ISSN: 2655-1780 (Elektronik)
ISSN: 2654-8534 (Cetak)

METACOGNITION-BASED ACADEMIC WRITING WITH GENERATIVE-AI SCAFFOLDING: STUDENT AND INSTRUCTOR NEEDS

Tofan Stofiana[1*], Dadang Sunendar[2], Yeti Mulyati[3], Andoyo Sastromiharjo[4]

Pendidikan Bahasa Indonesia, Universitas Pendidikan Indonesia, Bandung, Indonesia[1234]
E-mail[*]: tofanstofiana@upi.edu

ABSTRAK

This study examines metacognition-oriented academic writing needs and the supportive role of generative AI among cross-disciplinary master's students. A convergent mixed-methods design combined an online survey (n = 120; 20 Likert items), semi-structured interviews (12 students, 8 faculty), and document analysis of syllabi, rubrics, and anonymized student texts to compare intended scaffolding in curricula with scaffolding experienced in practice. Results show that metacognitive awareness exceeded the scale midpoint, with the highest scores in Linguistics, while difficulties in constructing coherent arguments were more evident in Management and Law. Attitudes toward generative AI were generally positive in these professional programs, though ethical policies and reflective opportunities varied across courses. Thematic and document analyses indicate that paragraph-level argument quality improves when rubric criteria are accompanied by worked examples of claim–evidence–warrant patterns and when feedback cycles require explicit articulation of reasoning links and transitions. Pilot reliability (Cronbach's $\alpha < .70$ on some subscales) suggests several items require refinement; quantitative data were therefore interpreted descriptively and triangulated with qualitative and documentary findings. Recommendations include embedding brief, rubric-aligned reflective checkpoints, using worked-example and self-explanation exercises to stabilize claim–evidence–warrant connections, and providing explicit ethical guidance that positions generative AI as a reasoning companion rather than a substitute for reflection.

Keywords: Academic Writing; Argumentation; Generative AI; Higher Education; Metacognition

INTRODUCTION

Graduate-level academic writing requires students to engage in deliberate planning, monitoring, and evaluation of their own composing while meeting academic expectations for clarity, coherence, and evidential rigor. The process reflects a high level of metacognitive regulation, in which writers take responsibility for managing their learning, setting goals, and evaluating progress. Metacognition, as both awareness of one's cognitive processes and regulation of those processes, enables learners to transform writing from a mechanical act into a reflective and goal-directed endeavor. Research has consistently shown that students with stronger

metacognitive awareness demonstrate higher writing self-efficacy, stronger coherence in argumentation, and greater responsiveness to feedback (Rosdiana et al., 2023; González et al., 2020; Huang & Zhang, 2022; Prihandoko et al., 2024). These capacities are strengthened through guided activities such as reflective journals, structured peer review, and writing retreats that cultivate autonomy and sustained engagement with the writing process (Bonnamy et al., 2024; Busby & Malone, 2023; Sudirman et al., 2021).

Argument quality represents a distinct yet related dimension of academic writing. It reflects not only linguistic accuracy but also the reasoning quality through which claims are justified by evidence. The capacity to connect claims, evidence, and warrants forms the foundation of coherent argumentation. Instructional approaches that explicitly guide learners in reasoning organization, such as modeling argument structures and prompting reflection, have been shown to foster greater analytical precision and reflective awareness (Rivas et al., 2022; Dennis & Somerville, 2022). Learners who consciously evaluate the logic linking their claims and evidence tend to produce more coherent and well-supported arguments (Baumanns & Rott, 2022; Sethares & Asselin, 2021). The ability to articulate these logical connections is therefore an indicator of both argumentation skill and metacognitive maturity.

Theoretical models of argument quality provide further insight into how writers construct and evaluate reasoning. The Toulmin Argument Pattern describes argument strength in terms of logical relations among claims, data, and warrants, clarifying how justifications mediate between evidence and assertion (Warren, 2010; Yang, 2023). The knowledge-transforming model proposed by Bereiter and Scardamalia (1987) positions writing as an iterative process of reflection and reorganization, where conceptual understanding and rhetorical goals co-evolve through self-monitoring. More recent perspectives, such as the Epistemic Cognition Model, view argumentation as an epistemic act in which writers assess the nature, credibility, and justification of knowledge itself (Chinn, Buckland, & Samarapungavan, 2011). Argument quality, therefore, depends on writers' ability to apply metacognitive regulation to epistemic evaluation, deciding what counts as valid evidence, how strongly it supports a claim, and whether reasoning remains coherent and ethically sound.

