



AI-Assisted Academic Writing in Arabic Education: Causal Evidence and Ethical Implications from MAN 4 Hulu Sungai Tengah

Beni Suranto¹, Khalilurrahman², Irfan³, Syahabuddin Nur*⁴, M. Syihabuddin⁵

¹Auckland University Of Technology, New Zealand, ²IAI Darussalam Martapura, Indonesia,

³Universitas Negeri Makassar, Indonesia, ^{4,5}STIQ Amuntai, Indonesia

*Corresponding Author: sahabudin54690@gmail.com

Article Info:

Submitted: 1 Mei 2026

Published: 26 June 2026

Keywords:

AI-Assisted Writing;

Arabic Academic Writing;

Academic Integrity

Abstrak

This multi site study investigates both the effectiveness and ethical implications of AI-assisted academic writing in Arabic Language Education at MAN 4 Hulu Sungai Tengah. Employing a mixed-methods design, the study involved 90 students divided into two groups: an AI-assisted writing group and a conventional learning group. Quantitative findings indicate a statistically significant improvement in writing performance among students in the AI-assisted group, along with a notable difference in overall achievement compared to the conventional group. The large effect size suggests that properly guided AI support can accelerate drafting and revision processes while substantially enhancing writing quality. Qualitative results further highlight stage specific patterns of AI utilization. Generative AI tools were primarily employed during the early stages of writing, such as idea generation, outlining, overcoming writer's block, and improving academic tone. In contrast, editing tools were more frequently used for grammatical correction and stylistic refinement. From an ethical perspective, the findings reveal uneven risk distribution. Lower-performing students demonstrated a higher tendency toward overreliance on AI, often accompanied by a diminished sense of authorial voice. Conversely, higher-performing students were more likely to retain control over their writing and exhibited stronger awareness of academic integrity. Additionally, limitations of AI in handling culturally and disciplinarily specific terminology within Arabic academic discourse emphasize the continued necessity of instructor feedback to ensure conceptual accuracy and rhetorical appropriateness. Overall, this study proposes an effective and ethically grounded hybrid instructional model, positioning AI as a targeted scaffolding tool while integrating transparency standards and context-sensitive pedagogical safeguards.

INTRODUCTION

Generative AI and AI enabled writing assistants have rapidly reshaped how academic writing is taught, practiced, and assessed in higher education worldwide. (Tusino et al. 2025) Generative AI and AI enabled writing assistants are rapidly reshaping academic writing in higher education worldwide, offering guided feedback that can speed revision and improve clarity and organization while simultaneously heightening concerns over integrity, authorship transparency, and the limitations of detection based enforcement prompting quality assurance bodies to call for redesigned assessment and stronger AI literacy so pedagogical gains do not erode academic standards. (Mahapatra 2024)

Within language education, academic writing development requires not only grammatical accuracy but also argumentation quality, genre control, and sustained revision



discipline, which makes AI support particularly attractive. In Arabic Language Education, the challenge intensifies because academic tone, rhetorical legitimacy, and culturally embedded terminology can be difficult to master and easy to distort when simplified. This study contextualizes the issue within a single institutional setting, namely Madrasah Aliyah Negeri (MAN) 4 Hulu Sungai Tengah, in order to capture variations in learning dynamics within a secondary-level educational environment. This setting is significant, as students' writing practices, levels of exposure to AI tools, and understandings of academic integrity can vary even within comparable curricular structures.

Accordingly, the material object of this study is AI-assisted academic writing in Arabic language education, examined through both measurable changes in writing performance and the processes of tool-mediated writing among students at MAN 4 Hulu Sungai Tengah.

Recent studies commonly report that generative AI can improve students' writing products, especially when integrated as feedback support rather than full automation of composing. However, many of these studies emphasize end scores while giving limited attention to *mechanisms*, such as how different tools serve different stages of writing (planning, drafting, revising). (Dewi et al. 2024) This limitation matters because "AI use" is not a single treatment, but a portfolio of micro interventions that can shift cognitive effort across tasks and potentially inflate outcomes if process visibility is weak. Prior work also warns that AI-generated prose can resemble student writing at a surface level, complicating how improvement should be interpreted without complementary process evidence. Therefore, a clear gap is the need to connect measurable performance gains with stage specific patterns of tool use in authentic classroom writing tasks. (Kurniawan et al. 2025)

