R., & Sumarwati. (2023). Utilization of artificial intelligence technology in an academic writing class: How do Indonesian students perceive? *Contemporary Educational Technology, 15*(4), ep450. https://doi.org/10.30935/cedtech/13419

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard: Harvard University Press.

Woo, D. J., Wang, D., Guo, K., & Susanto, H. (2024). Teaching EFL students to write with ChatGPT: Students' motivation to learn, cognitive load, and satisfaction with the learning process. *Education and Information Technologies, 29*(18), 24963-24990. https://doi.org/10.1007/s10639-024-12819-4

Yasmin, M., Naseem, F., & Raza, M. (2025). Evaluating ChatGPT's effectiveness in enhancing argumentative writing: A quasi-experimental study of EFL learners in Pakistan. *Sustainable Futures, 10,* 100809. https://doi.org/10.1016/j.sftr.2025.100809

Zh, M. H. R., Sani, N. L., Kuswandi, D., & Fadhli, M. (2024a). Needs analysis of development FBO media as a support for blended learning in Al-Qur'an Hadits lesson. *Jurnal Pendidikan Agama Islam Al-Thariqah, 9*(1), 16-32. https://doi.org/10.25299/althariqah.v9i1.15383

Zh, M. H. R., Putra, M. F. B., Kuswandi, D., Wedi, A., & Ardiansyah, A. (2024b). Developing Wordwall evaluations in blended Islamic education using the Smith and Ragan model. *Al-Aulia: Jurnal Pendidikan dan Ilmu-Ilmu Keislaman, 10*(1), 89-104. https://doi.org/10.46963/aulia.v10i1.1745

Zh, M. H. R., Pradana, M. I. Y., Purnomo, Soepriyanto, Y., & Budiman, F. (2025). Comparative analysis of student learning outcomes in Al-Qur'an Hadith lessons based on learning media. *Al-Afkar: Journal for Islamic Studies, 8*(1), 241-250. https://doi.org/10.31943/afkarjournal.v8i1.1312