Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An AI Literacy Checkpoint Framework for Critical Evaluation of AI-Generated Content in Higher Education

Muhammad Subhan

msubhan@utas.edu.au

University of Tasmania

Ali Raza

University of Tasmania

Soonja Yeom

University of Tasmania

Soo-Hyung Kim

Chonnam National University

Hyung-Jeong Yang

Chonnam National University

Research Article

Keywords: AI literacy, Generative artificial intelligence, Higher education, Automated assessment, Hallucination detection, Academic integrity

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11206695/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

An AI Literacy Checkpoint Framework for Critical Evaluation of AI-Generated Content in Higher Education

Muhammad Subhan[1*], Ali Raza[1], Soonja Yeom[1*], Soo-Hyung Kim[2], Hyung-Jeong Yang[2]

[1]School of ICT, University of Tasmania (UTAS), 7001, Australia
[2]Department of Artificial Intelligence Convergence, Chonnam National University, Korea

[*]Corresponding author: Soonja Yeom, **soonja.yeom@utas.edu.au**
[*]Corresponding author: Muhammad Subhan, **msubhan@utas.edu.au**

ORCID iDs:

Muhammad Subhan: 0009-0009-7939-6004
Soonja Yeom: 0000-0002-5843-101X
Soo-Hyung Kim: 0000-0003-3575-5035
Hyung-Jeong Yang: 0000-0003-3024-5060

Abstract

The rapid adoption of generative artificial intelligence (GenAI) in higher education has created a significant assessment gap, as students increasingly submit AI-assisted work without demonstrating competencies required to verify facts, authenticate sources, identify hallucinations, assess bias, uphold academic integrity, and critically reflect on AI use. Although several AI literacy frameworks have been proposed, including those by Fan et al. (2025), AI4K12 (Touretzky et al., 2019), and international policy reports (UNESCO, 2019, 2021, 2022), no automated and empirically validated system has existed to assess AI-assisted academic submissions systematically. This study introduces the AI Literacy Checkpoint Framework (ALCF), a six-checkpoint automated sequential pipeline that evaluates Source Verification (CP1), Fact Verification (CP2), Hallucination Detection (CP3), Bias and Fairness Assessment (CP4), Academic Integrity (CP5), and Reflection and Critical Thinking (CP6). The framework was evaluated through three progressively larger sets of constructed submissions to examine feasibility, robustness, and scalability. Phase 1 ($N = 10$ submissions) established proof of concept; Phase 2 ($N = 20$ submissions) introduced more complex error scenarios and structured reporting; and Phase 3 ($N = 50$ submissions) provided the primary evaluation with comparative baseline systems, approval strategies, and inter-rater reliability analysis. Phase 3 achieved an end-to-end accuracy of 0.980, with precision of 0.964, recall of 1.000, F1-score of 0.982, and MCC of 0.960. Deterministic checkpoints (CP1–CP4) achieved strong performance (0.92–0.96 accuracy), whereas subjective checkpoints (CP5–CP6) remained limited at 0.72 accuracy, highlighting challenges in rubric-based assessment. Comparative evaluations showed that the complete ALCF outperformed citation-only, retrieval-only, and majority-vote alternatives. These findings establish ALCF as a flexible and extensible framework that can be adapted as AI capabilities, hallucination profiles, and educational assessment practices evolve.

Keywords: AI literacy; Generative artificial intelligence; Higher education; Automated assessment; Hallucination detection; Academic integrity

1. Introduction

1.1. The Challenge of AI-Generated Content in Higher Education

Large language models (LLMs) are now rapidly transforming the creation, evaluation, and sharing of academic knowledge within the context of higher education. The use of generative AI by students has become more common in various tasks, such as drafting essays, synthesizing literature, creating arguments, and generating references (Brown et al., 2020; Ouyang et al., 2022). The technologies provide great pedagogical potential, but at the same time, they have revealed an important weakness of the current assessment landscape: institutions are increasingly rating the products of the use of AI without having a scalable process to verify, evaluate, and responsibly use the outputs from AI use to ensure that the students are equipped with the epistemic skills to do so.

The challenge is not limited to issues with text generation only. Empirical evidence shows that modern LLMs are very often able to produce invented academic references, incorrect factual statements, and rhetorically guided texts, while remaining coherent and fluent. The percentages of citation hallucination were found to be 15% to 47% in scientific writing tasks by Ansari (2026), meaning that unverified submissions assisted by AI can be plagiarized with a high degree of risk. Likewise, Huang et al. (2025) pointed out hallucination as a long-standing architectural challenge to generative models: generating factually incorrect information with confidence and surface plausibility. In addition to their factual reliability, scholars have shown that web-scale training corpora contain historical, demographic, and ideological biases that can be reproduced or amplified in generated content (Bender et al., 2021; Dodge et al., 2021). The use of AI writing tools can be beneficial, but it should be done carefully, responsibly, and with a critical mindset, as students must ensure that they do not plagiarize or fabricate information, and respect the integrity, transparency, and critical scholarship of academic work.

While many of these issues are being recognized, institutional approaches are still limited in scope and are reactive. Current technological interventions target only elements of the problem: text similarity detection tools assess the degree of overlap in content, AI-text detection tools try to determine the authorship of a piece of text, and guidelines on responsible AI usage offer conceptual clarity and direction. These approaches, however, do not address the full range of AI literacy skills needed to undertake academic work responsibly, nor offer empirically supported methods for gauging the reliability, integrity, and critical quality of AI-aided work before submission for evaluation. This has led to significant disconnects between the forward-looking goals outlined in AI literacy initiatives and the tools available to bring those goals to life in real-world educational environments.

This study addresses the research gap by designing, implementing, and progressively testing the AI Literacy Checkpoint Framework (ALCF), a six-step automated assessment architecture that assesses AI-supported academic work across source verification, factual accuracy, hallucination detection, bias and fairness assessment, academic integrity, and reflective critical engagement. This research uses a staged methodology in which the framework is examined through a three-phase design that progressively adds participants ($N = 10$,20, and 50) and methodological complexity. Progressive methodological improvements, larger evidence sets, and more stringent validity procedures are introduced in each phase based on limitations identified in earlier versions.

The novelty of this research lies not only in developing an AI literacy assessment framework but also in addressing limitations identified across existing AI literacy, educational-assessment, and responsible-AI approaches. It defines a longitudinal validation process through which automated educational assessment systems can progress from proof-of-concept prototypes to methodologically sound and scalable tools. This study outlines the performance trajectory, methodological maturation, comparative baseline assessments, and limitations across the stages of the research, offering a practical model for the empirical development of AI literacy technologies in higher education.

1.2. Research Questions

Six main research questions were formulated for this research:

> **RQ1:** Can a six-checkpoint sequential automated pipeline reliably classify AI-assisted academic submissions as meeting or failing AI literacy standards, and how does this reliability evolve as the evaluation dataset and evidence corpus scale from $N = 10$ to $N = 50$?
>
> **RQ2:** How does per-checkpoint classification performance vary across the six checkpoint stages, and is this performance gradient stable or volatile as sample size increases across the three phases?
>
> **RQ3:** Does the choice of retrieval strategy for fact verification (hybrid BM25+TF-IDF versus single-method configurations) affect end-to-end classification performance at Phase 3 scale?
>
> **RQ4:** How does the complete six-checkpoint design compare against simplified internal baseline systems — including citation-only, retrieval-only, technical-subset, majority-vote, and weighted-fusion alternatives — in terms of end-to-end classification accuracy?
>
> **RQ5:** What do the formally specified approval-strategy equations (conjunction, majority, weighted, confidence) reveal about the precision–flexibility trade-off inherent in the framework's final decision rule?
>
> **RQ6:** What is the illustrative inter-rater reliability of the subjective checkpoints (CP5 Academic Integrity, CP6 Reflection), and what does the associated infrastructure imply for the design of a future expert-validated study?

2. Literature Review

2.1. AI Literacy Competency Frameworks

The conceptual foundations of the AI Literacy Checkpoint Framework (ALCF) are drawn from recent research in AI literacy, instructional technology, and explainable AI (XAI). The principal foundation is the AI literacy model developed by Fan et al. (2025), which combined seventeen dimensions of AI competence from education, human–computer interaction, and public-policy research. Their definition presents AI literacy as a multifaceted construct that includes understanding how AI systems work, identifying their capabilities and limitations, critically assessing outputs, using AI responsibly, and maintaining ethical awareness. ALCF translates these abstract skills into six measurable checkpoints. Specifically, CP1 addresses awareness of the limitations of AI-generated citations; CP2 and CP3 address the critical evaluation of factual claims and the identification of hallucinations; CP4 addresses bias and fairness awareness; and CP5 and CP6 address transparent, responsible, and reflective engagement with AI-assisted writing.

The wider scope of education offered by ALCF is also informed by the work of Touretzky et al. (2019), who outlined five key principles of AI literacy and education about artificial intelligence systems: perception, representation and reasoning, learning, natural interaction, and societal impact. Although created for pre-university use, these principles offer a valuable conceptual framework for understanding AI literacy as a more complex concept than technology alone, involving critical analysis of AI systems and their impact on society. This echoes the ongoing recommendations of UNESCO regarding AI competency in education, as expressed in its Beijing Consensus on Artificial Intelligence and Education (UNESCO, 2019), Recommendation on the Ethics of Artificial Intelligence (UNESCO, 2021), and AI curriculum guidance (UNESCO, 2022). Together, these efforts highlight transparency, accountability, moral considerations, and responsible use of AI, which are fundamental principles guiding CP5's approach to AI disclosure and attribution requirements.

The framework also draws on research exploring the pedagogical risks of using AI in education. Holmes et al. (2019) identified key issues with educational AI systems, including opacity, bias, and learner dependency. ALCF has a checklist for source verification, hallucination detection, bias assessment, and reflective review, all of which

are central to addressing these concerns. The framework also acknowledges that AI-generated outputs will not necessarily be trustworthy, prompting students to undertake systematic verification processes to develop deeper critical thinking and minimize reliance on AI.

Another theoretical influence relates to the emergence of Explainable Artificial Intelligence in Education (XAIED). López-Pernas et al. (2025) argued that explainable educational technologies can help develop metacognition by allowing learners to comprehend the process of automated decision-making. This principle is integrated into ALCF's architecture by providing clear, understandable checkpoint-level evaluations, with evidence trails and clearly defined criteria for each evaluation outcome. Therefore, the framework is not only an assessment tool, but also a learning tool that promotes reflection on the products of AI. Lastly, the framework addresses a gap in the overall Artificial Intelligence in Education (AIED) literature. In their extensive review of 20 years of research on AIED, Chen et al. (2022) found substantial scholarly work on intelligent tutoring systems, adaptive learning environments, and automated feedback mechanisms, but comparatively little work on systematically assessing AI literacy itself. Few existing studies are based on staged, multi-phase validation methods that allow for demonstrating the stability of performance over time and the maturity of the method. Supporting this finding, Shukor and Osman (2025) proposed that fostering AI literacy should involve not just conceptual training but active participation in evaluating and critically reflecting on AI-generated outputs. ALCF operationalizes this pedagogical principle by offering formative, checkpoint-specific feedback on each piece of work, making assessment a structured opportunity to develop AI literacy.

These theoretical frameworks place ALCF at the crossroads of AI literacy research, educational assessment, explainable AI, and responsible AI governance. The framework represents an initial effort to translate high-level competency models and ethical considerations into a practical, quantifiable, and empirically evaluated assessment framework. It responds to challenges reported across the literature, including difficulties in verifying AI-generated information, detecting hallucinations, assessing bias, maintaining academic integrity, and supporting students' critical reflection on AI-assisted work.