The second gap concerns ethics and integrity, because recent scholarship shows that AI related misconduct risk is driven by motivation, perceived norms, and self control rather than by tool availability alone. While institutions increasingly publish guidance on responsible Gen AI use, classroom level evidence linking integrity perceptions to measurable writing outcomes remains uneven, especially outside Anglophone contexts. (Alghazo, Al Salem, and Alrashdan 2021) Furthermore, studies on academic integrity in AI mediated writing highlight persistent concerns about authorship ambiguity and paraphrasing misuse, but often treat "loss of voice" as a conceptual worry rather than an observable gradient across student capability levels.

However, recent integrity scholarship emphasizes that scaffolding becomes educational only when students retain agency and verification practices, because outsourcing thinking undermines learning and authorship. (Alghazo et al. 2021) Quality assurance perspectives similarly argue that pedagogical integration must be paired with assessment design and transparency norms to maintain standards in the Gen AI era Finally, discipline sensitive views of voice suggest that academic writing quality includes rhetorical identity and conceptual precision, which require human oversight in culturally and theologically loaded Arabic discourse. (Ahmed and Rezk 2026)

In secondary level institutions such as Madrasah Aliyah Negeri (MAN) 4 Hulu Sungai Tengah, students are increasingly exposed to AI tools while still requiring substantial instructional guidance to meet academic writing standards. This situation creates a practical demand for scalable forms of support. More broadly, limitations in individualized teacher feedback often lead students to rely on AI tools for drafting, revising, and paraphrasing, particularly under time constraints.

At the same time, academic discourse in Indonesia has begun to highlight AI use as a potential source of ambiguity in authorship and risks related to improper paraphrasing, especially in contexts where policy frameworks and AI literacy remain uneven. In the domain of Arabic academic writing, the challenge extends beyond grammatical correctness to include mastery of academic tone and culturally embedded terminology. As a result, students may produce linguistically fluent texts that nonetheless lack disciplinary legitimacy without adequate teacher supervision.

These conditions justify examining AI-assisted academic writing and its ethical boundaries within the specific context of MAN 4 Hulu Sungai Tengah, rather than treating AI

use as a generalized, context-independent phenomenon. This study aims to test whether AI-assisted writing support can improve students' academic writing performance in Arabic while also documenting how specific AI tools are utilized across different stages of the writing process, thereby linking outcomes with underlying mechanisms. A second objective is to empirically examine ethical boundaries by identifying patterns of academic integrity concerns, tendencies toward overreliance, and perceptions of "loss of authorial voice" as observable behaviors rather than abstract issues.

By focusing on MAN 4 Hulu Sungai Tengah, this study contributes contextually grounded evidence that refines broader claims about AI integration in education. Its scholarly contribution lies in combining causal evidence of effectiveness with qualitative insights into tool usage patterns, ethical negotiation, and discipline-specific constraints in Arabic academic writing. Ultimately, the study proposes a pedagogically actionable hybrid model of AI integration that enhances writing development while safeguarding academic integrity and preserving authorial ownership.

RESEARCH METHODS

This study employed a mixed-methods quasi-experimental pretest–posttest control group design conducted in intact classes at Madrasah Aliyah Negeri (MAN) 4 Hulu Sungai Tengah. This design was selected because random assignment was not feasible in the school context, while still enabling reasonably robust causal inference through careful control of internal validity threats. (Creswell and Creswell 2018)

Participants were drawn using intact class sampling from Arabic academic writing-related subjects and assigned into an experimental group and a control group with comparable instructional objectives and task exposure. The experimental group received structured AI-assisted writing scaffolding utilizing tools such as ChatGPT for idea generation and outlining, and Grammarly for post draft refinement, along with rephrasing tools when appropriate whereas the control group completed equivalent writing tasks without AI support. (Hedges 2025)

Data collection consisted of rubric based writing assessments administered as pretests and posttests, complemented by questionnaires and short reflective responses capturing stage specific tool usage, perceived benefits, and ethical concerns. Quantitative analysis involved assumption testing followed by within group gain analysis and between group posttest comparisons, including effect size estimation. (Capili and Anastasi 2024)

RESULT

This section reports the results of a mixed-methods study that examined whether AI writing assistants support (as cognitive scaffolds) can improve students' academic journal writing, while also mapping how students negotiate ethical boundaries during AI use. Quantitative evidence was derived from pre-test and post-test writing scores, assumption checks (Shapiro Wilk normality and Levene's homogeneity of variance), paired and independent-samples t-tests, and effect size estimates (Cohen's d). Qualitative evidence was drawn from post-intervention survey responses and thematic coding of students' reflections to describe tool use patterns across writing stages, self-regulation, overreliance, and issues of authorial voice and academic integrity.