Table 1. Conceptual Mapping of Argument Components and Metacognitive Regulation

Argumentation Component	Function in Toulmin / Bereiter–Scardamalia Models	Corresponding Metacognitive Regulation	Typical Student Action
Claim	Main assertion or stance the writer intends to defend.	Planning – setting goals and organizing argument trajectory.	Formulate thesis statement; anticipate counter-arguments.
Evidence (Data)	Facts, citations, or examples supporting the claim.	Monitoring – checking relevance, adequacy, and credibility of support.	Select sources; verify consistency between data and claim.
Warrant	Logical or conceptual bridge linking evidence to claim.	Evaluation – judging sufficiency of justification and coherence.	Reflect on reasoning chain; revise transitions or logical connectors.

The mapping shows how argument components align with metacognitive phases, emphasizing that coherent reasoning in writing emerges through conscious regulation of thought. Writers who effectively manage their planning, monitoring, and evaluation processes demonstrate stronger argumentative clarity and reflective judgment.

Recent developments in educational technology have added new layers to this dynamic. Generative artificial intelligence (AI) tools such as ChatGPT and automated feedback systems now provide writers with opportunities to externalize their reasoning and receive immediate coherence checks or structural suggestions. When used critically, these tools can enhance metacognitive engagement by triggering reflective assessment and encouraging revision cycles (Ahn & Alkhaqani, 2024; Mahapatra, 2024; Fontenelle-Tereshchuk, 2024). Studies in AI-supported self-regulated learning highlight that generative systems can act as adaptive scaffolds that complement learners' planning, monitoring, and evaluation behaviors (Lai, 2023; Zawacki-Richter et al., 2023). In these contexts, AI functions as a metacognitive partner that helps writers focus on reasoning quality while reducing cognitive overload.

Researchers have also pointed out that uncritical reliance on AI-generated feedback may undermine self-regulation and reflective judgment. Responsible integration thus requires pedagogical and institutional frameworks that emphasize human oversight, ethical reflection, and transparent disclosure (Chan, 2023; UNESCO, 2023, 2025; Xia et al., 2024). Educational policies that couple AI use with metacognitive prompts, asking students to evaluate, not merely accept, machine-generated suggestions, can maintain reflective agency and deepen epistemic understanding (Yang & Xia, 2023; Xu et al., 2025). These structured reflective cycles position AI not as a replacement for cognition, but as a catalyst for metacognitive reasoning and critical thinking.

Synthesizing these strands suggests that metacognitive regulation, argumentation quality, and AI scaffolding operate as an integrated system. Metacognition provides the internal regulation needed to plan, monitor, and evaluate writing; AI offers external scaffolding that both supports and challenges these processes; and argumentation quality represents the observable outcome of this interaction. This study conceptualizes these relationships within a unified framework linking internal and external regulation to epistemic reasoning outcomes.

The following conceptual framework illustrates the relationships among metacognitive regulation (planning–monitoring–evaluating), generative-AI scaffolding as a supportive or reflective challenge, and argumentation quality (claim–evidence–warrant reasoning). See Figure 1.

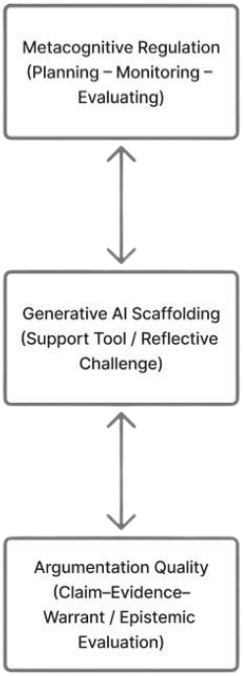

Figure 1. Conceptual Framework of the Study

The framework positions graduate writing as a reflective interaction between human regulation and intelligent assistance, where metacognition drives reasoning, AI acts as a scaffold for reflection, and argument quality serves as the developmental outcome of this interplay.

Guided by this framework, the study addressed the following research questions:

- What metacognitive practices, difficulties, and training needs do graduate students report in argumentative academic writing across programs?
- How do lecturers describe scaffolding strategies and constraints for paragraph-level argument construction, and how do these align with rubric criteria and worked examples?
- How do syllabi, rubrics, and anonymized student work reflect intended scaffolding relative to students' experienced scaffolding?
- How do students and lecturers perceive the role and ethical boundaries of generative-AI support in academic writing?