2.2. Large Language Model Hallucination and Citation Fabrication

Recent studies have shown that hallucination is not simply an implementation problem of large language models (LLMs), but a natural by-product of their probabilistic construction. Unlike systems that verify claims against an authoritative representation of external reality (Bommasani et al., 2021; Brown et al., 2020; Petroni et al., 2019), LLM outputs are generated by predicting statistically likely token sequences based on patterns observed during training. As a result, linguistic fluency and factual correctness are not necessarily tied together: models can produce content that is coherent, convincing, and academically sound, but factually incorrect. This is a major difficulty in educational settings, as a student may lack the skills necessary to distinguish verifiable information from credible sources from statements that merely sound plausible.

This phenomenon has been reported extensively in the hallucination literature. In a comprehensive review of question answering, summarization, dialogue systems, and machine translation, Huang et al. (2025) identified two types of hallucination: faithfulness hallucination, in which the generated output does not match the given source material, and factuality hallucination, in which the output contradicts existing world knowledge and modality-specific constraints. Following this, Zhang et al. (2025) proposed a more detailed categorization of hallucinations — input-conflicting, context-conflicting, and fact-conflicting — each requiring specific detection and mitigation approaches. Hallucinations are especially challenging in academic contexts because they can appear linguistically coherent and convincing while containing unsupported or incorrect information.

CP3's design is further guided by evidence on the effectiveness of Natural Language Inference (NLI)-based verification approaches. Studies by Falke et al. (2019) and Maynez et al. (2020) demonstrated that entailment

relationships between generated statements and trusted evidence sources offer a strong tool for detecting unsupported or contradictory claims. ALCF implements this by evaluating the strength of evidence supporting each generated claim and converting evidence deficits into an interpretable hallucination risk score.

Citation hallucination is a related form of factuality failure that has become one of the greatest dangers of AI writing assistants in academia. Citation hallucinations are distinct from ordinary factual errors, producing a false impression of academic authority through citations that mimic typical academic style but correspond to nonexistent or incorrect sources. Ansari (2026) showed that current LLM systems can generate full bibliographic metadata — author names, journal names, publication details, page information, and digital object identifiers (DOIs) — often in ways that are hard to detect at a glance. Sakai et al. (2026) reported fabrication rates ranging from 15% to 47% across models and disciplinary areas.

Citation hallucination is especially dangerous from an educational perspective, as it undermines both academic honesty and the transparency of research, and may enable academic dishonesty even when students do not intend to deceive. ALCF addresses this challenge with Source Verification (CP1), an automated bibliographic validity checking process that uses real-time cross-referencing with authoritative scholarly databases, replacing manual and error-prone verification with a formalized assessment tool.

2.3. Retrieval-Augmented Generation for Fact Verification

The growing incidence of hallucination in LLM outputs has increased interest in retrieval-based methods that connect generated content with externally verifiable evidence. One prominent approach is Retrieval-Augmented Generation (RAG), which extends the parametric memory of language models with information retrieved from external knowledge repositories, thereby reducing unsupported claims (Singh et al., 2025). RAG enables AI systems to move beyond their training data, improving transparency and factual grounding.

The RAG paradigm has been extended in later studies. In a thorough analysis of retrieval-augmented architectures, Zhao et al. (2026) distinguished between naive, advanced, and modular RAG systems, finding that hybrid sparse retrieval approaches remain among the best and least computationally expensive methods for domain-specific knowledge verification. Importantly, much of the RAG literature is dedicated to improving text generation. The present research, by contrast, reuses retrieval augmentation for assessment: retrieval is a process of discovering evidence to verify, rather than generating new content. Given a student submission, ALCF retrieves the top-n most relevant passages from a curated collection of scholarly works and assesses how well the claims in the submission correspond with independently sourced evidence.

The retrieval architecture used in ALCF combines two complementary sparse retrieval methods: BM25 and Term Frequency–Inverse Document Frequency (TF-IDF). BM25 relies on the probabilistic retrieval tradition (Robertson & Zaragoza, 2009) to estimate document relevance, incorporating term frequency, inverse document frequency, and a saturation function that prevents excessively frequent terms from overwhelming the ranking score. TF-IDF (Salton & McGill, 1983) constructs a weighted vector-space model of documents and queries and uses cosine similarity for ranking. These methods complement each other, offering different perspectives on retrieval and increasing the likelihood of retrieving relevant evidence across academic domains.

To combine the strengths of both retrieval methods, ALCF uses Reciprocal Rank Fusion (RRF) (Cormack et al., 2009), a rank aggregation method that fuses multiple ranked retrieval results without supervised learning or manually set weights. RRF sums the reciprocal ranks that each document receives across retrieval systems, giving higher rank to documents that are consistently close to the top of multiple ranked lists. This hybrid retrieval architecture is supported by prior empirical studies: Chen et al. (2017) showed that TF-IDF retrieval combined with downstream evidence evaluation can be competitive with modern systems in open-domain question

answering, while Dai and Callan (2020) demonstrated that context-aware term weighting significantly enhances retrieval effectiveness in domain-specific corpora.

Retrieval is thus the crucial link between the assessment of AI literacy and evidence-based reasoning: assessment in ALCF is grounded not in linguistic or stylistic evaluation, but in independently verifiable evidence.

2.4. Academic Integrity and Disclosure-Based Assessment

The advent of generative artificial intelligence has shaken the foundations of traditional academic honesty. Traditional institutional efforts on integrity have relied on plagiarism detection via text-similarity analysis, premised on the idea that plagiarism involves unauthorized reuse of existing human-created text. In the age of large language models, this model is increasingly obsolete, as LLMs can produce original language that may nonetheless contain inaccuracies, fabricated citations, or lack clear evidence of student authorship and critical interpretation (Azevedo et al., 2025; Perkins, 2023). The integrity issue at the heart of AI-driven learning environments is therefore broader than ensuring originality; it concerns transparency, accountability, authorship, and responsible use.

Recent research has begun to expose the limitations of detection techniques. Yang et al. (2025) developed a statistical model (DetectGPT) for identifying AI-generated text, but follow-up testing revealed significant drawbacks, including high false-positive rates for non-native English speakers and domain-specific academic writing. This has led scholars to argue that the central question should not be whether AI was used, but how it was used and whether such use was disclosed and critically evaluated.

Perkins (2023) articulated a disclosure-based approach to academic integrity, suggesting that responsible AI use should be evaluated through transparent documentation of AI use, verification practices, and substantive student contribution. ALCF operationalizes this via Academic Integrity Checkpoint 5 (CP5), which assesses explicit disclosure of AI use, evidence of independent verification, documentation of the revision process, and demonstration of meaningful student contribution.

Self-regulated learning theory provides additional theoretical grounding for CP5. Winne and Hadwin (1998) identified metacognitive monitoring — the ability of learners to assess, control, and explain their own thinking — as central to effective learning. CP5 reflects this by requiring students to demonstrate understanding of the creation, revision, and review of AI-generated outputs prior to submission. Kizilcec (2016) showed that transparency in assessment processes enhances student trust, fairness, and acceptance of automated systems, motivating CP5's explicit, rubric-based, interpretable evaluation model. Cognitive Load Theory (Sweller, 1988) further informs the checkpoint's design: rather than requiring a single holistic judgement of integrity, CP5 decomposes the construct into four independently evaluable dimensions (disclosure, verification, authorship, and reflection), improving consistency and reducing assessor and learner burden. Finally, Floridi et al. (2021) identified transparency, accountability, and explainability as necessities for socially responsible AI systems — principles that CP5 operationalizes as measurable assessment criteria.

2.5. Bias Detection and Rhetorical Analysis

One of the most prevalent issues with LLMs is bias. Trained on enormous volumes of internet text, LLMs inevitably learn patterns, assumptions, and representational imbalances present in their training data, and can perpetuate or amplify social, cultural, ideological, and demographic biases in ways that appear natural and authoritative. Bender et al. (2021) drew attention to the connection between training-data composition and biased model behavior, while Dodge et al. (2021) showed that large-scale web corpora contain significant bias, exclusion, and problematic content that can be propagated by language models trained on them.

Linguistic bias has long been studied in NLP. Tomar et al. (2025) introduced lexicon-based techniques for identifying hateful and biased language, while Cignarella et al. (2025) proposed a more comprehensive approach

based on subjective language, framing effects, evaluative expressions, and perspective markers — moving beyond explicit prejudicial language toward the subtler rhetoric of unbalanced argumentation, selective framing, absolutist statements, and missing alternative viewpoints. Gallegos et al. (2024), in a comprehensive review of LLM bias assessment and mitigation, highlighted the value of lightweight lexicon-based methods for their transparency, interpretability, and computational efficiency — properties that are often more important than predictive competence in educational assessment contexts. Ferrara (2024) further showed how LLM outputs can bias user perception in education, information, and decision-making contexts through subtle rhetorical mechanisms, underscoring the need for scalable bias-detection solutions.

ALCF's Bias and Fairness Assessment checkpoint (CP4) builds directly on this literature. Rather than relying on an opaque machine-learning classifier, CP4 uses a transparent, rubric-based multi-lexicon architecture evaluating five dimensions of rhetorical balance: absolutist language, viewpoint diversity, demographic bias markers, subjective/evaluative expressions, and negation-balance indicators. This multi-lexicon design is theoretically and practically motivated: bias is a multidimensional phenomenon that cannot be captured by a single indicator, and lexicon-based methods remain transparent, inexpensive, and domain-independent across the 21 disciplinary domains covered in the Phase 3 evidence corpus.

2.6. Reflective Critical Thinking in AI-Supported Learning

The final checkpoints of ALCF are grounded in educational theories holding that reflection, evaluation, and metacognitive regulation represent the highest stages of intellectual development. CP6 builds on the earlier checkpoints by emphasizing a fundamental educational goal: ensuring that students do not passively accept AI-generated content.

One of the most influential foundations of CP6 is the revised taxonomy of educational objectives proposed by Anderson and Krathwohl (2001), which places evaluation and creation — assessing validity, providing justification, and producing informed responses — at the top of the cognitive hierarchy. CP6 operationalizes these higher-order skills via an epistemic reflection rubric requiring students to explain why information is trustworthy, identify supporting evidence, and document substantive changes to AI-generated drafts.

Such reflective activities are supported by research on retrieval-based learning: Franzoi et al. (2025) showed that retrieval practice produces deeper conceptual learning than passive review. Kaddoura (2013) found that structured, guided reflective exercises improve students' ability to analyze information, evaluate alternatives, and make reasoned judgments. The broader instructional design of ALCF draws on Vygotsky and Cole's (1978) Zone of Proximal Development, scaffolding students from concrete verification tasks (early checkpoints) toward evaluative judgement, ethical reasoning, and metacognition (later checkpoints).

Concerns about cognitive offloading motivate CP6's design: Gerlich (2025) found a strong negative correlation between frequent AI tool use and measured critical-thinking performance, mediated by cognitive offloading. However, Gonsalves (2026) showed that structured, evidence-based reflection significantly improved critical-thinking outcomes for students using generative AI, compared to students using AI without reflection — indicating that thoughtfully designed reflective interventions can mitigate offloading risks. CP6 assesses the quality of students' reflective thinking, evidence-based explanations, and intellectual engagement, treating reflection as a manifestation of cognitive process rather than a mandatory administrative step.

2.7. Statistical Foundations for Small-Sample Educational Technology Evaluation

Because ALCF was validated using three increasingly sizeable pilots rather than a single large authentic dataset, the choice of statistical evaluation methods is of critical methodological importance, particularly given performance values near the extremes of the metric range and non-ideal class distributions.

