International Journal of Learning, Teaching and Educational Research
Vol. 25, No. 8, pp. 946-967, August 2026
https://doi.org/10.26803/ijlter.25.8.37
Received Jul 5, 2026; Revised Aug 4, 2026; Accepted Aug 12, 2026

Development and Preliminary Validation of the ARROW-AI Framework for Responsible AI-Assisted Research Writing in Higher Education

Luijim S. Jose* **and Reynaldo A. Cabual**
Nueva Ecija University of Science and Technology
Cabanatuan City, Philippines

Abstract. Generative artificial intelligence is increasingly used in higher-education research writing, yet institutions often lack operational procedures to protect authorship, ensure research alignment, verify, and disclose. This study developed and preliminarily evaluated the ARROW-AI Framework and Toolkit through a sequential multiphase developmental-validation design with embedded mixed-methods formative evaluation and a bounded pilot. Participants comprised 12 experts, 84 student-researchers, 18 advisers/instructors, and 32 pilot cases. Descriptive expert-domain ratings ranged from 3.38 to 3.58 for the framework and from 3.20 to 3.53 for the toolkit, while all 15 components met the prespecified I-CVI and CVR criteria. Student-researcher domain ratings ranged from 4.05 to 4.18, and adviser/instructor ratings ranged from 3.82 to 4.24; documentation and workload burden were moderate. Pilot completion averaged 88.91%; six initially high-risk requests were down-scoped, and four cases required substantive revision. Because the instruments covered conceptually distinct formative domains, ratings were interpreted descriptively by domain rather than as scores from unidimensional psychometric scales. Integrated evidence supported six revisions, including shorter records, completed examples, clearer risk decisions, and stronger adviser guidance. ARROW-AI should therefore be treated as a developmental process guide rather than evidence of effectiveness or misconduct prevention.

Keywords: academic integrity; generative artificial intelligence; higher education; research writing; responsible AI use

Citation:
Jose, L. S., & Cabual, R. A. (2026). Development and Preliminary Validation of the ARROW-AI Framework for Responsible AI-Assisted Research Writing in Higher Education. *International Journal of Learning, Teaching and Educational Research, 25*(8), 946–967.
https://doi.org/10.26803/ijlter.25.8.37

*Corresponding author: Luijim S. Jose; *luijimjose@ineust.ph.education*

1. Introduction

Generative artificial intelligence (AI) can support ideation, outlining, revision, feedback, and source-search preparation, but it can also produce false claims, fabricated citations, privacy risks, over-reliance, and opaque authorship. UNESCO emphasizes human agency, transparency, data protection, and pedagogically grounded use (Miao & Cukurova, 2024; Miao & Holmes, 2023). Students and faculty similarly report educational benefits alongside concerns about accuracy, integrity, and the need for explicit guidance (Chan & Hu, 2023; Chiu, 2024; Rajabi et al., 2024). These tensions are especially consequential in research writing, where claims must remain traceable to methods, data, evidence, and human judgment.

AI assistance is not inherently unethical, yet students may overlook the integrity implications of its use in assessment (Gruenhagen et al., 2024). Academic integrity requires honesty, trust, fairness, respect, responsibility, and courage (International Center for Academic Integrity, 2021). Publication guidance likewise excludes AI from authorship and places responsibility for accuracy, attribution, verification, and disclosure on human authors (Committee on Publication Ethics, 2023; International Committee of Medical Journal Editors, 2026). Research-writing instruction, therefore, requires operational procedures, not merely permission or prohibition. Fluent AI output may still be false or unverifiable: systems fabricate references (Walters & Wilder, 2023), while detectors produce uncertain judgments and may disadvantage non-native English writers (Liang et al., 2023; Perkins et al., 2024). Responsible practice should instead require prospective task planning, source and claim verification, adviser review, and transparent disclosure.

This study defines responsible AI-assisted research writing as documented, human-controlled, ethically bounded, and verifiable support that does not transfer authorship, weaken methodological alignment, bypass verification, or conceal assistance. ARROW-AI comprises Authorial Control, Research Alignment, Risk-Stratified AI Use, Output Verification, and Written Disclosure and Defense. Its ten-tool workflow supports planning, review, verification, disclosure, and defense; it is not a detector, disciplinary instrument, or effectiveness-tested intervention. These operational boundaries synthesize guidance on publication, education, and research governance (Committee on Publication Ethics, 2023; Miao & Holmes, 2023; Smith et al., 2026).

ARROW-AI's proposed contribution is a research-writing-specific sequence that converts responsible-AI principles into auditable writer–adviser decisions. The study asked: (1) How was ARROW-AI developed? (2) What expert-validation and content-validity evidence emerged? (3) How did student-researchers and advisers/instructors evaluate it, and what occurred during the pilot? (4) What integrated strengths, limitations, risks, and revisions produced the revised framework? No confirmatory hypotheses were tested. The contribution was positioned against current policy, assessment, and research-governance frameworks (O'Sullivan et al., 2025; Perkins et al., 2025; Smith et al., 2026).

2. Literature Review

Generative AI has shifted higher education from simple debates over prohibition toward pedagogical purpose, governance, integrity, and accountable use. Reported benefits include feedback, ideation, and language support, whereas documented risks include fabrication, loss of privacy, overreliance, inequity, and uncertain authorship (Bittle & El-Gayar, 2025; Chiu, 2024; Cotton et al., 2024; Crompton & Burke, 2023; Rajabi et al., 2024). The literature is therefore not uniformly affirmative: perceived usefulness does not resolve questions of authorship, evidence quality, or disclosure. In research writing, these risks threaten traceability to the approved problem, method, evidence, and human judgment.

AI literacy and human agency underpin Authorial Control. UNESCO requires ethical judgment, critical evaluation, and responsible design (Miao & Cukurova, 2024; Miao et al., 2024), while the Higher Education Authority foregrounds integrity, transparency, oversight, privacy, and sustainable pedagogy (O'Sullivan et al., 2025). These principles keep intellectual decisions, the use of evidence, and the final text under human responsibility.