METHODS

Design

A convergent mixed-methods design was employed to identify baseline needs among graduate students and lecturers regarding metacognitively oriented academic writing supported by generative-AI scaffolding. Quantitative, qualitative, and documentary data were collected concurrently and integrated in a single research phase. Integration used joint displays that linked findings across strands and produced transparent meta-inferences connecting quantitative signals and qualitative insights (Creswell & Plano Clark, 2018; Fetters et al., 2013; Guetterman et al., 2015). The approach provided a comprehensive perspective on participants' learning experiences and reflective engagement with AI-assisted writing.

Setting and Participants

The study was conducted at two anonymized institutions, referred to as Public University and Private University. Four master's programs participated: Indonesian Language and Literature Education, Linguistics, Management, and Law. The selection was guided by two rationales. The first concerned disciplinary variation, allowing comparison between linguistically oriented programs emphasizing reflective writing and professionally oriented programs stressing applied reasoning. The second concerned ethical and logistical accessibility, ensuring feasible data collection under approved research-ethics protocols.

Institution	Programs Included	Epistemic Orientation	Student Participants (Survey)	Interview Participants (Students)	Interview Participants (Lecturers)
Public University	• Indonesian Language & Literature Education • Linguistics	Language, pedagogy, reflective reasoning	60	6	4
Private University	• Management • Law	Professional, applied reasoning	60	6	4
Total	4 programs	Cross-disciplinary comparison	**120**	**12**	**8**

A purposive sampling strategy recruited 120 master's students (30 per program) actively engaged in academic writing tasks. The qualitative sample consisted of 12 students (three per program) and 8 lecturers (two per program), selected through maximum-variation sampling to capture diverse scaffolding experiences. Inclusion criteria required current enrollment, recent experience with major academic writing within 12 months, and provision of written informed consent.

Instruments and Materials

Quantitative data were collected through an online questionnaire containing 20 Likert-scale items (1–5) distributed across five dimensions: (1) metacognitive awareness of planning, monitoring, and evaluation; (2) academic-writing difficulties; (3) perceived lecturer scaffolding; (4) attitudes toward generative AI; and (5) training needs. Item construction was adapted from the Metacognitive Awareness Inventory for the graduate-writing context (Schraw & Dennison, 1994). Internal consistency yielded Cronbach's $\alpha = 0.71–0.84$, demonstrating acceptable reliability for an exploratory instrument.

Semi-structured interviews were designed for both students and lecturers. The student protocol explored personal writing experiences, coping strategies, scaffolding received, and ethical perspectives on AI. The lecturer protocol focused on supervisory strategies, instructional constraints, feedback practices, and reflections on integrating generative AI responsibly.

Documentary data consisted of anonymized syllabi, analytic rubrics, and de-identified student texts. These were analyzed to compare intended scaffolding in curriculum design with experienced scaffolding in practice, emphasizing reflection opportunities, clarity of argument criteria, revision feedback, and explicit mention of AI use.

Brief Note on Instrument Reliability

The quantitative instrument was treated as a pilot measure. Internal consistency was examined using Cronbach's alpha at the dimension level as an initial indicator of reliability. Several items with low item–total correlations were earmarked for refinement through future iterative testing.

Procedures

Data collection spanned a single academic semester. The survey link was distributed through official mailing lists and course channels. Interviews were conducted in person or online, lasted 30 - 45 minutes, were audio-recorded with consent, and transcribed verbatim. Documents were obtained from instructors and program coordinators, anonymized, and coded by role, program, and institution to facilitate comparative analysis.

Data Analysis

Quantitative data were summarized through means and standard deviations for each dimension and program. Reliability was verified via Cronbach's alpha at the subscale level. Qualitative data underwent a six-phase thematic analysis using NVivo 12: familiarization, initial coding, theme development, review, definition, and reporting (Braun & Clarke, 2006). The coding framework combined deductive categories, metacognition, writing challenges, scaffolding, AI ethics, and training needswith inductive codes emerging from participant narratives.

Document analysis applied 1–5 rating indicators across materials with coder annotations, producing summaries of gaps between intended and enacted scaffolding. Integration occurred through a joint-display matrix aligning quantitative patterns, qualitative themes, and document indicators to yield explicit cross-strand inferences (Fetters et al., 2013; Guetterman et al., 2015).

Coder Training and Reliability

Two coders (the first author and a trained research assistant) independently analyzed a shared subset of 20% of transcripts to establish inter-rater consistency. Training involved two calibration rounds using sample excerpts to align code definitions and boundary judgments. Coding discrepancies were discussed until conceptual alignment was reached. Inter-rater reliability was computed using Cohen's $\kappa = 0.82$, indicating substantial agreement (Landis & Koch, 1977). The

remaining data were coded independently, with periodic peer debriefing to ensure consistency.