Traditional normal-approximation (Wald) confidence intervals are known to have significant coverage deficiencies for small sample sizes and proportions near 0 or 1. Because end-to-end accuracy ranged from 0.900 to 0.980 across the three validation phases, this study instead uses the Wilson score interval (Wilson, 1927), which provides more stable and statistically sound estimates of uncertainty for binomial proportions under small-sample conditions by adapting to asymmetry near the boundaries of the probability distribution.

Matthews' correlation coefficient (MCC; Matthews, 1975) is used as a key summary statistic, incorporating all four outcomes of the confusion matrix into a single balanced measure. This is especially important given the approval/rejection class distributions across phases (6:4 for Phase 1, 11:9 for Phase 2, 27:23 for Phase 3), where accuracy alone can produce overly optimistic interpretations.

Phase 3 also introduces infrastructure for evaluating the reliability of subjective checkpoint judgements (CP5, CP6), grounded in Cohen's (1960) Kappa (two raters) and Fleiss' (1971) Kappa (three or more raters), interpreted according to the Landis and Koch (1977) benchmark scale. This study explicitly distinguishes between reliability *infrastructure* (as implemented and verified) and reliability *outcomes* (as would be obtained from genuine expert ratings); the illustrative kappa analyses in Phase 3 demonstrate methodological readiness rather than confirmed assessment reliability, with comprehensive instructor-based validation identified as a key objective for future work.

3. Methodology

3.1. Research Design: A Three-Phase Staged Validation Program

This study uses a staged mixed-methods design, combining systems engineering with empirical evaluation and statistical analysis, to design and validate the AI Literacy Checkpoint Framework (ALCF). The study integrates the design and implementation of an automated assessment pipeline, benchmark-based assessment against an increasing set of gold-standard labels, and performance, comparative, and reliability analyses. The research adopts a post-positivist perspective, treating AI literacy as an empirically measurable construct while recognizing that certain dimensions (academic integrity, reflective thinking) are inherently subjective and involve judgement.

A distinctive feature of the study is its three-phase validation approach, through which the evaluation sample and methodological tools are incrementally expanded. Phase 1 ($N = 10$) demonstrated the feasibility of the six-checkpoint framework and implemented the retrieval-augmented verification pipeline and benchmark workflow. Phase 2 ($N = 20$) expanded the evidence base, introduced borderline and multi-error cases, and improved research-reporting functionality. Phase 3 ($N = 50$) further expanded the dataset and evidence corpus and introduced six comparative baseline systems, four alternative approval strategies, a formal decision-theoretic framework, and an inter-rater reliability module using Cohen's and Fleiss' kappa statistics.

The evolution of the design follows recommendations for fact-verification and educational AI research (Thorne et al., 2018), which advocate first testing systems on purpose-built benchmark datasets before moving to larger-scale, higher-quality datasets. All three phases provide evidence of final performance as well as methodological maturation, performance stability, and framework development over time.

3.2. Pipeline Architecture and Design Principles

ALCF uses a six-checkpoint assessment process that sequentially evaluates AI-supported academic submissions. Under the default conjunction-based approval strategy, a submission is approved only if all six checkpoints return a PASS result. This design follows a conservative assessment philosophy: approval requires satisfactory performance in source verification, factual accuracy, hallucination detection, bias assessment, academic integrity, and reflective critical engagement.

The framework also supports alternative approval models — majority-vote, weighted-fusion, and confidence-based decision rules (Section 3.10) — and can terminate at the first checkpoint failure to reduce computational overhead. The default implementation, however, runs all six checkpoints regardless of intermediate outcomes, following formative-assessment principles by providing diagnostic feedback across every assessment dimension rather than only those directly determining the final decision. The complete methodological structure — data flow, checkpoint sequencing, external evidence and verification sources, decision mechanisms, and research-output generation — is shown in Fig. 1.

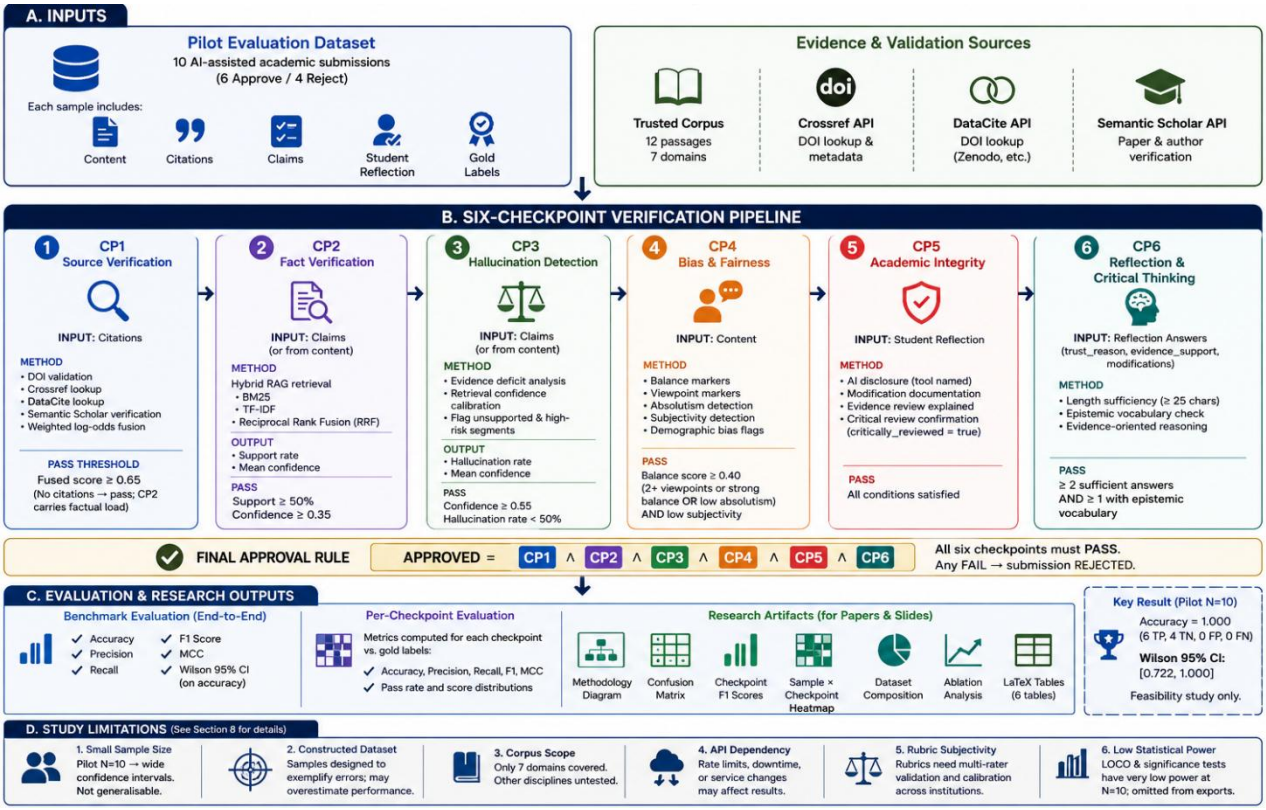


Figure 1: Methodological workflow of the AI Literacy Checkpoint Framework (ALCF)

3.3. Checkpoint 1 — Source Verification

The Source Verification checkpoint (CP1) evaluates the bibliographic authenticity of citations in a student submission. For each cited source — covering title, authors, publication venue, DOI, and publication year — the framework performs sequential verification against Crossref, DataCite, and Semantic Scholar. Crossref is the primary authority for DOI resolution and metadata validation, DataCite provides supplementary coverage for datasets and grey literature, and Semantic Scholar provides paper- and author-level confirmation when DOI-based verification is incomplete.

For each citation, six binary verification signals are generated: DOI validity, title correspondence, author plausibility, venue legitimacy, paper existence, and author confirmation. These signals are aggregated using weighted log-odds pooling to compute a citation-authenticity score:

$$L = \frac{\sum_{i=1}^{n} w_i r_i \log\left(\frac{s_i}{1-s_i+\varepsilon}\right)}{\sum_{i=1}^{n} w_i r_i} \tag{1}$$

where s_i denotes the verification signal, w_i is the signal weight, r_i is the reliability coefficient, and $\varepsilon = 10^{-6}$ is a numerical stability constant preventing division by zero. The aggregated score is transformed into a probabilistic confidence value using the sigmoid function:

$$\text{FusedScore} = \sigma(L) = \frac{1}{1+e^{-L}} \quad (2)$$

Signal weights are assigned based on evidential strength: DOI validity ($w = 1.2$) and title correspondence ($w = 1.0$) receive the highest influence, followed by author plausibility ($w = 0.9$), paper existence ($w = 0.85$), venue legitimacy ($w = 0.8$), and Semantic Scholar author confirmation ($w = 0.75$). The overall CP1 score for a submission is the mean fused score across all citations:

$$\text{CP1} = \frac{1}{m}\sum_{j=1}^{m} \text{FusedScore}_j \quad (3)$$

where m is the number of citations in the submission. A submission satisfies CP1 if:

$$\text{CP1} \geq 0.65 \quad (4)$$

Submissions containing no citations are assigned a default PASS outcome, as citation absence is treated primarily as an academic integrity concern managed in CP5. However, all factual claims remain subject to verification under CP2.

3.4. Checkpoint 2 — Fact Verification

The Fact Verification checkpoint (CP2) evaluates whether factual claims in a submission are supported by trusted scholarly evidence from the curated evidence corpus described in Section 4.1. Claims are extracted from the submission's explicit claim set, or, in its absence, from declarative sentences exceeding a predefined length threshold.

For each claim, the hybrid evidence retrieval module (Section 3.9) retrieves and ranks supporting passages using BM25, TF-IDF, and Reciprocal Rank Fusion (RRF). Let $S_{\max}$ denote the highest fused similarity score among the retrieved passages. A claim is classified as supported if:

$$S_{\max} \geq 0.35 \quad (5)$$

The CP2 score is computed by jointly considering the claim support rate (proportion of claims classified as supported) and the mean calibrated confidence across all claims. A submission passes CP2 only if both conditions are satisfied:

$$\text{Support rate} \geq 0.50 \quad \text{AND} \quad \text{Mean confidence} \geq 0.35 \quad (6)$$

3.5. Checkpoint 3 — Hallucination Detection

The Hallucination Detection checkpoint (CP3) assesses whether content segments exhibit elevated risk of fabricated or unsupported information. Unlike CP2, which evaluates evidential support for individual claims, CP3 focuses on segment-level hallucination risk based on calibrated uncertainty:

$$H_{\text{segment}} = 1 - \sigma_{\text{calibrated}}(\text{evidence confidence}) \quad (7)$$

A segment is flagged as a hallucination if it is unsupported by the evidence corpus, exceeds a minimum length threshold, and has hallucination risk exceeding 0.55. Flagged segments incur a penalty in checkpoint-level aggregation, reflecting the higher cost of false negatives relative to false positives in educational integrity settings. A submission passes CP3 only if both conditions are satisfied:

$$\text{Mean confidence} \geq 0.55 \quad \text{AND} \quad \text{Flagged-segment ratio} < 0.50 \quad (8)$$

3.6. Checkpoint 4 — Bias and Fairness Assessment

The Bias and Fairness Assessment checkpoint (CP4) evaluates rhetorical balance and neutrality using a multi-lexicon analysis framework. Five regular-expression lexicons identify linguistic indicators of bias, balance, and subjectivity: absolutism markers (e.g., *always*, *never*, *obviously*), negation-balance markers (e.g., *however*, *although critics argue*), viewpoint markers (e.g., *proponents*, *opponents*, *both sides*), demographic bias indicators, and subjectivity markers (e.g., *I believe*, *in my opinion*).