Academic integrity and publication guidance make this responsibility non-transferable: AI cannot qualify as an author, and humans must verify, attribute, and disclose AI-assisted work (Committee on Publication Ethics, 2023; International Center for Academic Integrity, 2021; International Committee of Medical Journal Editors, 2026). Research guidance further links responsible use to training, communication, infrastructure, and process change (Smith et al., 2026).
Research Alignment requires AI activity to remain consistent with the approved problem, questions, design, data, and warranted claims. Assessment-reform frameworks similarly emphasize authentic capability, process evidence, and alignment between permitted AI use and intended outcomes (Perkins et al., 2025; Tertiary Education Quality and Standards Agency, 2025). Polished text is still misaligned when it changes the study or exceeds its evidence.

Risk-Stratified AI Use provides a proportionate alternative to blanket rules. Current frameworks differentiate AI involvement and recommend context-sensitive permissions (O'Sullivan et al., 2025; Perkins et al., 2025; Smith et al., 2026). ARROW-AI classifies tasks before use and requires down-scoping when a request could transfer authorship, fabricate evidence, or compromise privacy. Output Verification is necessary because fluency is not reliability. Generative systems fabricate citations (Walters & Wilder, 2023), and detectors cannot replace source checking or academic judgment (Liang et al., 2023; Perkins et al., 2024). Verification, therefore, traces sources and evidence and records whether generated material was retained, revised, or rejected.

Written Disclosure and Defense extends transparency beyond a brief acknowledgment. Governance frameworks call for evolving institutional processes (Pinho et al., 2025; Smith et al., 2026); ARROW-AI adds writer–adviser records of the task, output, human revision, verification, and rationale that the researcher must explain.

Table 1 compares selected competency, policy, assessment, governance, and research frameworks with ARROW-AI.

Methodologically, the study combined educational design research, developmental validation, instrument-development guidance, framework-guided qualitative content analysis, and mixed-methods integration (DeVellis & Thorpe, 2021; Fetters & Tajima, 2022; Mayring, 2022; McKenney & Reeves, 2018; Richey & Klein, 2007). Validity was treated as evidence supporting bounded interpretations and uses, not as a permanent property of an instrument, while acceptability and feasibility were distinguished from effectiveness (American Educational Research Association et al., 2014; Kane, 2013; Proctor et al., 2011; Yusoff, 2019). This distinction is central to the present study because favorable expert or user ratings cannot establish improved writing, reduced misconduct, or institutional readiness.

Table 1: Comparison of selected responsible-AI frameworks and ARROW-AI

Framework or guidance	Primary scope	ARROW-AI components	Operational gap or distinction
UNESCO Guidance (Miao & Holmes, 2023)	Human-centered, ethical, safe educational use	AC, RS, OV, WDD	Broad guidance; no linked research-writing records.
UNESCO competency frameworks (Miao & Cukurova, 2024; Miao et al., 2024)	Student and teacher AI competencies	AC, OV	Competency focus; no case-level adviser workflow.
AI Assessment Scale (Perkins et al., 2025)	Permitted AI involvement in assessment	AC, RS	Assessment-centered; limited source and defense tracing.
HEA Policy Framework (O'Sullivan et al., 2025)	Institutional policy, literacy, integrity, privacy	AC, RS, OV, WDD	System-level; no linked writer–adviser forms.
Living GenAI Governance Model (Pinho et al., 2025)	Dynamic governance across research levels	AC, OV, WDD	Governance model; not a task-level toolkit.
University research framework (Smith et al., 2026)	Principles and institutional research support	AC, OV, WDD	Institutional framework; limited writing workflow.
ARROW-AI Framework and Toolkit	Research-writing process and auditable writer–adviser workflow	AC, RA, RS, OV, WDD	Five components linked to ten tools and preliminary validation.

Note. AC = Authorial Control; RA = Research Alignment; RS = Risk-Stratified AI Use; OV = Output Verification; WDD = Written Disclosure and Defense.

Existing guidance operates mainly at competency, policy, assessment, or governance levels. ARROW-AI's bounded originality is a linked research-writing workflow for authorial control, alignment, risk classification, verification, disclosure, and defense supported by ten reviewable records—not evidence of improved writing or misconduct prevention.

3. Methodology

3.1 Research Design

The study used a sequential multiphase developmental-validation design with embedded mixed-methods formative evaluation and a bounded pilot. This design supported iterative development, expert review, user evaluation, implementation testing, and evidence-based revision rather than causal testing (McKenney & Reeves, 2018; Richey & Klein, 2007). Quantitative evidence summarized validation, perceptions, burden, and implementation; qualitative evidence explained strengths, risks, and revision needs; and integration linked evidence directly to redesign decisions.

Five phases produced linked outputs: (1) conceptual mapping, (2) framework and tool development, (3) expert and content validation, (4) user review and bounded pilot implementation, and (5) evidence integration and final refinement. Effects on writing quality, misconduct, defense performance, publication readiness, and institutional adoption were outside the scope of the design. This phase structure permits replication of the developmental process while preserving the distinction between formative evaluation and effectiveness testing.

3.2 Research Setting, Participants, and Sampling

The study was conducted over six weeks during Academic Year 2025–2026 at a public university in Central Luzon, Philippines, which offers undergraduate and graduate research-writing courses. Participants were 12 experts, 84 student-researchers, and 18 advisers/instructors; 32 pilot cases generated 320 tool audits and 180 qualitative meaning units. Experts were selected through purposive expert sampling and were required to hold a relevant advanced qualification, have at least 5 years of experience, and demonstrate publication, validation, supervision, measurement, ethics, or institutional policy work. The panel comprised three methodologies, two AI/educational-technology, two integrity/publication-ethics, two measurements, and three research-writing/supervision specialists.

All experts reviewed the content; three additionally reviewed the language, and two additionally reviewed the layout/usability. Student-researchers were adults engaged in research writing, had prior academic exposure to generative AI, and consented to tool completion; 52 were undergraduates, and 32 were graduate students. The adviser/instructor group comprised 10 advisers, 5 research instructors, and 3 panel evaluators with at least 3 years of relevant experience. Pilot cases varied by study level, task, and initial risk. Purposive and convenience sampling sought information-rich developmental feedback rather than population representation.