Quality, Reliability, and Trustworthiness

Credibility was ensured through triangulation of data sources (students, lecturers, and documents), peer-debriefing, and detailed audit trails of coding decisions. Dependability and confirmability were reinforced through maintenance of a comprehensive codebook, analytic memos, and iterative reviews of thematic stability. Quantitative reliability was supported by internal-consistency measures, and qualitative reliability was demonstrated through high coder agreement, providing methodological coherence across strands.

Ethical Considerations

All procedures followed institutional ethical standards and obtained prior approval. Participants received information sheets outlining study aims, confidentiality measures, and voluntary participation rights. Written informed consent was obtained before data collection. Identifiers were removed from all datasets, which were stored in encrypted drives with restricted access. Any reference to AI use by participants was anonymized, and any generative-AI support used by the authors will be transparently disclosed in accordance with journal policy and academic-integrity principles.

RESULTS AND DISCUSSION

Results

This section presents findings from survey data (n = 120), semi-structured interviews (12 students and 8 lecturers), and document analysis of syllabi, rubrics, and anonymized student work. Each subsection corresponds to the study's research questions and integrates quantitative results, illustrative qualitative excerpts, and document-based evidence. Descriptive and inferential statistics are complemented with effect sizes and thematic syntheses. Supporting materials appear in Appendix Tables 1–7 and Appendix Figures 1–8. The analytic integration follows conventions for convergent mixed-methods reporting (Creswell & Plano Clark, 2018; Fetters et al., 2013).

RQ1. Metacognitive Awareness Patterns

Mean scores for metacognitive awareness were above the midpoint across all programs (Appendix Table 3; Figure 1). The Linguistics program reported the highest mean ($M = 4.08$, $SD = 0.42$), followed by Indonesian Language and Literature Education ($M = 3.88$, $SD = 0.47$). Management ($M = 3.54$, $SD = 0.53$) and Law ($M = 3.59$, $SD = 0.48$) showed moderate but stable values. A one-way ANOVA indicated significant differences, $F(3, 116) = 4.26$, $p < .05$, with a small-to-medium effect size ($\eta^2 = 0.09$).

Further item-level detail presented in Appendix Table 4 provides a finer-grained view of individual indicators of planning, monitoring, and evaluation. Although overall means remained above the scale midpoint (Appendix Tables 2–3), within-dimension variation suggests that certain regulatory behaviors, particularly evaluative reflection, were less consistently maintained across programs. These internal contrasts inform the qualitative interpretation that follows, clarifying how students' metacognitive control fluctuated under revision pressure.

Interview narratives confirmed the quantitative trend. Students described reliance on plan–monitor–evaluate routines that often broke down under heavy revision loads. One explained, "I outline my argument and check coherence, but repeated revisions make the focus shift" (S02–Indonesian Language & Literature). Lecturers noted similar lapses in sustained reflection under deadline pressure.

Document evidence (Appendix Tables 6–7; Figure 6) showed that Reflective Space was most explicit in the Indonesian Language & Literature syllabi (mean = 4.00), with other programs ranging 3.00–3.50. The triangulated findings highlight the need to institutionalize metacognitive reflection rather than leaving it to individual initiative.

RQ2–RQ3. Scaffolding Needs and Argumentation Quality

Survey data indicated notable differences in perceived writing difficulty (Appendix Table 3; Figure 2). Law (M = 3.45, SD = 0.51) and Management (M = 3.38, SD = 0.49) students reported greater difficulty than their peers in Linguistics and Indonesian Language & Literature (M = 3.05, SD = 0.45). Betwen-group differences produced a medium effect size (d = 0.64).

Qualitative results offered context for these disparities. Students described struggling to maintain the claim–evidence–warrant relationship through multiple revisions: "It is difficult to sustain a coherent argument after several feedback rounds" (S05–Law). Lecturers echoed this, observing that many students presented data without articulating the underlying reasoning chain.

Rubric analysis (Appendix Tables 6–7; Figure 7) showed strong Argument and Coherence criteria in Linguistics (4.50; 4.50) and Indonesian Language & Literature (4.50; 4.00), while Worked Examples were inconsistently present across programs. Management syllabi emphasized applied examples without explicit reasoning models, while Law rubrics prioritized citation accuracy over reasoning depth.