A composite fairness score combines the balance score, normalized viewpoint frequency, the complement of the absolutism ratio, and the complement of the subjectivity ratio; higher scores indicate greater rhetorical balance. Submissions with short, neutral text below a minimum word count, and exhibiting neither absolutism nor demographic bias indicators, are automatically assigned a PASS outcome. A submission passes CP4 if both conditions are satisfied:

$$\text{Composite score} \geq 0.40 \quad \text{AND} \quad \text{Subjectivity ratio} < 0.12 \tag{9}$$

3.7. Checkpoint 5 — Academic Integrity

The Academic Integrity checkpoint (CP5) evaluates student-provided reflection metadata using a four-dimension rubric grounded in disclosure-based AI integrity principles and self-regulated learning theory. The rubric assesses whether the student critically reviewed the AI-generated output; documented substantive modifications; explicitly disclosed the AI tool used; and reviewed the factual basis of AI-generated content. CP5 passes only when all the following are satisfied:

$$\text{critically_reviewed} = \text{True} \tag{10}$$

$$\text{ai_disclosed} = \text{True} \tag{11}$$

$$(\text{modifications_documented} \vee \text{evidence_reviewed}) = \text{True} \tag{12}$$

3.8. Checkpoint 6 — Reflection and Critical Thinking

The Reflection and Critical Thinking checkpoint (CP6) evaluates the epistemic quality of three structured reflection responses: *trust_reason*, *evidence_support*, and *modifications*. Each response is assessed for length sufficiency (substantive reflection) and the presence of epistemic vocabulary (e.g., *source*, *study*, *data*, *citation*, *DOI*, *research*, *evidence*, *peer-reviewed*). CP6 passes when at least two of the three reflection responses are substantive and at least one contains epistemic vocabulary:

$$\text{Substantive responses} \geq 2 \tag{13}$$

$$\text{Epistemic-vocabulary presence} \geq 1 \tag{14}$$

Both conditions must be satisfied for the checkpoint to pass.

3.9. Hybrid Retrieval Architecture

The evidence retrieval component underpinning CP2 and CP3 employs a non-parametric hybrid retrieval architecture combining BM25 and TF-IDF ranking via Reciprocal Rank Fusion (RRF). BM25 tokenizes queries and corpus passages into lowercase alphanumeric terms and computes relevance scores using the Okapi BM25 formulation with standard parameters ($k_1 = 1.5, b = 0.75$). TF-IDF represents queries and passages using unigram and bigram features with English stop-word removal and L2 normalization, ranking by cosine similarity.

Given a claim query, BM25 and TF-IDF independently rank all corpus passages; the ranked lists are merged via:

$$\mathrm{RRF}(d) = \sum_{m=1}^{M} \frac{1}{k + r_m(d)} \tag{15}$$

where $r_m(d)$ denotes the rank of document d under retrieval method m, M is the number of retrieval methods, and k is a smoothing constant. The highest-scoring passages under the resulting RRF ranking are returned as supporting evidence. Equal weighting is assigned to BM25 and TF-IDF during fusion, reflecting established findings that symmetric rank fusion provides a robust baseline in the absence of domain-specific calibration data.

3.10. Approval Strategies: A Formal Decision Model

A key methodological contribution introduced in Phase 3 is the formalization of the framework's final approval mechanism as a set of decision-theoretic strategies, extending the implicit rule-based conjunction approach used in Phases 1 and 2. Let $p_i \in \{0,1\}$ denote the pass status of checkpoint i, δ_i its continuous score, and w_i its assigned weight. Four approval strategies are implemented:

$$A_{\text{conj}} = \prod_{i=1}^{n} p_i \tag{16}$$

A submission is approved only if all checkpoints pass.

$$A_{\text{maj}} = \begin{cases} 1, & \sum_{i=1}^{n} p_i \geq k, \\ 0, & \text{otherwise}, \end{cases} \tag{17}$$

where $k = 4$ for the six-checkpoint framework.

$$A_w = \begin{cases} 1, & \frac{\sum_{i=1}^{n} w_i \delta_i}{\sum_{i=1}^{n} w_i} \geq \theta, \\ 0, & \text{otherwise}, \end{cases} \tag{18}$$

where $\theta = 0.72$.

$$A_c = \begin{cases} 1, & \bar{s} \geq \tau_c \ \wedge \ \sum_{i=1}^{n} p_i \geq k_{\min}, \\ 0, & \text{otherwise}, \end{cases} \tag{19}$$

where $\bar{s}$ denotes the mean checkpoint confidence score, τ_c is a configurable confidence threshold, and $k_{\min}$ specifies the minimum number of checkpoints that must pass.

The conjunction strategy is used as the primary approval mechanism throughout this study, as it maximizes precision by requiring satisfactory performance across all dimensions of AI literacy. The majority, weighted, and confidence strategies are fully implemented to support comparative analysis; however, consistent with the limitations discussed in Section 6, these alternative strategies have not yet been empirically validated using authentic student submissions and are reported as theoretical specifications of expected decision trade-offs.

3.11. Internal Baseline Systems

To assess the contribution of the complete six-checkpoint architecture, Phase 3 introduces six internal baseline systems evaluated on the same fifty-submission dataset and fifty-passage evidence corpus, differing only in checkpoint selection and approval logic: ALCF-Full (all six checkpoints, conjunction approval), Citation-Only (CP1), Retrieval-Only (CP2–CP3), Technical-3CP (CP1–CP3), Majority-Vote-6CP (approval when ≥ 4 of 6 checkpoints pass), and Weighted-Fusion-6CP (weighted aggregation of checkpoint scores). These baselines represent architectural simplifications of the overall framework and enable direct evaluation of whether the full six-checkpoint design provides measurable benefits over less comprehensive alternatives.

3.12. Ablation Study Design

An ablation study isolates the contribution of the retrieval component within CP2 and CP3, while keeping the six-checkpoint architecture and conjunction approval strategy unchanged. Three retrieval configurations were evaluated on the identical Phase 3 dataset and evidence corpus: ALCF-Full (BM25 + TF-IDF with RRF), RAG-TF-IDF-Only (TF-IDF retrieval without BM25), and RAG-BM25-Only (BM25 retrieval without TF-IDF).

3.13. Inter-Rater Reliability Infrastructure

Because CP5 and CP6 incorporate rubric-based judgement, Phase 3 introduces an inter-rater reliability module to support future validation against independent human assessors. The framework computes Cohen's Kappa (pairwise agreement) and Fleiss' Kappa (agreement among three or more raters), with interpretation based on established benchmark categories. To verify implementation correctness, the module was evaluated using an illustrative set of five constructed samples independently rated by three simulated assessors. The resulting reliability coefficients confirm the computational validity of the implementation; they are not intended as empirical validation, and collection of ratings from independent academic assessors remains a key objective for future work.

3.14. Evaluation Metrics

Framework performance is evaluated using accuracy, precision, recall, F1-score, specificity, and Matthews Correlation Coefficient (MCC) against binary gold-standard labels. MCC is reported as a primary summary measure because it incorporates all elements of the confusion matrix and remains informative under class imbalance. To quantify statistical uncertainty, 95% Wilson confidence intervals are reported for end-to-end accuracy across all evaluation phases. In addition to end-to-end evaluation, each checkpoint is assessed independently by comparing its PASS/FAIL outcome against checkpoint-specific gold labels using the same classification metrics.

4. Results

4.1. Framework Definition and Evidence Corpus Composition

Table 1 summarizes the research objective, verification method, data source, and pass criterion associated with each checkpoint (CP1–CP6). Table 2 summarizes the composition of the trusted evidence corpus used by CP2 and CP3, comprising fifty manually curated passages distributed across twenty-one academic and professional domains. Fig. 2 visualizes the domain distribution of the Phase 3 evidence corpus, demonstrating intentional coverage across STEM, social sciences, humanities, business, education, law, and digital technologies.

Table 1: Six-checkpoint framework definition, verification methods, data sources, and pass criteria.

CP	Stage	Research Question	Method	Data Source	Pass Criterion
1	Source Verification	Are citations authentic?	Weighted log-odds multi-signal fusion	Crossref, DataCite, Semantic Scholar APIs	Mean fused score ≥ 0.65
2	Fact Verification	Are claims evidentially supported?	Hybrid BM25 + TF-IDF + RRF retrieval	Trusted evidence corpus (50 passages, 21 domains)	Support rate $\geq 50\%$ AND mean conf. ≥ 0.35
3	Hallucination Detection	Is content likely fabricated?	Calibrated evidence-deficit scoring	Same trusted evidence corpus as CP2	Mean conf. ≥ 0.55 AND flag rate $< 50\%$
4	Bias & Fairness	Is discourse balanced and fair?	Multi-lexicon balanced rubric (5 lexicons)	Built-in lexicons — no external API	Balance score ≥ 0.40 AND subjectivity < 0.12

CP	Stage	Research Question	Method	Data Source	Pass Criterion
5	Academic Integrity	Is AI use disclosed and critically reviewed?	Weighted four-dimension integrity rubric	Student reflection metadata only	All three gate conditions satisfied
6	Reflection & Critical Thinking	Is reasoning substantive and evidence-based?	Epistemic vocabulary scoring + length check	Student reflection metadata only	≥ 2 sufficient answers AND ≥ 1 epistemic marker

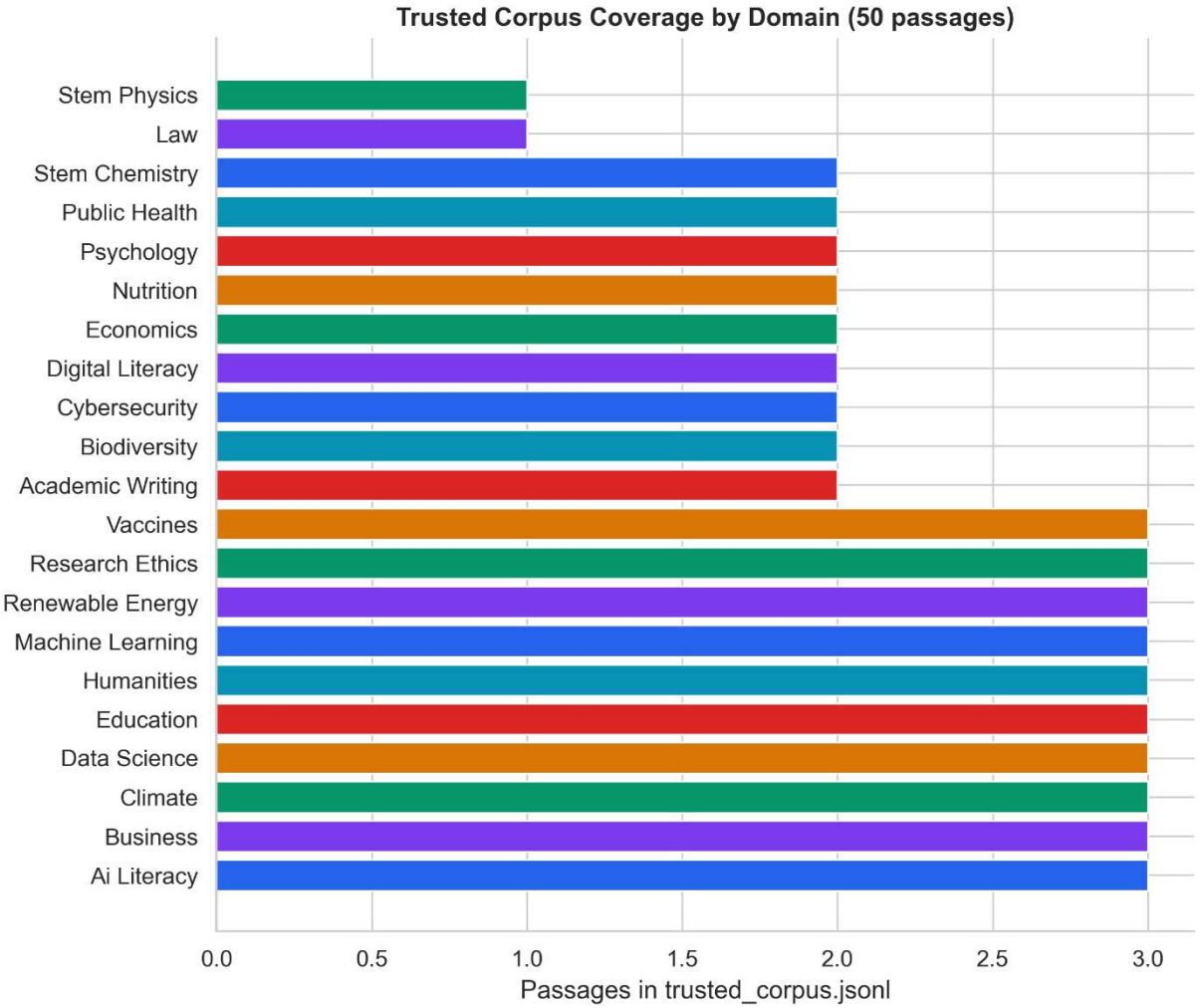


Figure 2: Domain distribution of the Phase 3 trusted evidence corpus (50 passages across 21 domains)

Table 2: Phase 3 evidence corpus composition across twenty-one domains and fifty passages.