3.3 ARROW-AI Framework and Toolkit

ARROW-AI synthesized responsible-AI guidance, publication ethics, academic integrity, research alignment, verification of literature, and supervision practice into five sequential components. Table 2 summarizes their functions.

Table 2: ARROW-AI components and operational functions

Component	Operational function
Authorial Control	Retains human responsibility for decisions, claims, evidence, and final text.
Research Alignment	Checks consistency with the approved problem, questions, design, data, and evidentiary limits.
Risk-Stratified AI Use	Classifies tasks and restricts or down-scales authorship, evidence, privacy, or scholarly-work risks.
Output Verification	Requires source tracing, factual review, and documented retention, revision, or rejection.
Written Disclosure and Defense	Records AI assistance and prepares the researcher to explain and defend decisions.

Ten tools operationalized planning, execution, verification, supervision, disclosure, and defense: the AI Use Agreement, AI Task Risk Matrix, Prompt Planning Sheet, Prompt–Output–Revision Log, Human Defense Checklist, Research Alignment Matrix, Output Verification Ledger, Adviser AI Review Form, AI Disclosure Statement Template, and Final AI-Assisted Manuscript Audit Checklist.

3.4 Instruments

The instruments were mapped to the five components, intended users, and bounded interpretations and were reviewed for alignment, wording, burden, and usability following instrument-development guidance (DeVellis & Thorpe, 2021). Expert instruments included a 10-item profile form, a 28-item framework form, a 25-item toolkit form, and a 45-judgment content-validity sheet covering relevance, clarity, and essentiality. User and implementation instruments included a 24-item student-researcher survey, a 20-item adviser/instructor survey, a six-prompt feedback form, a 12-indicator observation form, an eight-indicator audit guide, and coding/integration records. Experts used a four-point scale without a neutral midpoint because ratings of 3 or 4 represented content-validity agreement. Student-researchers and advisers/instructors used five-point scales because neutral perceptions were plausible after limited exposure. Burden domains were analyzed separately because higher scores represented less favorable documentation or workload experiences.

3.5 Data-Gathering Procedure

Phase 1 mapped the responsible AI, integrity, publication, research writing, and methodological requirements to components and tools. Phase 2 drafted the framework, removed redundancy, and clarified claim boundaries. In Phase 3, all experts independently rated the framework and toolkit, judged essentiality, and supplied item-level comments. Values below the prespecified criteria or recurring concerns triggered revision; expert-profile data were used to distinguish content, language, and layout observations.

In Phase 4, participants received a 45-minute orientation, and 32 cases completed a four-week pilot, each using one approved research-writing task. Initially, high-risk requests were reviewed with an adviser, down-scoped, and not implemented as originally proposed. In Phase 5, ratings, comments, observations, and audit

records were integrated in a joint evidence-to-revision matrix; a serious concern or convergence across at least two evidence sources prompted revision (Fetters & Tajima, 2022). This procedure created an auditable link between evidence and each change in Version 2.0.

3.6 Research Questions, Data Sources, and Analyses

Table 3 aligns each revised research question with its evidence and analysis.

Table 3: Research questions, principal data sources, and analysis procedures

Research question	Principal evidence	Analysis
RQ1: How was ARROW-AI developed?	Guidance map, design, and revision records	Developmental mapping and process analysis
RQ2: What validation evidence emerged?	Expert forms, content-validity sheet, comments	Descriptive statistics, I-CVI, CVR, comment analysis
RQ3: How was ARROW-AI evaluated and piloted?	User surveys, observations, and tool audits	Descriptive/burden summaries, sensitivity analysis, coding
RQ4: What integrated revisions resulted?	All quantitative, qualitative, and pilot evidence	Content analysis and joint evidence-to-revision integration

3.7 Scoring and Data Analysis

Descriptive statistics included valid n, missing n, mean, standard deviation, median, and range. The instruments were designed as multidimensional formative-evaluation forms composed of conceptually distinct domains rather than as unidimensional psychometric scales. Domain means were therefore interpreted separately as descriptive summaries of participant judgments; no total or omnibus score was treated as a latent-scale measure. Burden domains were analyzed separately because higher scores represented less favorable documentation or workload experiences. No inferential tests were used.

I-CVI was the proportion of 12 experts rating relevance or clarity as 3 or 4; .83 was the retention criterion. CVR used $(ne - N/2)/(N/2)$, with .56 as the 12-expert critical value; lower values prompted review rather than automatic deletion (Ayre & Scally, 2014; Lawshe, 1975; Yusoff, 2019).

Internal-consistency coefficients were not used to support score interpretation because the instruments served as multidimensional formative-evaluation forms rather than as measures of a single underlying construct. Framework and toolkit revisions were instead based on the convergence of content-validity indices, domain-level descriptive ratings, expert and user comments, qualitative findings, and bounded-pilot records. This analytical decision restricted the numerical findings to descriptive interpretations of the individual domains.

Qualitative records underwent framework-guided content analysis using the five components, along with inductive codes (Mayring, 2022). Two coders calibrated 36 meaning units; one coded the corpus, and the second reviewed ambiguous or high-consequence units. Disagreements were resolved by documented consensus.

A joint display examined convergence, complementarity, dissonance, and gaps across quantitative, qualitative, and pilot evidence (Fetters & Tajima, 2022). Sensitivity analysis excluded initially high-risk requests that had been down-scoped.

3.8 Ethical Considerations

Written consent, voluntary participation, withdrawal before de-identification, separate encrypted identifiers, and protection from academic consequences were specified. Survey responses were withheld from evaluators until grading or decision milestones had passed. Writing samples, prompts, outputs, and logs were treated as sensitive records, and participants were instructed not to upload identifiable or confidential data. The developer-researcher role, independent review, unfavorable findings, and claim limits were disclosed. Records were scheduled for secure disposal after five years.

4. Results and Findings

Findings follow the four research questions and remain developmental; they do not establish causal effectiveness, reduced misconduct, improved writing or defense performance, or institutional adoption.