Survey means for Perceived Lecturer Scaffolding were moderate-to-high overall (M = 3.56), with Management (3.62) and Law (3.68) slightly higher (Appendix Table 3; Figure 3). Students expressed the need for explicit paragraph models: "I need examples that show claim, evidence, and reasoning so the standards are clear" (S09–Management). Lecturers confirmed the value of pre-writing activities to clarify warrants: "Rubrics help, but structured exercises make their reasoning visible" (L03–Linguistics).

Document data showed the richest Worked Examples in Management (mean = 4.50) and the least in Law (3.00) (Appendix Tables 6–7; Figure 7). These findings converge on the importance of modeling argument structure and providing iterative feedback loops to stabilize metacognitive control and improve argument quality.

RQ4. AI Integration and Ethical Considerations

Attitudes toward generative AI varied significantly across programs (Appendix Table 3; Figure 4). Management students showed the highest positive attitude (M = 3.82, SD = 0.40), followed by Law (M = 3.68, SD = 0.42), whereas Indonesian Language & Literature (M = 3.26, SD = 0.51) and Linguistics (M = 3.15, SD = 0.47) expressed more cautious views. ANOVA results revealed a large effect size ($\eta^2 = 0.14$, $p < .01$).

Qualitative insights indicated selective and ethical AI use. Students in professional programs described using ChatGPT to generate ideas or check coherence, emphasizing critical verification: "AI enriches ideas, but I always confirm validity and sources" (S11–Law). In contrast, humanities students voiced concerns about academic integrity and overreliance on automation.

Lecturers acknowledged AI's dual potential, as a reflective support tool and as a possible shortcut, calling for clear institutional guidance. Document analysis (Appendix Tables 6–7; Figure 6) revealed that only one syllabus explicitly addressed AI ethics, while the rest provided general statements on originality or plagiarism. Reflective-Ethics scores ranged from 3.00 to 4.00.

The synthesis across strands indicates that ethical AI engagement improves when learners are prompted to evaluate AI outputs within metacognitive cycles rather than simply adopting them. Structured disclosure guidelines at course level could transform AI from a mechanical assistant into a reflective scaffold.

Cross-RQ Synthesis: Convergence and Divergence

Patterns across data strands demonstrated both convergence and divergence (Appendix Figures 6–8). Quantitatively, Linguistics and Indonesian Language & Literature ranked highest in metacognitive awareness and argument coherence,

supported by explicit reflective components in syllabi. Qualitative data corroborated this, showing stronger student articulation of planning and evaluation strategies.

In contrast, Management and Law displayed stronger AI engagement and scaffolding perceptions but lower reflective consistency. Document analysis confirmed that professional programs emphasized output quality and citation compliance rather than internal reasoning processes.

Overall, the integrated pattern reveals that metacognitive regulation functions as the internal engine of argument quality, while AI scaffolding acts as an external amplifier that can enhance or displace regulation depending on ethical framing. Programs that embedded reflection routines and explicit reasoning models demonstrated better alignment between intended and experienced scaffolding.

Reliability Note

Pilot-stage reliability testing (Appendix Table 1) showed Cronbach's alpha values ranging from .71 to .84 across dimensions, indicating satisfactory internal consistency for exploratory use. Qualitative inter-rater reliability reached Cohen's $\kappa = 0.82$, classified as substantial agreement (Landis & Koch, 1977). These results confirm robustness and balance between quantitative and qualitative strands.

Discussion

Overview of Principal Findings

Findings reveal a coherent yet differentiated pattern across programs. Students demonstrated moderate-to-high metacognitive awareness but struggled to sustain planning–monitoring–evaluation cycles under revision pressure. This fragility was most visible during intense feedback rounds, suggesting that metacognitive control fluctuates when external demands accelerate (Appendix Tables 2–3; Figures 1 and 6). Program-level contrasts showed that Indonesian Language & Literature and Linguistics embedded reflection more explicitly within course structures, while Management and Law emphasized product-oriented quality controls rather than regulatory routines. Document evidence confirmed that explicit reflective space and paragraph-level exemplars strengthen argument structure, though their availability remains uneven across curricula (Appendix Tables 6–7; Figure 7).

The cross-strand synthesis therefore highlights that argument quality improves when planned scaffolds (rubrics, exemplars, and feedback routines) align with students' lived scaffolding experiences. This alignment determines whether metacognitive regulation functions as an internalized habit or remains an instructional aspiration.