Domain	Passages	Representative Sources
Climate Science	3	IPCC AR6 assessment summaries; NASA climate factsheets
Vaccines & Immunology	3	WHO and CDC immunization guidance documents
AI Literacy	3	UNESCO AI competency framework; academic integrity guidelines
Renewable Energy	3	IEA and NREL renewable energy technical reports
Machine Learning	3	LLM limitation studies; RAG methodology literature
Research Ethics	3	COPE publication ethics; preregistration and pilot-reporting norms
Education	3	Pedagogical theory and curriculum design literature

Domain	Passages	Representative Sources
Business	3	Management theory and organizational behavior sources
Humanities	3	Critical theory and textual-analysis scholarship
Data Science	3	Statistical methodology and applied analytics literature
Biodiversity	2	IPBES global extinction and ecosystem-service assessments
Nutrition	2	WHO dietary guideline documents
Public Health	2	WHO hydration and beverage consumption guidance
Digital Literacy	2	EU digital competence framework; misinformation verification guides
Cybersecurity	2	NIST cybersecurity framework and standards documentation
Psychology	2	Cognitive and behavioral science reference literature
Economics	2	Macroeconomic theory and policy-analysis sources
STEM — Chemistry	2	Chemical sciences reference material
Academic Writing	2	Graduate-level synthesis and evidence-integration guides
STEM — Physics	1	Physical sciences reference material
Law	1	Legal reasoning and case-analysis source material

4.2. End-to-End Classification Performance (Phase 3, N = 50): Addressing RQ1

Table 3 presents the end-to-end and per-checkpoint performance of ALCF on the Phase 3 evaluation dataset ($N = 50$). The framework achieved an overall accuracy of 0.980, precision of 0.964, recall of 1.000, F1-score of 0.982, and MCC of 0.960. The confusion matrix (Fig. 3) indicates 27 true positives, 22 true negatives, 1 false positive, and 0 false negatives: all twenty-seven approval-intended submissions were correctly approved, and twenty-two of the twenty-three rejection-intended submissions were correctly rejected. Notably, the framework maintained perfect recall, ensuring that no approval-intended submission was incorrectly rejected. The single false positive is examined in Section 4.6.

The Wilson 95% confidence interval for Phase 3 accuracy was [0.895,0.996], as illustrated in Fig. 4. Compared with earlier phases, this interval is narrower, reflecting increased statistical precision from progressive dataset expansion: Phase 1 ($N = 10$) exhibited considerable uncertainty due to its small sample; Phase 2 ($N = 20$) improved estimate stability; and Phase 3 ($N = 50$) further reduced uncertainty while maintaining high classification performance.

Table 3: End-to-end and per-checkpoint performance of ALCF-Full across the Phase 3 evaluation dataset (N=50).

Stage	Accuracy	Precision	Recall	F1	MCC	95% CI / Mean Score
End-to-end ($N = 50$)	0.980	0.964	1.000	0.982	0.960	Wilson [0.895, 0.996]
CP1 Source Verification	0.940	0.938	1.000	0.968	0.612	Mean = 0.860 (SD = 0.150)
CP2 Fact Verification	0.940	0.940	1.000	0.969	0.000†	—
CP3 Hallucination Detection	0.920	0.920	1.000	0.958	0.000†	—
CP4 Bias & Fairness	0.960	1.000	0.955	0.977	0.846	—
CP5 Academic Integrity	0.720	1.000	0.682	0.811	0.452	—
CP6 Reflection	0.720	1.000	0.689	0.816	0.426	—

†MCC is affected by checkpoint-level class imbalance resulting from the intentional corpus-coverage design of the evaluation dataset.

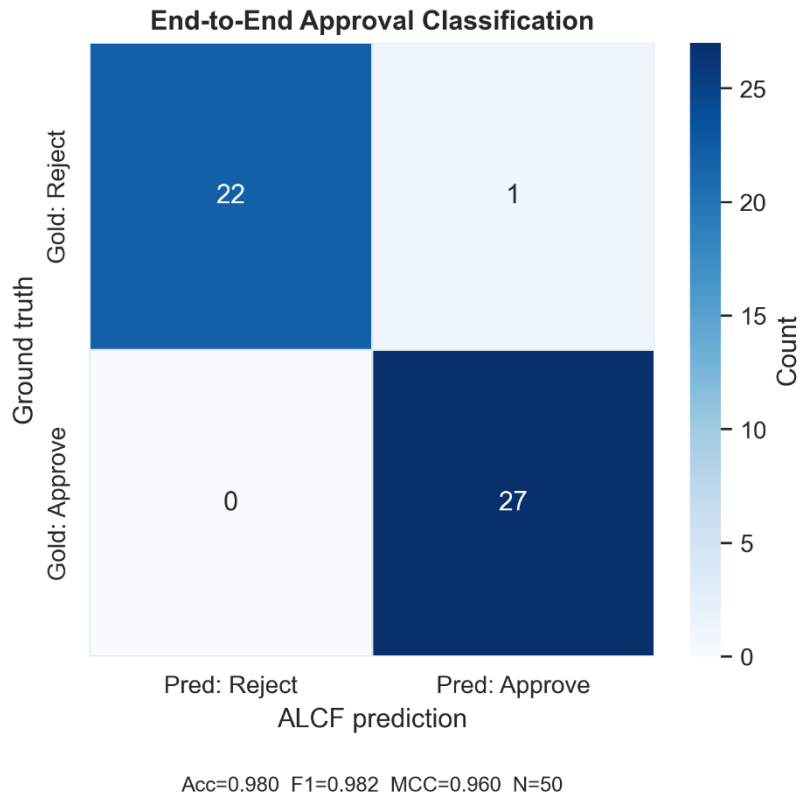


Figure 3: Confusion matrix for ALCF on the Phase 3 dataset (TP = 27, TN = 22, FP = 1, FN = 0)

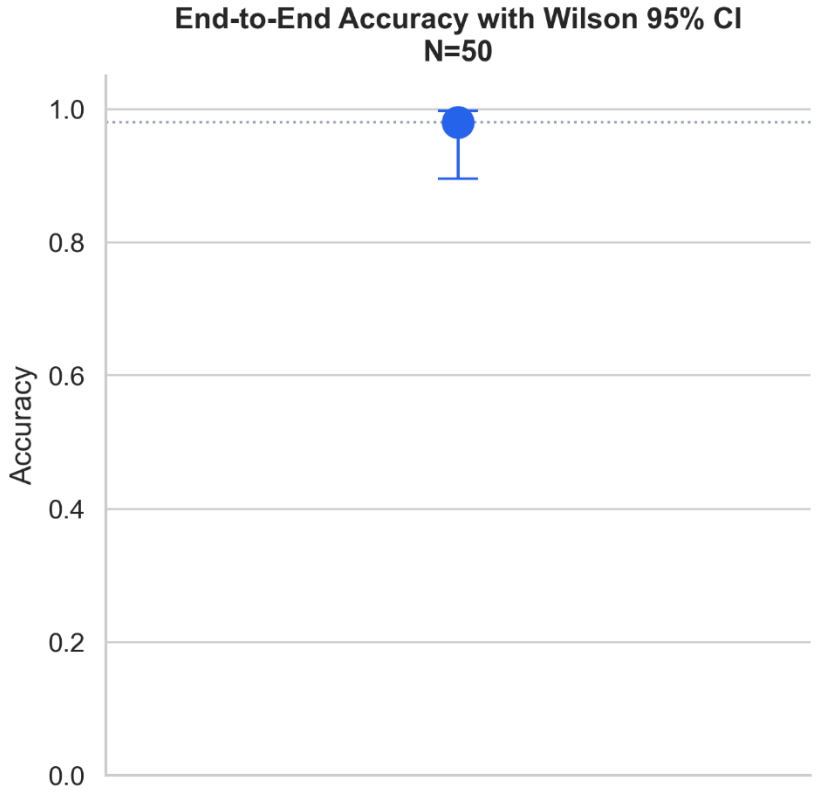


Figure 4: Phase 3 end-to-end accuracy with its Wilson 95% confidence interval (N=50)

4.3. Cross-Phase Accuracy and Confidence-Interval Trajectory: Addressing RQ1

Table 4 summarizes the performance, dataset scale, evidence-corpus composition, and methodological infrastructure across the three research phases. Three trends are apparent. First, end-to-end performance improved steadily: accuracy increased from 0.900 (Phase 1) to 0.950 (Phase 2) and 0.980 (Phase 3); precision (0.857, 0.909, 0.964), F1-score (0.923, 0.952, 0.982), and MCC (0.816, 0.905, 0.960) followed similar trends, while recall remained perfect (1.000) throughout. Second, Wilson 95% confidence intervals became increasingly precise: interval width decreased from approximately 0.44 (Phase 1, $N = 10$) to 0.227 (Phase 2, $N = 20$) and 0.101 (Phase 3, $N = 50$), consistent with the well-known narrowing of binomial confidence intervals with increasing sample size. Third, the framework's error profile remained stable: despite substantial growth in dataset size and domain diversity, the number of false positives did not increase, and the persistence of one false-positive case across all three phases suggests a recurring structural pattern rather than random noise (examined in Sections 4.6 and 5.1).

Table 4: Cross-phase comparison of dataset scale, evidence corpus composition, performance metrics, and methodological infrastructure across Phases 1–3.