4.1 Framework Development

Source mapping linked authorship, research alignment, risk classification, verification, and disclosure requirements to five components. Prototype review produced ten tools and explicit claim boundaries; expert, user, and pilot evidence subsequently generated six Version 2.0 revisions. This sequence answers RQ1 by showing how literature-derived requirements were translated into tools and subsequently modified based on empirical formative evidence.

4.2 Sample Flow and Data Integrity

Table 4 reports the analysis populations: 12 experts, 84 student-researchers, 18 advisers/instructors, 32 pilot cases, 320 audits, and 180 qualitative meaning units. No record was excluded for consent, completeness, duplication, range, or linkage problems. The complete case flow supports internal traceability of the descriptive analyses, but the purposive, single-institution samples remain unsuitable for prevalence estimates or population generalization.

Table 4: Sample flow and analysis populations

Analysis population	Included n	Missing n	Excluded n
Expert validators	12	0	0
Student-researchers	84	0	0
Advisers/instructors	18	0	0
Bounded pilot cases	32	0	0
Toolkit-audit records	320	0	0
Qualitative meaning units	180	0	0

Experts represented methodology (3), AI/educational technology (2), integrity/publication ethics (2), measurement (2), and research writing/supervision (3), with 6–24 years of experience. Students included 52 undergraduates and 32 graduate students; the adviser/instructor group included

10 advisers, five instructors, and three panel evaluators. This role diversity strengthened formative coverage of content, supervision, and implementation issues, although it did not remove the limitations of small expert and adviser samples.

4.3 Expert Validation of the Framework and Toolkit

Table 5 presents the descriptive ratings for the individual framework domains. Domain means ranged from 3.38 to 3.58 on the four-point scale. Originality received the highest descriptive rating (M = 3.58), whereas relevance received the lowest (M = 3.38). Every domain mean exceeded 3.00, supporting retention and further refinement of the framework structure. These domain ratings were interpreted as expert judgments on distinct evaluation criteria rather than as components of a single psychometric score. Expert comments were also used to clarify purpose, research alignment, privacy safeguards, and claim boundaries.

Table 5: Framework expert-validation descriptive statistics

Domain	n	M	SD
Clarity	12	3.42	0.42
Relevance	12	3.38	0.43
Originality	12	3.58	0.36
Ethical adequacy	12	3.42	0.42
Research-integrity protection	12	3.46	0.40
Usability	12	3.50	0.30
Feasibility	12	3.42	0.47

Table 6 presents the descriptive ratings for the individual toolkit domains. Domain means ranged from 3.20 to 3.53. Operational usefulness (M = 3.53) and coherence with the framework (M = 3.50) received stronger descriptive ratings than documentation adequacy (M = 3.20) and implementation feasibility (M = 3.22). This contrast indicates that the experts valued the toolkit's operational logic while identifying concerns about documentation requirements and implementation demands. The toolkit was therefore retained but streamlined through completed examples, shorter fields, and a clearer adviser workflow.

Table 6: Toolkit expert-validation descriptive statistics

Domain	n	M	SD
Operational usefulness	12	3.53	0.33
Coherence with the framework	12	3.50	0.33
Scoring clarity	12	3.42	0.43
Documentation adequacy	12	3.20	0.36
Implementation feasibility	12	3.22	0.48

4.4 Content-Validity Evidence

Table 7 shows that all 15 framework and toolkit components met the .83 I-CVI retention criterion. Relevance I-CVI ranged from .83 to 1.00 (M = .97), clarity from .83 to 1.00 (M = .93), and CVR from .67 to 1.00 (M = .94), above the .56 critical value used for 12 experts. These indices support item retention and targeted revision; they do not establish construct validity, internal consistency, causal effectiveness, or the instrument's permanent quality.

Table 7: Content-validity evidence

Framework component/tool	Relevance I-CVI	Clarity I-CVI	CVR	Decision
Authorial Control	1.00	1.00	1.00	Retain
Research Alignment	1.00	.92	1.00	Retain; examples
Risk-Stratified AI Use	1.00	.92	1.00	Retain; boundaries
Output Verification	1.00	1.00	1.00	Retain
Written Disclosure and Defense	.92	.92	.83	Retain; expand
AI Use Agreement	1.00	1.00	1.00	Retain
AI Task Risk Matrix	1.00	.92	1.00	Retain; notes
Prompt Planning Sheet	1.00	.92	1.00	Retain; example
Prompt-Output-Revision Log	.92	.83	.83	Revise for brevity
Human Defense Checklist	1.00	1.00	1.00	Retain
Research Alignment Matrix	1.00	1.00	1.00	Retain
Output Verification Ledger	.92	.83	.83	Simplify fields
Adviser AI Review Form	1.00	.92	1.00	Retain; add guidance
AI Disclosure Statement Template	1.00	1.00	1.00	Retain
Final Manuscript Audit Checklist	.83	.83	.67	Retain with revision

The Final AI-Assisted Manuscript Audit Checklist had the lowest indices (.83 relevance, .83 clarity, .67 CVR). Because it still met the prespecified criteria and experts considered it essential, it was retained but shortened and paired with a completion example. The decision illustrates how quantitative thresholds and qualitative comments were combined rather than using a mechanical retain-or-delete rule.

4.5 Student-Researcher and Adviser/Instructor Perceptions

Table 8 presents the separate student-researcher domain ratings. The non-burden domain means ranged from 4.05 to 4.18. Acceptability received the highest descriptive rating (M = 4.18), whereas usability received the lowest (M = 4.05). Documentation burden was moderate (M = 2.99). These findings indicate that student-researchers generally valued the framework's safeguards but did not regard the documentation process as effortless. The findings support simplification and risk-proportionate documentation rather than universal use of the full set of forms.

Table 8: Student-researcher survey results

Domain	n	M	SD
Usability	84	4.05	0.35
Relevance	84	4.16	0.34
Acceptability	84	4.18	0.35
Verification support	84	4.12	0.38
Disclosure and defense	84	4.10	0.36
Documentation burden	84	2.99	0.37

Table 9 presents the separate adviser/instructor domain ratings. The non-burden domain means ranged from 3.82 to 4.24. Output-verification support received the highest descriptive rating (M = 4.24), whereas feasibility received the lowest (M = 3.82). Workload burden was moderate (M = 3.04). The contrast between usefulness and feasibility indicates that implementation requires orientation, completed examples, role clarity, and workload safeguards, rather than assuming that favorable perceptions will automatically translate into sustained adoption.