Interpreting Tension in Light of Prior Work

Results reaffirm that metacognition serves as a practical lever for complex academic writing when regulatory phases are routinized rather than incidental (Stanton, 2021; Teng & Zhang, 2021). Yet the observed "slippage" of regulation under time constraints introduces theoretical tension with Bjork's notion of desirable difficulties (Bjork & Bjork, 2011). The idea that cognitive strain can deepen learning contrasts with students' accounts of metacognitive fatigue during extended revisions. This divergence suggests that difficulty is productive only when writers possess explicit reflective tools to interpret that difficulty as information, not as failure. In this sense, poorly scaffolded challenge becomes overload rather than learning opportunity.

Evidence that paragraph-level modeling supports claim–evidence–warrant reasoning aligns with Toulmin-based pedagogy and extends prior classroom findings by embedding exemplars directly within rubrics (Warren, 2010; Yang, 2023). The advantage of worked examples and self-explanation clarifies why concrete models, rather than abstract criteria, more reliably influence drafting decisions in authentic graduate contexts (Atkinson et al., 2000; Chi, 2009). These findings position paragraph exemplars as a mediating device that converts evaluation into metacognitive awareness.

Generative-AI Integration: Risk, Utility, and Reflection

Patterns of AI use reveal that generative tools now participate in students' metacognitive ecology rather than sit outside it. Classifying these uses into functional categories clarifies both opportunity and risk:

- Acceptable use – Idea generation and coherence checking.
  Students reported employing ChatGPT to expand topical ideas or review paragraph flow. When accompanied by critical verification, these uses complement the planning and monitoring stages of metacognition by offloading routine cognitive load while maintaining reflective control (Ahn & Alkhaqani, 2024; Mahapatra, 2024).
- Risky use – Plagiarism and source fabrication.
  Both lecturers and students identified ethical gray zones in overreliance on AI-generated evidence or paraphrases. Such automation risks eroding evaluative reasoning by substituting machine confidence for human judgment. Ethical frameworks emphasize explicit disclosure, verification of references, and awareness of algorithmic bias as safeguards (UNESCO, 2023, 2025; Chan, 2023).

- Reflective scaffolding – AI as metacognitive partner.
  Emerging practices showed students prompting AI to critique argument logic or to identify weak transitions, then using that feedback for revision. This behavior mirrors metacognitive monitoring and evaluation phases, suggesting that AI can operate as an externalized reflection partner when learners critically mediate its feedback (Xu et al., 2025).

Across programs, professional disciplines (Management and Law) exhibited higher openness to such reflective scaffolding, whereas language-based disciplines displayed caution rooted in authorship ethics. This disciplinary divergence reflects differing epistemic cultures: applied fields frame AI as pragmatic tool use, while humanities view it through authenticity and authorship lenses. Cross-disciplinary dialogue may therefore help normalize transparent AI engagement without compromising integrity.

Pedagogical and Policy Implications

Pedagogical implications center on embedding regulation routines and AI reflection within assessment design. Weekly checklists aligned with rubric criteria can operationalize the plan–monitor–evaluate cycle, ensuring students focus on reasoning moves rather than surface edits (Nicol & Macfarlane-Dick, 2006). Supervision meetings become more diagnostic when students complete short "micro-drills" on warranting and cohesion prior to consultation, allowing feedback to target paragraph logic rather than generic advice.

Rubric rows should include fully developed paragraph exemplars illustrating claim–evidence–warrant connections, paired with self-explanation prompts that ask writers to name their warrants explicitly. Empirical evidence links such exemplars and reflective prompts to stronger transfer from evaluative criteria to drafting decisions (Atkinson et al., 2000; Chi, 2009; Hattie & Timperley, 2007).

Policy frameworks can articulate clear AI guardrails through course-level statements detailing (1) acceptable purposes (idea generation, coherence checking), (2) minimal documentation or disclosure logs, and (3) verification steps to ensure factual reliability. Reflection prompts inviting students to record when AI assisted or disrupted reasoning translate policy language into teachable metacognitive practice (Ogunleye et al., 2024).

International guidance emphasizing human oversight and transparency supports viewing AI as an intellectual collaborator that complements, rather than replaces, human reasoning (UNESCO, 2023, 2025). Embedding reflective AI routines within writing courses thus promotes both ethical awareness and cognitive independence.

Cross-Disciplinary Targeting

Disciplinary comparisons offer tailored directions for instructional design. Management and Law programs require exemplars that emphasize evidence selection under uncertainty, warrant calibration for high-stakes claims, and sentence-level citation patterns that render source authorization explicit (Appendix Table 3; Figures 2 and 5).