The experimental group ($n = 45$) showed an increase in the mean writing score from 68.4 ($SD = 5.1$) at the pre-test to 81.7 ($SD = 4.4$) at the post-test. Shapiro–Wilk tests indicated that the data were normally distributed (pre-test $W = 0.972$, $p = 0.34$; post-test $W = 0.976$, $p = 0.41$). A paired-samples t-test confirmed a statistically significant improvement, $t(44) = 11.63$, $p < 0.001$, with Cohen's $d = 1.26$ (a large effect). Substantively, this pattern suggests that after the intervention the participants were able to produce better-structured drafts with more consistent academic language, implying that AI-supported drafting and revision helped reduce bottlenecks in planning, sentence formulation, and iterative editing.

Table 1. Pre-test and Post-test Writing Scores

Test	Mean	SD
Pre-test	68.4	5.1
Post-test	81.7	4.4

Table 1 summarizes the shift in mean scores before and after the intervention. The reduction in the standard deviation at post-test (from 5.1 to 4.4) indicates that scores became more concentrated, suggesting that the improvement was shared by most students rather than driven by a small subset. Post-test normality assumptions were met (Shapiro–Wilk $p = 0.297$ for the experimental group; $p = 0.187$ for the control group), and variances were homogeneous (Levene $p = 0.260$). An independent-samples t -test showed a significant difference: the experimental group ($M = 87.13$, $SD = 4.95$) outperformed the control group ($M = 80.71$, $SD = 5.42$), $t(88) = 6.423$, $p < 0.001$, with Cohen’s $d = 1.04$ (a large effect). This result indicates that at the post-intervention point, students who received AI-assisted support produced higher-quality writing than those who relied on conventional support alone. The magnitude of the effect suggests that the difference is educationally meaningful, not merely statistically detectable.

Table 2. Assumption Tests and Main Between-Group Comparison

Test	Group(s)	Statistic	p-value	Decision
Shapiro–Wilk	Experimental	$W = 0.981$	0.297	Normal
Shapiro–Wilk	Control	$W = 0.974$	0.187	Normal
Levene	Exp vs Ctrl	$F = 1.287$	0.260	Homogeneous
Independent t -test	Exp vs Ctrl	$t(88) = 6.423$	< 0.001	Significant
Effect size	—	$d = 1.04$	—	Large

Table 2 reports the statistical assumptions and the main between-group test. Both groups satisfy normality and homogeneity conditions, supporting the appropriateness of the independent-samples t -test. The large effect size reinforces that the intervention’s impact is substantial in practice. The qualitative results below describe how students used AI across the writing process and how these practices relate to self-regulation and ethical risk. The integrated reading of survey patterns and thematic codes suggests that AI support can strengthen autonomy and reduce writing anxiety, but it may also increase the likelihood of overreliance and a weakened authorial voice when students lack clear guidance and monitoring. Survey data indicate that AI use was strategic rather than uniform: ChatGPT dominated the ideation stage (outlining, clarifying ideas, setting an academic tone), while Grammarly dominated post-drafting revision (grammar, sentence structure, and register consistency). A significant association was found between writing stage and tool preference ($\chi^2 = 28.35$, $p < 0.01$), with a moderate strength of association (Cramer’s $V \approx 0.45$ – 0.52 , depending on the tested pairing).

In practical terms, students appear to allocate different AI tools to different cognitive needs: generative tools to scaffold planning and content organization, and automated feedback tools to scaffold surface-level editing. This division is consistent with a scaffolding interpretation, where supports change as task demands change across stages.