Table 9: Adviser/instructor survey results

Domain	n	M	SD
Usefulness	18	4.22	0.35
Feasibility	18	3.82	0.37
Supervision support	18	4.21	0.34
Output verification	18	4.24	0.39
Workload burden	18	3.04	0.41

4.6 Bounded Pilot Implementation

Table 10 shows a mean toolkit-completion rate of 88.91% (SD = 8.20), with an average of 2.22 missing required fields, 72.44 minutes of task time, and 0.88 implementation issues per case; four cases required substantive revision. The sensitivity estimates changed little after initially high-risk requests were excluded, suggesting that the main descriptive pattern was not driven solely by those cases. High completion nevertheless remained incomplete and should not be equated with accurate disclosure or effective writing.

Table 10: Bounded pilot implementation summary

Pilot metric	n	M	SD	Median	Range	Sensitivity result
Toolkit completion (%)	32	88.91	8.20	90.00	70.00–100.00	89.23
Missing required fields	32	2.22	1.64	2.00	0.00–6.00	2.15
Time on task (minutes)	32	72.44	17.13	74.00	46.00–110.00	71.58
Implementation issues	32	0.88	0.87	1.00	0.00–3.00	0.81

Nine requests were initially low-risk, 17 moderate-risk, and six high-risk. All six high-risk requests were down-scoped, and none were implemented as originally proposed. Advisers accepted 12 cases unchanged, 16 after minor revisions, and 4 after substantive revisions. Defense-checklist status changed from 20 ready, eight needing clarification, and four needing revisions to 28 ready and four needing minor clarification. These records show a documented review process and preparedness changes, not evidence that oral-defense performance improved.

4.7 Measurement Interpretation

The expert and user instruments covered conceptually distinct formative-evaluation domains and were not intended to produce a single latent score. Consequently, ratings were analyzed and interpreted separately at the domain level. No total or omnibus score was treated as an established psychometric measure, and internal-consistency coefficients were not used as evidence for validation. Content-validity indices, domain-level descriptive ratings, qualitative findings, and bounded-pilot records served as the principal evidence for retaining and revising the framework and toolkit. These boundaries mean that the numerical results describe participant judgments within the present developmental study and do not establish structural validity or psychometric scale quality.

4.8 Qualitative Findings

The qualitative dataset comprised 180 meaning units derived from expert comments, student-researcher feedback, adviser/instructor feedback, observation notes, and toolkit-audit records. Each meaning unit was assigned to one of the following primary categories: strength, weakness, revision suggestion, or risk. As shown in Table 11, 87 meaning units were categorized as strengths, 37 as weaknesses, 30 as revision suggestions, and 26 as risks. These four primary-category counts sum to the complete corpus of 180 meaning units.

The frequent codes presented beneath the primary categories represent recurring issues identified within the corpus. They are subordinate code frequencies and should not be added to the primary-category totals. Calibration coding was conducted on 36 meaning units and yielded 88.89% agreement and a Cohen's κ of .82. Coding disagreements were reviewed and resolved through documented consensus. The frequencies describe the composition of the qualitative corpus and

the relative prominence of revision priorities; they should not be interpreted as prevalence estimates for a wider population.

Table 11: Qualitative categories and frequent codes

Classification	Category or frequent code	Count
Primary category	Strength	87
Primary category	Weakness	37
Primary category	Revision suggestion	30
Primary category	Risk	26
Frequent code	Useful source-verification process	37
Frequent code	Clear human authorship responsibility	28
Frequent code	Documentation can feel lengthy	23
Frequent code	Helpful risk classification	22
Frequent code	Prompt log needs examples	14
Frequent code	Simplify output-verification fields	14
Frequent code	Students may underreport AI use	13
Frequent code	Adviser workload may increase	13
Frequent code	Add allowed/restricted-use examples	9
Frequent code	Add orientation checklist	7

The qualitative findings identified verification and human authorship responsibility as the most frequently reported strengths. At the same time, participants raised concerns about the length of documentation, incomplete disclosure, and additional workload for advisers. Revision-oriented comments particularly emphasized the need for completed examples, simpler verification fields, clearer permitted and restricted uses, and structured orientation.

De-identified excerpts illustrate these findings. Student-researcher S17 reported that the Output Verification Ledger prompted source checking, explaining, "It reminded me to check whether the source actually supported the AI-generated claim." Adviser/instructor A06 similarly indicated that the AI Task Risk Matrix supported revision of a proposed task before AI use, stating, "The matrix showed that the original task needed clearer limits before students used AI."

However, participants also identified areas requiring simplification and clarification. Student-researcher S42 described full documentation as repetitive for minor editing activities, noting, "For simple grammar editing, recording every step felt repetitive and unnecessary." Expert E09 requested clearer documentation of the respective contributions of AI and the human author, emphasizing, "The form should state what AI contributed and what the human author personally verified."

These excerpts help explain why the framework's verification and accountability requirements were retained, while the documentation forms, examples, disclosure prompts, and adviser procedures were revised to improve clarity, proportionality, and practical usability.

4.9 Mixed-Methods Integration and Revision Decisions

Quantitative ratings, pilot records, and qualitative findings were integrated through an evidence-to-revision matrix. The purpose of the integration was not to determine effectiveness but to identify where evidence converged, complemented other findings, revealed implementation gaps, or required caution. Table 12 uses the same qualitative-code labels and frequencies reported in Table 11 to maintain traceability across the findings.

Convergent evidence supported retaining Authorial Control, Output Verification, and Risk-Stratified AI Use. Participants valued these functions, and the qualitative findings emphasized human responsibility, source verification, and pre-use risk classification. In contrast, documentation burden and adviser workload showed convergent concerns. Prompt logging and the verification ledger also demonstrated implementation gaps that required completed examples and simplified fields.