Indonesian Language & Literature and Linguistics can capitalize on stronger reflective traditions by adding lightweight graded reflection checkpoints at interim deadlines and requiring students to label warrants directly within paragraph drafts (Appendix Tables 6–7; Figures 1 and 6).

Document analysis confirmed that reflective space and argument criteria alone are insufficient without sustained, course-embedded practice. A combined training package, comprising weekly checklists, worked paragraph exemplars, and structured feedback sessions, would help close the gap between intended and experienced scaffolding (Appendix Figures 6–8; Appendix Table 7).

Limitations and Future Research

Several methodological and measurement constraints should be acknowledged to contextualize interpretation. The study involved two anonymized universities and four master's programs, which limits generalizability across institutional and disciplinary boundaries. The sample size, while adequate for exploratory mixed-methods purposes, remains insufficient for confirmatory modeling or robust inferential testing. These boundary conditions position the findings as contextually informative rather than statistically generalizable (Creswell & Plano Clark, 2018).

Although internal consistency coefficients for several quantitative subscales fell below the conventional threshold of .70, these indicators were explicitly framed as exploratory rather than psychometrically definitive. Their purpose was to complement, not to validate, qualitative and documentary evidence by signaling directional tendencies across programs. This interpretive stance reflects early-stage scale development practice, where conceptual sensitivity is prioritized over reliability precision during instrument piloting (Hattie & Timperley, 2007).

Self-report data may have inflated perceived strategy use due to recall or social desirability bias, particularly regarding ethical AI engagement. Qualitative sampling, while diverse, was modest and constrained by postgraduate scheduling realities. Document ratings necessarily simplified complex instructional enactments into ordinal scales, which may not fully represent the richness of scaffolding practices observed in classrooms.

Triangulation mitigated these weaknesses by connecting quantitative patterns with interview narratives and documentary exemplars, yielding a multi-perspectival understanding of the same phenomena. Nonetheless, subsequent studies should refine the quantitative scales by adding homogeneous items, removing weak indicators, and examining construct validity through larger cross-institutional samples. Test–retest reliability and measurement invariance analyses are needed to establish temporal stability. Longitudinal designs could further clarify how metacognitive regulation, argument quality, and reflective AI use evolve across iterative writing cycles and disciplinary settings.

Future research may also experiment with classroom interventions that integrate reflective prompts, AI-use logs, and rubric–example pairings to test the efficacy of the proposed scaffolding model. Comparative and equity-focused inquiries would be particularly valuable in assessing how institutional context, digital access, and disciplinary epistemologies shape the ethical and cognitive dynamics of AI-supported writing.

Concluding Synthesis

The discussion underscores a productive tension between regulation stability and productive struggle. Graduate writers learn most when challenge is accompanied by structured reflection, and AI can amplify this balance when positioned as a reflective partner within the metacognitive cycle. The integration of metacognitive regulation, argument modeling, and ethical AI scaffolding forms a coherent framework for developing autonomous, critically reflective academic writers.

CONCLUSION

This convergent mixed-methods study examined baseline needs for metacognition-oriented academic writing supported by generative-AI scaffolding across four master's programs in two universities. Integrated evidence from surveys, interviews, and document analyses converged on three patterns. Students demonstrated moderate-to-high metacognitive awareness yet struggled to maintain the plan–monitor–evaluate cycle when facing intensive revision demands. Paragraph-level argumentation remained uneven, particularly in Management and Law, unless rubric criteria were accompanied by worked examples and feedback cycles prompting students to articulate warrants and transitions explicitly. Attitudes toward generative AI were cautiously positive, though ethical guidance and reflective opportunities varied considerably across programs. These findings suggest that argument quality depends on aligning intended scaffolding embedded in syllabi and rubrics with the scaffolding students actually experience during drafting and revision.

A strand-integrated scaffolding model is proposed to strengthen this alignment by:

- embedding short, graded reflective checkpoints linked to rubric rows;
- incorporating worked examples that operationalize full claim–evidence–warrant chains;
- implementing micro-drills prior to supervision meetings to surface reasoning choices; and
- formalizing AI-use disclosure and verification routines so that generative tools act as reflective companions rather than substitutes for reasoning.

These recommendations build on evidence that worked examples and self-explanation foster transfer (Atkinson et al., 2000; Chi, 2009), that formative feedback enhances regulation and performance (Hattie & Timperley, 2007; Nicol & Macfarlane-Dick, 2006), and that sustained metacognitive routines require deliberate institutionalization to endure authentic task pressures (Zimmerman, 2002; Teng et al., 2021). They also align with current global guidance emphasizing human oversight, transparency, and accountability in the educational use of generative AI (UNESCO, 2023, 2025).