Table 3. Distribution of AI Tool Utilization Across Writing Stages

Writing stage	Tool	Dominant function	Usage
Ideation	ChatGPT	Idea generation; outlining; translating/reformulating	65.6%
Ideation	Grammarly	Tone refinement	59.4%
Ideation	QuillBot	Rephrasing; style adjustment	53.3%
Post-drafting	Grammarly	Grammar & syntax correction	78.9%
Post-drafting	QuillBot	Sentence restructuring	66.7%

Table 3 shows that tool choice shifts with writing demands. Generative assistance is most prominent at ideation, while automated feedback becomes dominant after drafting. This pattern supports the view that students treat AI as a set of stage specific supports rather than a single end-to-end replacement for writing.

High-performing students were more likely to use AI responsibly and to express stronger ethical concerns, whereas lower-performing students were more prone to overreliance under deadline pressure and more frequently reported a ‘loss of authorial voice’. This gradient suggests that differences in self-regulation and AI literacy may shape whether AI operates as a productive scaffold or a risky substitute. In instructional terms, the most vulnerable group requires explicit training in verification strategies, authorship norms, and transparent reporting of AI assistance.

Table 4. Ethical and Cognitive Engagement with AI Tools by Performance Level

Category	Used AI responsibly	Expressed ethical concern	Reported overreliance	Felt loss of voice
High (≥ 85)	82.6%	78.3%	17.4%	13.0%
Mid (75–84)	61.4%	55.1%	33.3%	28.9%
Low (< 75)	39.7%	26.5%	61.8%	52.9%

Table 4 documents a clear gradient: as performance decreases, responsible AI use and ethical concern fall, while overreliance and loss of voice rise sharply. The pattern implies that integrity safeguards and metacognitive scaffolds should be prioritized for lower performing students. Most respondents perceived AI generated text as often grammatically correct but less aligned with the demands of academic tone and disciplinary precision, particularly for epistemically dense and culturally bound content. This finding positions AI as strong at surface-level form but limited for disciplinary judgment in Arabic academic contexts, especially when terminology is sensitive and meaning depends on scholarly tradition. Consequently, human feedback remains central for concept accuracy, rhetorical appropriateness, and discipline specific style.

Table 5. Student Evaluation of AI Output Quality in Arabic Academic Contexts

Indicator	Agreement
Grammatically correct but lacks academic tone	74.4%
Misrepresents culturally specific/theological terminology	68.9%
Teacher feedback remains essential	81.1%
Reliable for theology/classical Arabic content	26.7%
Prefer hybrid workflow (AI + human input)	71.1%

Table 5 indicates that students tend to value AI for linguistic correctness but remain cautious about its reliability for content that is sensitive to scholarly context. The strong preference for a hybrid workflow underscores that AI is viewed as complementary support rather than an authoritative source. Most students reported higher confidence and reduced writer’s block after the intervention. However, the perceived benefits were greatest when AI support was combined with human feedback from instructors or peers. From a learning perspective, AI appears to act as an immediate catalyst that lowers entry barriers to drafting and revision. Nevertheless, sustained quality improvement still depends on structured feedback cycles that reinforce argumentation, citation consistency, and the development of a stable authorial voice.

Table 6. Student Perceptions of AI-Assisted Writing After Intervention

Indicator	n	agreement
More confident in academic writing	74	82.2%
Helped overcome writer's block	67	74.4%
Improved sentence variation	63	70.0%
Stronger outcomes when combined with human feedback	61	67.8%
AI-only reliance reduced engagement	40	44.4%
AI seen as scaffold, not final solution	65	72.2%

Table 6 summarizes post-intervention perceptions that are broadly positive, while also warning against AI only dependence. The results support a pedagogy that emphasizes process documentation (reflection notes, revision trails) and structured consultation to keep AI assistance within ethical and educationally meaningful boundaries. The study investigated whether AI writing assistants function as cognitive scaffolds that improve academic journal writing among Arabic Language Education students. Quantitative data were generated from pre test and post test writing tasks assessed with an analytic rubric, while qualitative data came from questionnaires and students' reflective accounts of how AI tools supported drafting and revision. The analysis combined assumption checks (normality and homogeneity), inferential tests (paired and independent t-tests), effect size estimation (Cohen's d), and pattern mapping of tool use across writing stages. In short, the dataset captured both measurable changes in writing quality and the process-level mechanisms that plausibly explain those changes, allowing the findings to be interpreted as more than a simple "score increase."