Table 12: Mixed-methods evidence-to-revision matrix

ARROW-AI area	Quantitative or pilot evidence	Qualitative evidence	Integrated interpretation	Revision decision
Authorial Control	Framework domain means ranged from 3.38 to 3.58	Clear human authorship responsibility (28)	Convergent support for retaining explicit human responsibility	Retain
Output Verification	Student verification support M = 4.12; adviser output-verification support M = 4.24	Useful source-verification process (37)	Convergent support, with a need for clearer application examples	Add completed examples
Risk-Stratified AI Use	Six initially high-risk requests were down-scoped before implementation	Helpful risk classification (22)	Bounded evidence that the risk matrix supported pre-use review and revision	Add risk-decision notes
Written Disclosure and Defense	Student disclosure-and-defense rating M = 4.10	E09 requested clearer disclosure of AI contribution and human verification; this comment was not separately quantified	Favorable perception was complemented by a specific need for clearer disclosure guidance	Expand disclosure and defense prompts
Documentation burden	Student documentation-burden rating	Documentation can feel lengthy (23)	Convergent concern regarding the	Shorten and proportion forms

	M = 2.99; adviser workload-burden rating M = 3.04		amount of documentation required	according to task risk
Prompt logging	Toolkit completion was high but incomplete, and required fields were sometimes missing	Prompt log needs examples (14)	Pilot records and qualitative feedback identified a need for a completed model	Add a completed prompt-log exemplar
Output Verification Ledger	Toolkit implementation feasibility (M = 3.22) was lower than operational usefulness (M = 3.53)	Simplify output-verification fields (14)	The tool was considered useful but required greater usability and brevity	Shorten and simplify the ledger
Adviser workflow	Adviser workload burden was moderate (M = 3.04)	Adviser workload may increase (13)	Implementation requires role clarity, scheduling, and workload safeguards	Add an adviser workflow checklist
Misconduct prevention	Misconduct reduction was not a predefined or measured outcome	No qualitative evidence established misconduct prevention	Evidence gap; no effectiveness or prevention claim is warranted	Make no misconduct-prevention claim

The integrated findings supported retaining the five-component structure and all ten operational tools while making six principal revisions to the toolkit. Version 2.0 added permitted and restricted-use examples, risk-decision notes, a completed Prompt–Output–Revision Log, a shorter Output Verification Ledger, expanded disclosure and defense prompts, and an adviser workflow checklist. These revisions responded directly to identifiable quantitative, qualitative, and pilot evidence rather than to favorable ratings alone.

The integration also established clear claim boundaries. The evidence supports preliminary conclusions regarding developmental adequacy, perceived usefulness, content validity, usability, burden, feasibility, and revision priorities. It does not demonstrate improved writing quality, improved oral defense performance, reduced academic misconduct, scalability, or institutional adoption. Because Version 2.0 contains substantively revised tools, its comprehension, burden, score behavior, and implementation feasibility require renewed validation before wider use.

5. Discussion

5.1 Development, Validation, and Originality

Expert judgments, content-validity evidence, and integrated revision supported retaining five components and ten tools with six modifications. The defensible contribution is a preliminarily content-validated and formatively evaluated research-writing architecture developed through mapping, expert review, bounded implementation, and evidence integration—not an effectiveness-tested intervention. In educational design-research terms, the contribution lies in a context-responsive, revisable solution and its documented design logic (McKenney & Reeves, 2018; Richey & Klein, 2007).

Favorable domain-level ratings support bounded interpretations of clarity, relevance, developmental adequacy, usefulness, and feasibility; they do not confer permanent validity or establish the quality of a psychometric scale (American Educational Research Association et al., 2014; Kane, 2013). The result converges with UNESCO and institutional frameworks that emphasize human agency, transparency, verification, and accountability (Miao & Cukurova, 2024; O'Sullivan et al., 2025; Smith et al., 2026). It differs from broad policy guidance by connecting those principles to task-level records and writer–adviser decisions.

ARROW-AI therefore operationalizes established principles rather than proposing a new ethical doctrine. Its bounded originality is the linked sequence of research alignment, pre-use risk classification, output verification, adviser review, written disclosure, and defense. Table 1 shows that the selected competency, assessment, policy, and governance frameworks address several of these elements, but not the same integrated research-writing workflow. This comparison demonstrates novelty while avoiding an unsupported claim that the underlying ethical principles are unique.

The framework's process orientation also contrasts with post hoc detection. Detection tools are uncertain and may disadvantage non-native English writers (Liang et al., 2023; Perkins et al., 2024), whereas ARROW-AI emphasizes auditable human decisions before, during, and after AI use. This is a procedural alternative, not proof that all concealed use can be detected or prevented.

5.2 User Evaluation and Bounded Implementation

Student-researchers and advisers/instructors valued the principal domains but reported moderate documentation and workload burden; verification support was especially strong, while feasibility was lower than usefulness. The favorable perceptions converge with studies reporting perceived benefits of generative AI for feedback, ideation, and writing support (Chan & Hu, 2023; Rajabi et al., 2024). The burden and feasibility findings provide an important counterpoint: acceptance of AI-related guidance does not imply willingness or capacity to complete extensive documentation.

This tension also aligns with evidence that students may use generative AI without fully recognizing the associated integrity implications (Gruenhagen et al., 2024). Participants' favorable acceptability ratings, therefore, cannot be

interpreted as proof of accurate self-disclosure. Verification records may encourage source retrieval and evidence comparison in response to fabricated or unverifiable output (Walters & Wilder, 2023), but the framework still depends on honest reporting and active adviser review.

All initially high-risk requests were down-scoped; none were implemented as proposed. Completion was high but incomplete, and four cases required substantive revision. Risk classification thus functioned as a formative checkpoint and shared vocabulary. It is consistent with context-sensitive permission frameworks (O'Sullivan et al., 2025; Perkins et al., 2025), but the bounded pilot does not show that inappropriate use was eliminated or that the procedure would operate similarly in other disciplines or institutions.