Program-specific targeting emerges naturally from the data. Management and Law would benefit from exemplars emphasizing evidence selection under uncertainty, warrant calibration for consequential claims, and sentence-level citation patterns that clarify epistemic authority. Linguistics and Indonesian Language & Literature programs can leverage stronger reflective traditions by embedding lightweight, graded checkpoints at interim deadlines and requiring explicit warrant labeling within paragraph drafts to maintain argument coherence during revision.

Limitations include the focus on two institutions, modest interview samples, simplified document indicators, and pilot-stage quantitative scales exhibiting low internal consistency. Although several subscales scored below .70, they were interpreted as exploratory indicators complementing qualitative patterns. Future research should refine these scales through item–total analyses, remove weak items, and test the proposed intervention package experimentally, combining rubric-example pairing, self-explanation prompts, and feedback checkpoints, while examining generalizability, equity, and access in AI-mediated writing support. In summary, bridging the gap between intended and experienced scaffolding, while cultivating transparent and ethically aware AI practices, offers a pragmatic pathway toward more durable metacognitive regulation and stronger paragraph-level argumentation in graduate academic writing.

REFERENCES

Ahn, A., & Alkhaqani, A. (2024). Should academics be concerned about articles written by ChatGPT? *Developments in Health Sciences, 7*(1), 25–27. https://doi.org/10.1556/2066.2024.00062.

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked-examples research. *Review of Educational Research, 70*(2), 181–214. https://doi.org/10.3102/00346543070002181.

Bereiter, C., & Scardamalia, M. (Eds.). (1987). *The psychology of written composition* (1st ed.). Routledge. https://doi.org/10.4324/9780203812310.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa.

Chi, M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1*(1), 73–105. https://doi.org/10.1111/j.1756-8765.2008.01005.x.

Chinn, C. A., Buckland, L. A., & Samarapungavan, A. (2011). Expanding the dimensions of epistemic cognition: Arguments, values, and mechanisms. *Educational Psychologist, 46*(3), 141–167. https://doi.org/10.1080/00461520.2011.587722.

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE.

Fetters, M. D., Curry, L. A., & Creswell, J. W. (2013). Achieving integration in mixed methods designs: Principles and practices. *Health Services Research, 48*(6 Pt 2), 2134–2156. https://doi.org/10.1111/1475-6773.12117.

Guetterman, T. C., Fetters, M. D., & Creswell, J. W. (2015). Integrating quantitative and qualitative results in health science mixed methods research through joint displays. *The Annals of Family Medicine, 13*(6), 554–561. https://doi.org/10.1370/afm.1865.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487.

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.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199–218. https://doi.org/10.1080/03075070600572090.

Ogunleye, B., Zakariyyah, K. I., Ajao, O., Olayinka, O., & Sharma, H. (2024). A systematic review of generative AI for teaching and learning practice. *Education Sciences, 14*(6), 636. https://doi.org/10.3390/educsci14060636.

Rivas, S. F., Saiz, C., & Cornejo, C. O. (2022). Metacognitive strategies and development of critical thinking in higher education. *Frontiers in Psychology, 13*, 913219. https://doi.org/10.3389/fpsyg.2022.913219.

Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology, 19*(4), 460–475. https://doi.org/10.1006/ceps.1994.1033.

Stanton, J. D., Sebesta, A. J., & Dunlosky, J. (2021). Fostering metacognition to support student learning and performance. *CBE—Life Sciences Education, 20*(2), fe3. https://doi.org/10.1187/cbe.20-12-0289.

Teng, M. F., Qin, C., & Wang, C. (2021). Validation of metacognitive academic writing strategies and the predictive effects on academic writing performance in a foreign language context. *Metacognition and Learning, 17*(1), 167–190. https://doi.org/10.1007/s11409-021-09278-4.

UNESCO. (2023). *Guidance for generative AI in education and research.* UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000386693.

UNESCO. (2025). *Guidance for generative AI in education and research (updated brief).* UNESCO. https://www.unesco.org/en/articles/guidance-generative-aieducation-and-research.

Warren, J. E. (2010). Taming the warrant in Toulmin's model of argument. *Teaching English in the Two-Year College, 38*(2), 170–186.

Yang, R. (2023). Whole-to-part argumentation instruction: An action research study aimed at improving argumentative writing based on the Toulmin model. *SAGE Open, 13*(4), 1–13. https://doi.org/10.1177/21582440231207738.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice, 41*(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2.