The experimental group (n = 90) demonstrated a substantial improvement in writing quality from pre test to post test, and parametric assumptions supported the use of t-tests; the gain was statistically significant with a large effect size, indicating that structured AI assisted writing support was associated with meaningful progress over the intervention period. This pattern was reinforced by the between group comparison at post test, the experimental group outperformed the control group with a significant difference and a large effect, providing convergent evidence that guided AI scaffolding can produce better writing outcomes than conventional instruction alone.

Process and perception data clarify *how* the benefits occurred and *where* limits and risks emerged. Students reported function-specific tool use across stages Chat GPT mainly for ideation/planning, Grammarly for post draft refinement, and QuillBot as supplementary rephrasing with a significant stage tool association of moderate strength, suggesting AI was used as targeted scaffolding rather than a single "magic button." Ethically, risks were stratified by performance level lower performing students reported more overreliance and "loss of voice," while higher performing students more often maintained authorial control and integrity awareness, implying that ethical guidance should be differentiated. In Arabic academic contexts, participants noted a key boundary condition AI outputs were often grammatically acceptable but less reliable for academic tone and culturally/theologically embedded terminology, which explains strong preferences for instructor feedback and a hybrid workflow that combines AI enabled iteration with human expertise and explicit integrity norms.

DISCUSSION

The findings can be interpreted through a cognitive pedagogical lens generative AI functions as a cognitive offloading and orchestration aid that reduces planning and revision burden. By easing these demands, students can allocate more attention to higher order writing outcomes such as argument coherence, organization, and rhetorical fit. This helps explain why meaningful improvements can occur within a limited instructional window.(Tusino et al. 2025) Importantly, the gains are more defensibly read as changes in *process efficiency* than as automation of authorship. In other words, AI did not "write" in place of students, but supported them to move faster and more deliberately through composing decisions. The intervention

therefore operates as a scaffold that makes complex writing tasks more manageable. This framing keeps the interpretation aligned with a learning-oriented account of performance change. (Samanhudi and Rachmawati 2025)

This interpretation is consistent with empirical work showing that AI support is most educationally productive when students actively evaluate and transform suggestions. When learners treat AI outputs as hypotheses to test rather than text to copy revision becomes a cognitive activity instead of a mechanical one. From a feedback perspective, AI contributes timeliness and frequency of input, while instruction provides standards, conceptual guidance, and evaluative judgment. The central mechanism can be described as acceleration of the draft review revise cycle through repeated micro revisions. Over time, small improvements in clarity, structure, and phrasing accumulate into measurable quality gains. This also explains why guided use matters more than mere access to tools. Without explicit guidance, students may not convert AI feedback into learning-relevant revisions.

The process evidence strengthens the mechanism claim because it indicates purposeful, stage aligned tool use rather than undirected dependence. Students' preferences suggest they treated tools as functionally differentiated scaffolds across the writing pipeline. Generative AI was used primarily to initiate and structure ideas, while editing assistants were used to polish language after a draft existed. This division of labor mirrors how expert writers separate invention, drafting, and editing as distinct cognitive phases. It also reduces bottlenecks by giving support exactly where the task is most effortful at each stage. (Putri, Imaniah, and Nargis 2025) As a result, students can move from rough ideas to workable drafts and then to refined prose more efficiently. The implication is that instructional design should teach this sequencing explicitly to maximize learning benefits.

Ethically, the study contributes an interpretive nuance by showing that integrity risks are patterned rather than uniform across learners. Overreliance and "loss of voice" can be understood as self regulation vulnerabilities that intensify under lower proficiency and deadline pressure. (Wulandari 2024) In contrast, higher performing students more often preserved authorial control by using AI to test, refine, and iterate ideas. This implies that ethics instruction should be differentiated, with stronger guardrails and monitoring for students most susceptible to dependency. The Arabic academic writing context further clarifies a boundary condition: grammatical fluency does not guarantee disciplinary adequacy when culturally embedded or theological terminology demands precision and register sensitivity. Teacher feedback therefore remains decisive for conceptual validity and rhetorical legitimacy. Taken together, the findings support a hybrid model in which AI scaffolds planning and linguistic refinement, while pedagogy and assessment anchor transparency, voice, and disciplinary standards.