The difference between favorable usefulness and lower feasibility is best interpreted through implementation theory. Acceptability and feasibility are distinct from effectiveness and require separate evidence (Proctor et al., 2011). Accordingly, the improved defense-checklist records indicate only documented preparedness, not improved oral defense performance. Proportional records are warranted: abbreviated for low-risk tasks, complete for moderate-risk tasks, and redesigned or prohibited for high-risk tasks.

5.3 Content Validity and Measurement Boundaries

All 15 components met the I-CVI and CVR criteria, and the final audit checklist—the lowest-rated component—was retained with revision. This pattern shows how quantitative thresholds and expert comments jointly informed refinement. Under an argument-based view of validation, these results support only the intended developmental decisions and the bounded use of the ratings; they do not establish construct validity, causal impact, or the instrument's permanent quality (American Educational Research Association et al., 2014; Kane, 2013).

The expert and user instruments were designed for formative review across conceptually distinct criteria rather than for producing unidimensional latent-scale scores. The domain ratings were consequently interpreted only as descriptive evidence of how participants evaluated specific features of the framework and toolkit. No claim of internal consistency, structural validity, or psychometric scale reliability is made.

This measurement boundary does not invalidate the content validity, qualitative evidence, or implementation evidence, but it limits the interpretation of the numerical ratings. Framework refinement was therefore based on convergence among I-CVI and CVR results, separate domain ratings, expert and user comments, qualitative findings, and pilot records. Future studies should develop and test separate domain-specific scales if psychometric score interpretation is intended, including evaluation of internal consistency, factorial structure, temporal stability, and measurement invariance.

5.4 Integrated Revisions and Claim Boundaries

Qualitative evidence identified verification and accountability as strengths; repetition as the main weakness; underreporting and adviser capacity as risks;

and examples, simpler fields, and orientation as priorities. These findings explain why the framework structure was retained while the toolkit was modified. They also show that governance procedures can create new workload and reporting burdens even when users endorse their purpose.

Version 2.0 added permitted/restricted examples, risk notes, a completed prompt log, a shorter verification ledger, expanded disclosure/defense prompts, and an adviser checklist. Because these changes may alter burden, comprehension, and score behavior, the revised toolkit requires another validation cycle.

ARROW-AI's bounded originality is the integration of task permission, research alignment, verification, adviser review, disclosure, and defense into one research-writing sequence. It supplements rather than replaces policy, ethics, supervision, or publication standards and requires multisite validation.

5.5 Practical Implications

5.5.1 Implications for student-researchers

Student-researchers should encounter ARROW-AI before using AI in substantial ways. They should classify the task, retain intellectual decisions, verify sources and claims, and disclose what was retained, revised, or rejected. Verification records should identify the AI output, the authoritative evidence, the comparison result, and the action taken.

5.5.2 Implications for Advisers, Instructors, and Panel Evaluators

Advisers, instructors, and panel evaluators should use ARROW-AI as a structured conversation rather than a mechanical checklist. Early risk review can prevent high-risk delegation; later records can focus evaluation on reasoning and evidence. Because usefulness exceeded feasibility and burden was moderate, implementation should include a brief orientation, completed examples, a risk-based review, and an agreed-upon checking schedule.

5.5.3 Implications for Higher Education Institutions

Institutions should begin with a bounded pilot aligned with research manuals, integrity policy, privacy, ethics, and disclosure rules. Documentation should be short-form for low risk, complete for moderate risk, and redesigned or prohibited for high risk. Adoption also requires adviser training, disciplinary examples, protected disclosure procedures, and periodic policy review.

5.5.4 Implications for Further Research and Validation

Further studies should test Version 2.0 with larger, multisite, and more diverse samples; revalidate instruments; examine burden and workflow; and evaluate writing or defense outcomes only using suitable effectiveness designs.

5.6 Limitations

The developmental/formative design does not establish effects on writing quality, misconduct, defense performance, publication readiness, or adoption. Expert judgment, self-report, audit, and observation may also be affected by social desirability, incomplete disclosure, and researcher presence.

The single-institution purposive/convenience sample comprised 12 experts, 84 student-researchers, 18 advisers/instructors, and 32 short-term pilot cases. These features limit transferability and do not support population estimates.

The expert and user instruments were multidimensional formative-evaluation forms rather than established psychometric scales. Their domain means should therefore be interpreted only as descriptive judgments within the bounded study context. The study did not examine the internal consistency of independently developed domain scales, factorial structure, test–retest reliability, measurement invariance, or responsiveness to change. The qualitative corpus contained 180 meaning units, and only the initial calibration subset was independently double-coded. Rapidly changing AI capabilities and institutional policies also require periodic revision of the framework's risk categories, examples, and disclosure prompts.

6. Conclusion

This study answered four research questions by documenting ARROW-AI's development, content validation, formative evaluation, and evidence-based revision. The framework linked Authorial Control, Research Alignment, Risk-Stratified AI Use, Output Verification, and Written Disclosure and Defense into a single research-writing workflow. Expert-domain ratings were generally favorable at the descriptive level; all 15 components met the prespecified content-validity criteria; and student-researcher and adviser/instructor domain ratings indicated favorable perceptions, alongside moderate documentation and workload burden. The bounded pilot produced high but incomplete documentation, down-scoping of six initially high-risk requests, and four cases requiring substantive revision. Integrated evidence also identified repetition, possible underreporting, and adviser workload as implementation concerns.

Because the instruments contained conceptually distinct formative domains, their ratings were interpreted separately as descriptive judgments rather than as scores from established unidimensional psychometric scales. ARROW-AI should therefore be regarded as a developmental process guide, not an effectiveness-tested intervention, detector, or guarantee of preventing misconduct. Institutions may undertake bounded pilots using proportional documentation, completed examples, adviser orientation, approved local policies, and transparent disclosure. Before wider adoption, Version 2.0 requires renewed content and implementation evaluation with larger, multisite, and discipline-diverse samples. Future research should develop and test domain-specific measurement scales only when psychometric score interpretation is intended.