Metric	**Phase 1 ($N = 10$)**	**Phase 2 ($N = 20$)**	**Phase 3 ($N = 50$)**
Samples (N)	10	20	50
Corpus passages	12	16	50
Corpus domains	7	11	21
End-to-end accuracy	0.900	0.950	0.980
End-to-end precision	0.857	0.909	0.964
End-to-end recall	1.000	1.000	1.000
End-to-end F1	0.923	0.952	0.982
End-to-end MCC	0.816	0.905	0.960
Wilson 95% CI (low)	≈ 0.55	0.764	0.895
Wilson 95% CI (high)	≈ 0.99	0.991	0.996
True positives (TP)	6	10	27
True negatives (TN)	3	9	22
False positives (FP)	1	1	1
False negatives (FN)	0	0	0
CP1 accuracy	—	0.95	0.94
CP2 accuracy	—	0.90	0.94
CP3 accuracy	—	0.90	0.92
CP4 accuracy	—	0.95	0.96
CP5 accuracy	—	0.70	0.72
CP6 accuracy	—	0.75	0.72
Figures exported	12	12	13
Tables exported	6	6	7
External baselines	No	No	Yes (6 systems)
Approval strategies	1 (conjunction)	1 (conjunction)	4 strategies
Formal decision model	No	No	Yes

Metric	Phase 1 ($N = 10$)	Phase 2 ($N = 20$)	Phase 3 ($N = 50$)
Expert ratings	No	No	Template (illustrative)
Excel export	No	Yes	Yes
Unit tests	≈ 20	≈ 24	33

4.4. Per-Checkpoint Classification Performance and the CP5/CP6 Plateau: Addressing RQ2

Per-checkpoint classification performance on the Phase 3 evaluation set ($N = 50$) is summarized in Table 5; corresponding F1-scores are illustrated in Fig. 5, and pass rates in Fig. 6. A distinct performance continuum is seen across the six checkpoints. The technical checkpoints (CP1–CP4) consistently achieved high performance, with accuracy between 0.920 and 0.960 and F1-scores between 0.958 and 0.977; CP4 performed best overall (accuracy = 0.960, F1 = 0.977). CP1 achieved an accuracy of 0.940 and a mean continuous score of 0.860 (SD = 0.150). CP2 and CP3 achieved similar classification performance, although their MCC values were affected by checkpoint-level class imbalance arising from the intentional corpus-coverage design.

CP5 and CP6 were the weakest-performing checkpoints, both achieving an accuracy of 0.720, with F1-scores of 0.811 and 0.816, respectively. Both produced zero false positives (perfect precision) but many false negatives (lower recall), suggesting that CP5/CP6 thresholds are conservative: most submissions classified as failures genuinely failed, but some submissions that met the intended learning outcomes were also classified as failures. Qualitative examination shows that most false negatives arise from simple threshold-boundary phenomena rather than basic reasoning errors — e.g., a reflection falling just short of minimum length, a disclosure that references an AI-assisted process without naming an identifiable tool, or evidence-based reflection lacking epistemic keywords despite demonstrating meaningful thinking. Importantly, this pattern showed only minor change between Phase 2 ($N = 20$) and Phase 3 ($N = 50$), indicating that the accuracy ceiling reflects rubric calibration rather than sample size, and that further constructed samples alone are unlikely to significantly improve CP5/CP6 performance without revision of the disclosure and reflection scoring rubrics. This finding directly answers RQ2 by identifying CP5 and CP6 as the primary points for future calibration and human validation.

Table 5: Per-checkpoint confusion-matrix statistics, classification performance, and mean checkpoint scores for the Phase 3 dataset (N=50).

Checkpoint	TP	TN	FP	FN	Accuracy	F1	Mean Score (SD)
CP1 Source Verification	45	2	3	0	0.940	0.968	0.860 (0.150)
CP2 Fact Verification	47	0	3	0	0.940	0.969	1.000 (0.000)
CP3 Hallucination Detection	46	0	4	0	0.920	0.958	1.000 (0.000)
CP4 Bias & Fairness	42	6	0	2	0.960	0.977	0.733 (0.210)
CP5 Academic Integrity	30	6	0	14	0.720	0.811	0.700 (0.310)
CP6 Reflection	31	5	0	14	0.720	0.816	0.532 (0.290)

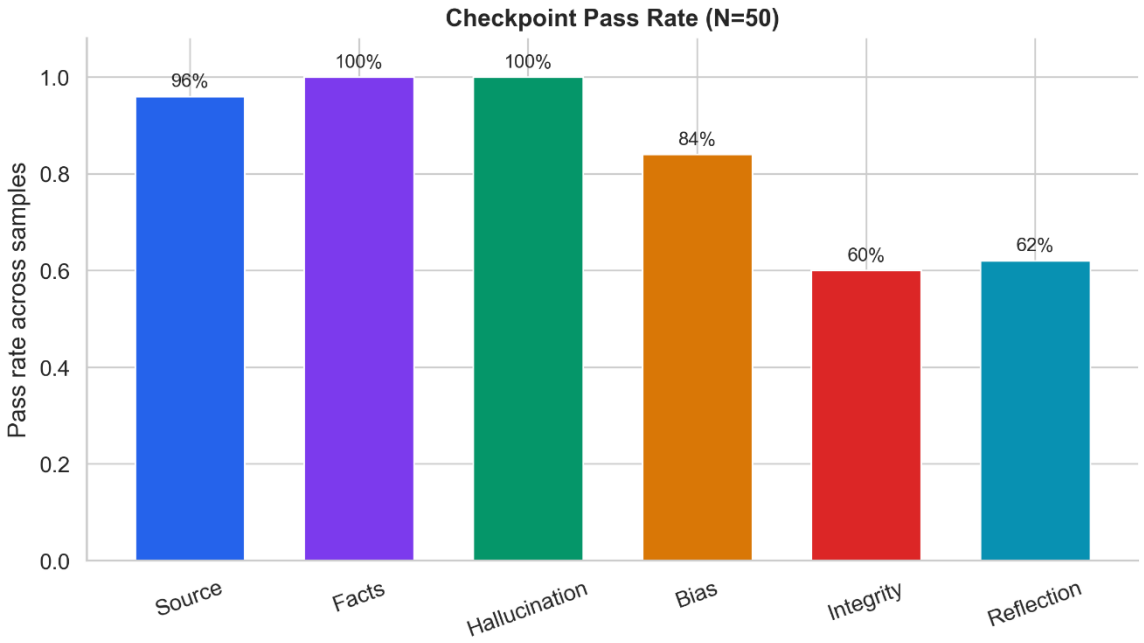


Figure 5: Per-checkpoint F1-scores for CP1–CP6 on the Phase 3 evaluation dataset

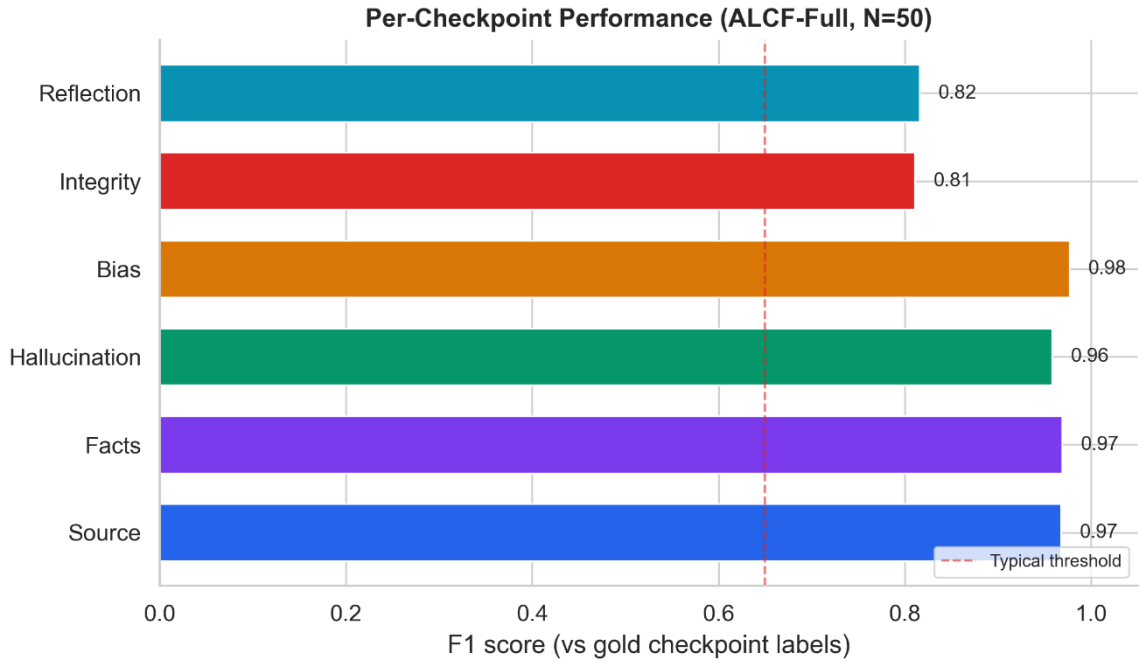


Figure 6: Per-checkpoint pass rates showing the proportion of submissions receiving a PASS verdict for each checkpoint

4.5. Score Distributions and Gold-Label Agreement

Fig. 7 and Fig. 8 provide distributional perspectives on checkpoint behavior and overall predictive agreement, complementing the classification metrics in Sections 4.2–4.4. As illustrated in Fig. 7, the score distributions of CP2 and CP3 cluster around the higher end of the scale, reflecting the deliberate corpus-coverage design of the evaluation corpus, where a high percentage of submissions presented factually supported content even when other checkpoint dimensions were not fulfilled. CP1 and CP4 show moderately dispersed distributions, reflecting variation in citation quality and rhetorical balance, while continuing to achieve high classification accuracy. CP5

and CP6 score distributions are wider and have lower central tendencies, consistent with Table 5 and further confirming that disclosure and reflection are the most difficult parts of the framework to automate.

Fig. 8 shows a high level of agreement between gold labels and pipeline predictions: of the 50 evaluated submissions, 98% received matching classifications. The single disagreement corresponds to the sole false-positive case identified in the end-to-end evaluation (Section 4.6), further supporting the framework's overall reliability and the findings associated with RQ1.

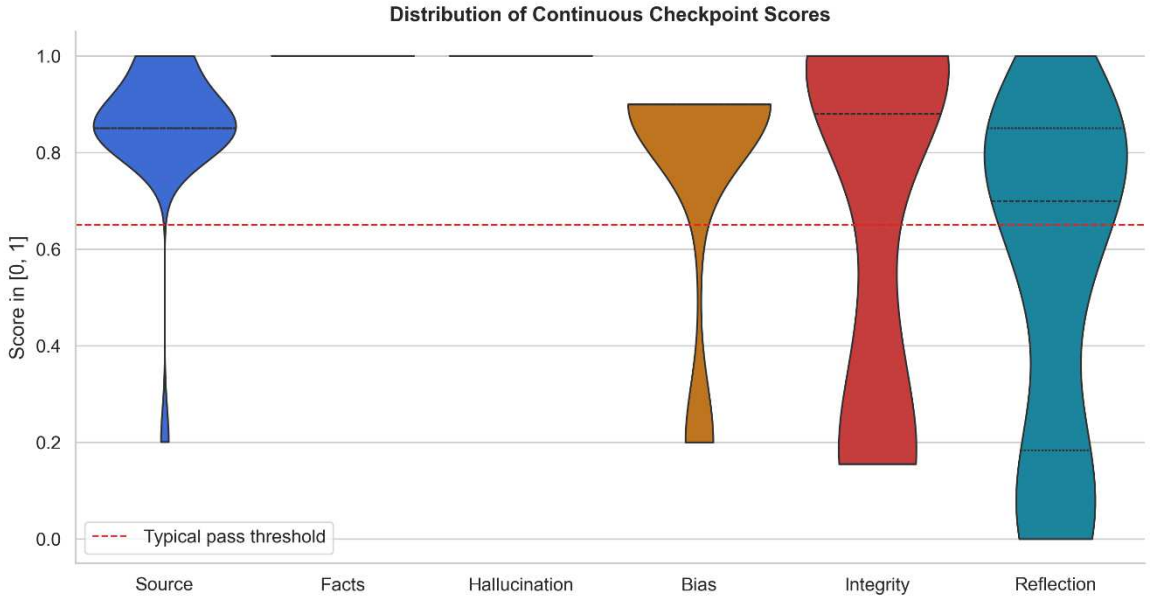


Figure 7: Distribution of continuous checkpoint scores across all Phase 3 submissions