CONCLUSION

The findings indicate that AI-assisted academic writing can substantially improve students' writing quality when deployed as a structured scaffold that strengthens planning and revision rather than replacing thinking. This study contributes an evidence-based, discipline-sensitive model showing that stage aligned tool use (generative AI for ideation/organization and editing assistants for refinement) can accelerate drafting revision cycles while exposing stratified integrity risks, particularly overreliance and loss of voice among weaker writers. The implications are clear: institutions should adopt a hybrid approach that keeps instructor feedback central for rhetorical legitimacy and culturally embedded terminology in Arabic academic writing, and implement explicit, differentiated integrity guidance and transparent assessment designs to preserve authorship and disciplinary standards.

REFERENCES

- Ahmed, Abdelhamid M., and Lameya M. Rezk. 2026. "Frame Markers in Qatari University Students' Arabic and English Argumentative Writing: A Corpus-Based Study of Metadiscourse and Metalinguistic Understanding." *Language and Education* 40(1): 1–23. doi:10.1080/09500782.2025.2608793.

- Alghazo, Sharif, Mohd Nour Al Salem, and Imran Alrashdan. 2021. "Stance and Engagement in English and Arabic Research Article Abstracts." *System* 103: 102681. doi:10.1016/j.system.2021.102681.
- Alghazo, Sharif, Mohd Nour Al Salem, Imran Alrashdan, and Ghaleb Rabab'ah. 2021. "Grammatical Devices of Stance in Written Academic English." *Heliyon* 7(11): e08463. doi:10.1016/j.heliyon.2021.e08463.
- Capili, Bernadette, and Joyce K. Anastasi. 2024. "An Introduction to the Quasi-Experimental Design (Nonrandomized Design)." *The American journal of nursing* 124(11): 50–52. doi:10.1097/01.NAJ.0001081740.74815.20.
- Creswell, John W., and J. David Creswell. 2018. *Research Design: Qualitative, Quantitative, And Mixed Methods Approaches*. Fifth edition. Los Angeles: SAGE.
- Dewi, Azzaria Mida Mustika, Muhammad Jafar Shodiq, Nurhikmah Sani, and Faizmailatus Sofa. 2024. "Digital Transformation of Kitabah Learning in Indonesia: A Sinta2-Indexed Journal Analysis." *Al-Arabi: Jurnal Bahasa Arab dan Pengajarannya = Al-Arabi: Journal of Teaching Arabic as a Foreign Language* 8(1): 67–80. doi:10.17977/um056v8i1p67-80.
- Hedges, Larry V. 2025. "Interpretation of the Standardized Mean Difference Effect Size When Distributions Are Not Normal or Homoscedastic." *Educational and Psychological Measurement* 85(2): 245–57. doi:10.1177/00131644241278928.
- Kurniawan, Citra, Ence Surahman, Deka Dyah Utami, and Rumaizah Mohd Nordin. 2025. "Attempts to Harness AI Model in Online Differentiated Learning on Students' Understanding Levels." *JTP - Jurnal Teknologi Pendidikan* 27(2): 587–97. doi:10.21009/jtp.v27i2.54400.
- Mahapatra, Santosh. 2024. "Impact of ChatGPT on ESL Students' Academic Writing Skills: A Mixed Methods Intervention Study." *Smart Learning Environments* 11(1): 9. doi:10.1186/s40561-024-00295-9.
- Putri, Intan Cahyani, Ikhfi Imaniah, and Nargis Nargis. 2025. "Using Chatgpt To Support Efl Writing: Student Insights And Experiences." *Eltin: Journal of English Language Teaching in Indonesia* 13(2): 423–32.
- Samanhudi, Udi, and Dina Rachmawati. 2025. "Towards a Structured AI-Integrated Academic Writing Guideline: Perspectives from the Academic Community." *Journal of English Language Studies* 10(2): 246–66. doi:10.62870/jels.v10i2.35395.
- Tusino, Tusino, Puspa Dewi, Mufid Lahiria Permata Hati, and Maria Joy C. Iglesia. 2025. "ChatGPT in Academic Writing Classes: Exploring EFL Students' Perspectives." *JOLLT Journal of Languages and Language Teaching* 13(4): 2162–72. doi:10.33394/jollt.v13i4.15293.
- Wulandari, Kurnia. 2024. "Research Trends in the Use of Automated Writing Evaluation Tools." *Lingua Pedagogia* 6(1). doi:10.21831/lingped.v6i1.76507.