Conflict of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

7. Acknowledgments

The authors acknowledge the institutional support of Nueva Ecija University of Science and Technology in the preparation and development of this scholarly

work. The University's commitment to research, innovation, and academic excellence provided an enabling environment for the completion of this study. The authors also acknowledge the limited use of Grammarly and ChatGPT during manuscript preparation. The tool was used only for language polishing, grammar checking, improving clarity, and refining academic phrasing. All substantive aspects of the study, including conceptualization of the framework, research design, data interpretation, and final scholarly judgment, were carried out by the authors. The manuscript remains the authors' original intellectual work.

8. References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. https://www.testingstandards.net/open-access-files.html

Ayre, C., & Scally, A. J. (2014). Critical values for Lawshe's content validity ratio: Revisiting the original methods of calculation. *Measurement and Evaluation in Counseling and Development, 47*(1), 79–86. https://doi.org/10.1177/0748175613513808

Bittle, K., & El-Gayar, O. (2025). Generative AI and academic integrity in higher education: A systematic review and research agenda. *Information, 16*(4), Article 296. https://doi.org/10.3390/info16040296

Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education, 20*(1), Article 43. https://doi.org/10.1186/s41239-023-00411-8

Chiu, T. K. F. (2024). Future research recommendations for transforming higher education with generative AI. *Computers and Education: Artificial Intelligence, 6*, Article 100197. https://doi.org/10.1016/j.caeai.2023.100197

Committee on Publication Ethics. (2023, February 13). *Authorship and AI tools*. https://publicationethics.org/cope-position-statements/ai-author

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International, 61*(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. *International Journal of Educational Technology in Higher Education, 20*(1), Article 22. https://doi.org/10.1186/s41239-023-00392-8

DeVellis, R. F., & Thorpe, C. T. (2021). Scale development: Theory and applications (5th ed.). SAGE Publications.

Fetters, M. D., & Tajima, C. (2022). Joint displays of integrated data collection in mixed methods research. *International Journal of Qualitative Methods, 21*, 16094069221104564. https://doi.org/10.1177/16094069221104564

Gruenhagen, J. H., Sinclair, P. M., Carroll, J.-A., Baker, P. R. A., Wilson, A., & Demant, D. (2024). The rapid rise of generative AI and its implications for academic integrity: Students' perceptions and use of chatbots for assistance with assessments. *Computers and Education: Artificial Intelligence, 7*, Article 100273. https://doi.org/10.1016/j.caeai.2024.100273

International Center for Academic Integrity. (2021). *The fundamental values of academic integrity* (3rd ed.). https://academicintegrity.org/images/pdfs/20019_ICAI-Fundamental-Values_R12.pdf

International Committee of Medical Journal Editors. (2026, January). *Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals*. https://www.icmje.org/recommendations/

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*(4), 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns, 4*(7), Article 100779. https://doi.org/10.1016/j.patter.2023.100779

Mayring, P. (2022). *Qualitative content analysis: A step-by-step guide*. SAGE Publications. https://doi.org/10.4135/9781036231798

McKenney, S., & Reeves, T. C. (2018). *Conducting educational design research* (2nd ed.). Routledge. https://doi.org/10.4324/9781315105642

Miao, F., & Cukurova, M. (2024). *AI competency framework for teachers*. UNESCO. https://doi.org/10.54675/ZJTE2084

Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO. https://doi.org/10.54675/EWZM9535

Miao, F., Shiohira, K., & Lao, N. (2024). *AI competency framework for students*. UNESCO. https://doi.org/10.54675/JKJB9835

O'Sullivan, J., Lowry, C., Woods, R., & Conlon, T. (2025). *Generative AI in higher education teaching and learning: National policy framework*. Higher Education Authority. https://doi.org/10.82110/PX37-MP48

Perkins, M., Roe, J., & Furze, L. (2025). Reimagining the artificial intelligence assessment scale: A refined framework for educational assessment. *Journal of University Teaching and Learning Practice, 22*(7). https://doi.org/10.53761/rrm4y757

Perkins, M., Roe, J., Postma, D., McGaughran, J., & Hickerson, D. (2024). Detection of GPT-4 generated text in higher education: Combining academic judgement and software to identify generative AI tool misuse. *Journal of Academic Ethics, 22*(1), 89–113. https://doi.org/10.1007/s10805-023-09492-6

Pinho, I., Costa, A. P., & Pinho, C. (2025). Generative AI governance model in educational research. *Frontiers in Education, 10*, Article 1594343. https://doi.org/10.3389/feduc.2025.1594343

Proctor, E., Silmere, H., Raghavan, R., Hovmand, P., Aarons, G., Bunger, A., Griffey, R., & Hensley, M. (2011). Outcomes for implementation research: Conceptual distinctions, measurement challenges, and research agenda. *Administration and Policy in Mental Health and Mental Health Services Research, 38*(2), 65–76. https://doi.org/10.1007/s10488-010-0319-7

Rajabi, P., Taghipour, P., Cukierman, D., & Doleck, T. (2024). Unleashing ChatGPT's impact in higher education: Student and faculty perspectives. *Computers in Human Behavior: Artificial Humans, 2*(2), Article 100090. https://doi.org/10.1016/j.chbah.2024.100090

Richey, R. C., & Klein, J. D. (2007). *Design and development research: Methods, strategies, and issues*. Routledge. https://doi.org/10.4324/9780203826034

Smith, S. M., Tate, M., Freeman, K., Walsh, A., Ballsun-Stanton, B., & Lane, M. (2026). A university framework for the responsible use of generative AI in research. *Journal of Higher Education Policy and Management, 48*(1), 17–36. https://doi.org/10.1080/1360080X.2025.2509187

Tertiary Education Quality and Standards Agency. (2025, September 24). *Enacting assessment reform in a time of artificial intelligence*. https://www.teqsa.gov.au/guides-resources/resources/corporate-publications/enacting-assessment-reform-time-artificial-intelligence

Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports, 13*(1), Article 14045. https://doi.org/10.1038/s41598-023-41032-5

Yusoff, M. S. B. (2019). ABC of content validation and content validity index calculation. *Education in Medicine Journal, 11*(2), 49–54. https://doi.org/10.21315/eimj2019.11.2.6