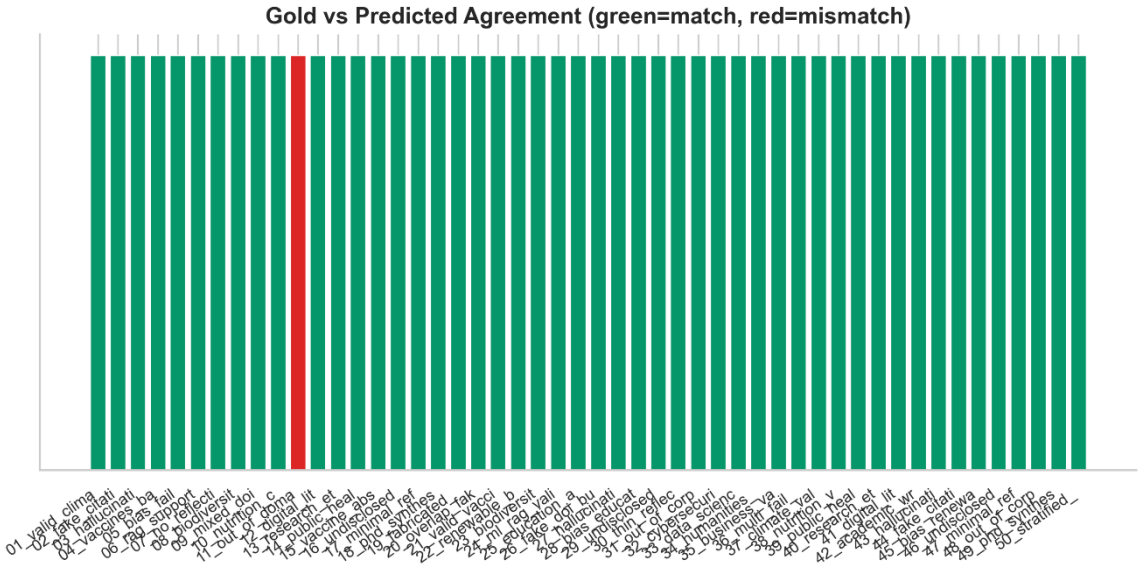


Figure 8: Agreement between gold labels and end-to-end pipeline predictions on the Phase 3 dataset

4.6. The Persistent False-Positive Case: A Structural Pattern

One interesting result across all three evaluation phases is a persistent false-positive classification: one reject-intended submission was incorrectly approved in Phase 1 ($N = 10$), Phase 2 ($N = 20$), and Phase 3 ($N = 50$) (Table 4). The persistence of this error across increasing dataset size and domain coverage indicates a structural property of the decision process rather than a sampling artifact.

As illustrated in Fig. 9, the false-positive case consistently meets all four technical checkpoints (CP1–CP4) with good margins — genuine citations, factual support, low hallucination risk, and acceptable rhetorical balance — while its CP5 and/or CP6 scores are remarkably close to the decision thresholds: high enough to pass, yet judged reject-intended by human annotators. This illustrates a key limitation of threshold-based decision making under conjunction approval, checkpoint outcomes are binarized (PASS/FAIL) prior to aggregation, so a score marginally above threshold is treated identically to a score far above threshold. Borderline cases that narrowly satisfy checkpoint requirements may therefore be approved despite lacking sufficient characteristics from a human evaluator's perspective.

This error motivates the alternative approval strategies presented in Section 3.10. Weighted and confidence-based approaches use continuous checkpoint scores rather than only binary threshold outcomes and may therefore be better suited to identifying submissions that are weak in integrity and reflection despite strong performance elsewhere.

Table 6 summarizes the four approval strategies formally specified and implemented within the proposed framework. The conjunction strategy served as the default approval mechanism for all primary results reported in this study, achieving strong end-to-end performance and providing a conservative decision policy designed to maximize precision. The remaining strategies — majority voting, weighted fusion, and confidence-based approval — were implemented to support future comparative evaluation but have not yet been empirically validated on the Phase 3 dataset; they are reported as formally specified decision models rather than experimentally verified alternatives.

Table 6: Approval strategies implemented within the framework and their corresponding decision rules.

Strategy	Decision Rule	Expected Behavior / Status
Conjunction (default)	Approve if all six checkpoints pass: $A = \prod_i p_i$	Maximizes precision; default strategy; validated on $N = 50$ pilot (this study)
Majority	Approve if $\geq k$ of 6 checkpoints pass ($k = 4$)	More permissive; expected to raise recall of borderline approvals at the cost of precision
Weighted	Approve if weighted score fusion exceeds threshold $\theta = 0.72$	Allows partial compensation between checkpoints; sensitive to weight calibration
Confidence	Approve if mean confidence $\geq \tau_c$ AND minimum pass count $k_{\min}$ satisfied	Combines score magnitude and coverage; requires calibration of τ_c and $k_{\min}$

4.7. Inter-Rater Reliability: Illustrative Demonstration — Addressing RQ6

Table 7 presents the results of the inter-rater reliability module (Section 3.13), evaluated using an illustrative five-sample template independently rated by three simulated assessors to verify the correctness of the reliability calculations implemented within the framework. For both CP5 and CP6, the module produced a Fleiss' Kappa of 1.000 and 100% agreement. Under conventional interpretation guidelines this would be classified as almost perfect; however, these results should not be interpreted as evidence of genuine rubric reliability, as the underlying ratings were intentionally constructed for software-validation purposes rather than collected from independent human assessors. The exercise was methodological rather than empirical: it demonstrates that the reliability infrastructure correctly computes agreement statistics and can support future validation studies involving multiple assessors. Establishing the true reliability of CP5 and CP6 requires blind evaluation by independent academic experts across multiple disciplines, remaining a key objective for future work and addressing RQ6 at the level of methodological readiness.



Figure 9: Sample-by-checkpoint pass/fail matrix for the Phase 3 evaluation dataset, highlighting the persistent false-positive

Table 7: Illustrative inter-rater reliability results for CP5 and CP6 using the validation template.

Checkpoint	N Samples	N Raters	Fleiss' Kappa	Percent Agreement
CP5 Academic Integrity	5	3	1.000	1.000
CP6 Reflection	5	3	1.000	1.000

4.8. Ablation Study: Retrieval Strategy Comparison — Addressing RQ3

Table 8 presents the retrieval ablation study conducted on the Phase 3 evaluation set ($N = 50$), isolating the retrieval component while holding the six-checkpoint architecture and conjunction approval strategy constant. All three retrieval configurations achieved identical end-to-end performance (accuracy = 0.980, precision = 0.964, F1 = 0.982, MCC = 0.960; Fig. 10). This result indicates that retrieval performance was not a limiting factor for the Phase 3 dataset, as the evaluation corpus was constructed to allow claims to match supporting evidence passages as closely as possible; both lexical (BM25) and vector-space (TF-IDF) retrieval are equally effective at locating correct evidence under these conditions, yielding identical downstream classification. The default framework configuration is retained as the hybrid BM25–TF-IDF model for theoretical reasons, given its established robustness in the retrieval literature and its blending of complementary lexical and semantic signals. Future testing with larger, more heterogeneous evidence corpora and authentic student work will be required to determine whether hybrid retrieval offers meaningful benefits under more realistic conditions.

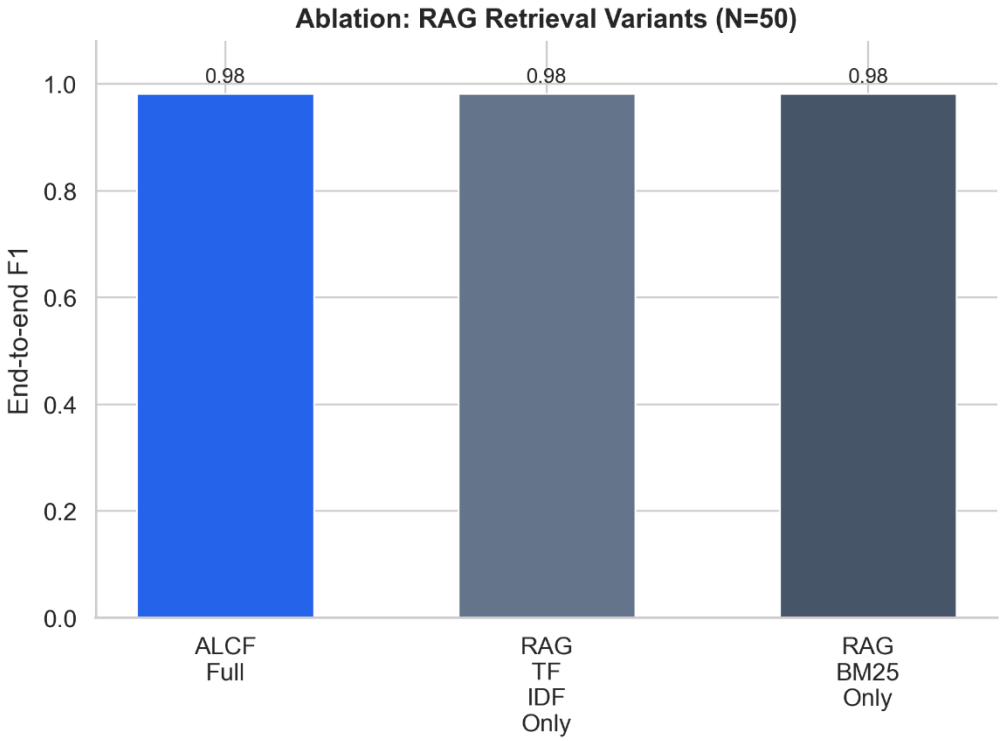


Figure 10: End-to-end F1-score comparison across the three retrieval configurations evaluated in the ablation study

Table 8: Retrieval ablation study comparing hybrid, TF-IDF-only, and BM25-only configurations on the Phase 3 dataset.

Variant	Description	Accuracy	Precision	F1	MCC
ALCF-Full	Hybrid BM25 + TF-IDF + Reciprocal Rank Fusion (default)	0.980	0.964	0.982	0.960
RAG-TF-IDF-Only	Dense-sparse TF-IDF vector retrieval, BM25 disabled	0.980	0.964	0.982	0.960
RAG-BM25-Only	Lexical BM25 probabilistic retrieval, TF-IDF disabled	0.980	0.964	0.982	0.960

4.9. Comparison Against Internal Baseline Systems: Addressing RQ4

Table 9 shows that ALCF-Full outperformed five internal baseline systems, all evaluated on the same submissions and evidence corpus, differing only in checkpoint composition and approval strategy. ALCF-Full achieved 0.980 accuracy, 0.982 F1, and 0.960 MCC. Weighted-Fusion-6CP ranked second (0.880 accuracy, 0.778 MCC); although it retained all six checkpoints, continuous score fusion did not match the discriminative capacity of the conjunction-based rule under the present evaluation conditions. All remaining baselines performed substantially worse: Majority-Vote-6CP achieved 0.700 accuracy, indicating that permitting approval despite multiple checkpoint failures reduced discriminative power; Citation-Only and Technical-3CP produced identical results (accuracy = 0.580), showing that technical verification alone is insufficient for reliable discrimination; and Retrieval-Only performed worst (accuracy = 0.540, MCC = 0.000), confirming that factual grounding and hallucination detection alone cannot provide a complete assessment of AI literacy.

These differences (Fig. 11) support the central architectural claim of this study: technical verification mechanisms must be combined with higher-order evaluative dimensions — bias awareness, academic integrity, and reflective reasoning — to assess AI literacy effectively. The results also reveal an important trade-off: ALCF-Full's conjunction strategy performs best because it requires competency across all checkpoint dimensions simultaneously, but this strict decision policy also contributes to the persistent borderline false-positive case discussed in Section 4.6. On a general level, the baseline comparison directly confirms that the six-checkpoint conjunction architecture has significant benefits over simpler architectures, addressing RQ4.

Table 9: Performance comparison between ALCF-Full and five internal baseline systems on the Phase 3 dataset.

System	Description	Accuracy	F1	MCC
ALCF-Full	Complete six-checkpoint pipeline, conjunction approval	0.980	0.982	0.960
Weighted-Fusion-6CP	All six checkpoints, weighted score fusion	0.880	0.900	0.778
Majority-Vote-6CP	All six checkpoints, approve if ≥ 4 of 6 pass	0.700	0.783	0.473
Citation-Only	CP1 (source verification) used in isolation	0.580	0.720	0.221
Technical-3CP	CP1 + CP2 + CP3 (technical checkpoints only)	0.580	0.720	0.221
Retrieval-Only	CP2 + CP3 (fact verification + hallucination only)	0.540	0.701	0.000

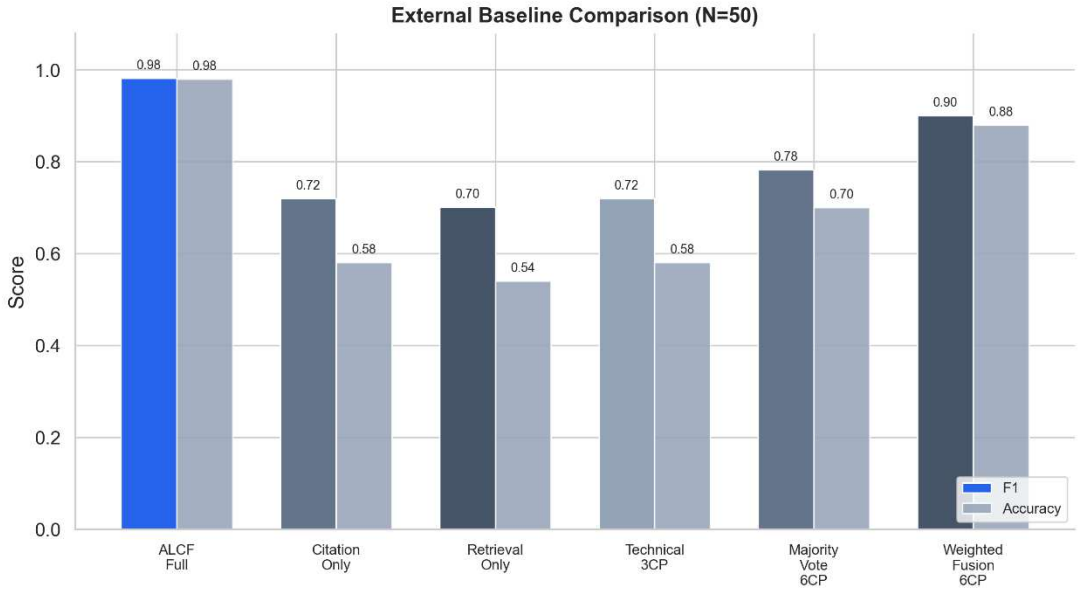


Figure 11: End-to-end accuracy and F1-score comparison of ALCF-Full and the internal baseline systems

4.10. Composite Results Summary

Fig. 12 and Fig. 13 provide an integrated overview of the principal Phase 3 findings. Fig. 12 combines key visualizations presented throughout this chapter, including the confusion matrix, per-checkpoint performance, checkpoint outcome patterns, and dataset composition, serving as a consolidated summary of framework behavior and performance across the complete evaluation pipeline. Fig. 13 summarizes the composition of the Phase 3 dataset: twenty-seven approve-intended and twenty-three reject-intended submissions, with the reject-intended subset spanning multiple failure categories, including fabricated citations, hallucinated content, biased reasoning, inadequate reflection, and deliberately constructed borderline or multi-error cases. This diversity of error types provided a challenging evaluation environment and ensured that all six checkpoints were exercised throughout the benchmark process.

Taken together, the results presented in Sections 4.2–4.10 demonstrate that the proposed framework achieves strong end-to-end performance while providing interpretable checkpoint-level diagnostics. The findings further indicate that integrating technical verification with rubric-based assessment is essential for reliable AI literacy evaluation.

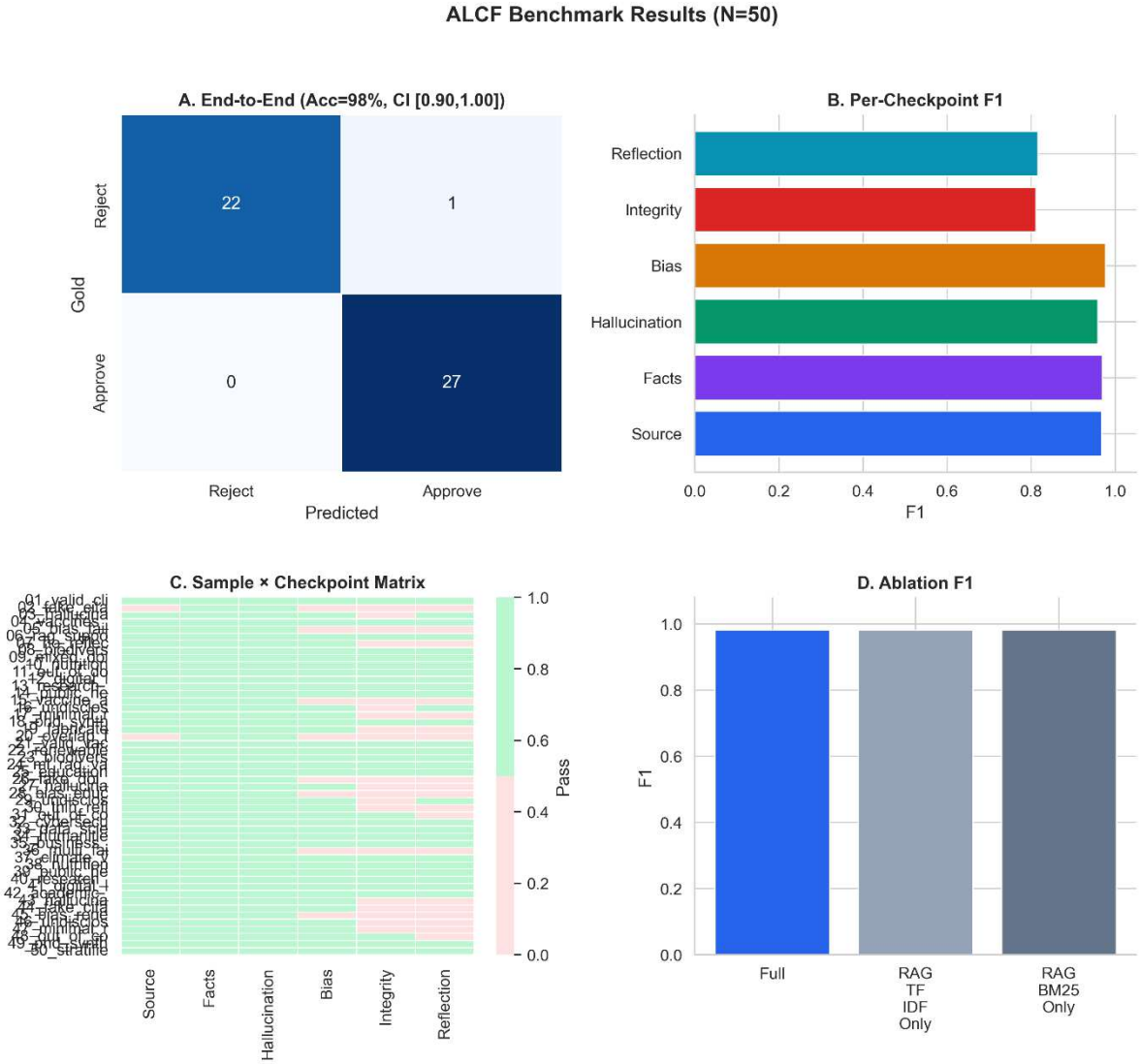


Figure 12: Composite summary of principal Phase 3 results across framework performance and checkpoint outcomes

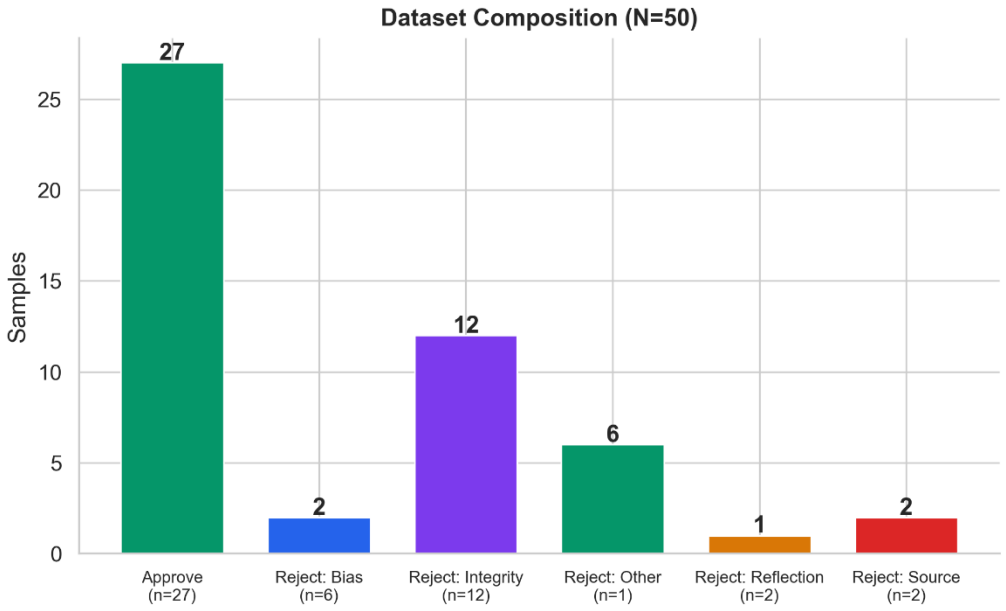


Figure 13: Composition of the Phase 3 evaluation dataset by approval status and primary error category

5. Discussion

5.1. The Conjunction Rule: A Precision–Compensation Trade-off

The results reveal a fundamental trade-off within the ALCF design. The conjunction approval rule contributes directly to the framework's strongest result — superior performance relative to all internal baselines (Section 4.9) — while also explaining the persistent false-positive case observed across all three phases (Section 4.6). Because approval requires every checkpoint to pass, the framework cannot distinguish between a score that marginally exceeds a threshold and one that exceeds it by a wide margin. This strictness improves precision but limits the system's ability to account for borderline cases. Rather than indicating a design flaw, it reflects an explicit policy choice that prioritizes minimizing false approvals in academic-integrity contexts.

5.2. Why CP5 and CP6 Plateau While CP1–CP4 Improve

Cross-phase results show continuous improvement in CP1–CP4, whereas CP5 and CP6 remain stable at approximately 0.70–0.72 accuracy. This difference reflects the nature of the underlying tasks. CP1–CP4 evaluate constructs verifiable against external evidence, allowing performance to improve as corpus coverage and dataset diversity increase. CP5 and CP6, by contrast, rely on fixed rubric thresholds governing disclosure and reflection quality; additional constructed samples expose threshold limitations rather than improving performance. Future gains will therefore depend on expert-informed rubric calibration and inter-rater validation rather than further dataset expansion alone.

5.3. Theoretical Contribution: An Automation Gradient

A key theoretical contribution of this study is the identification of an automation gradient across AI literacy dimensions. Technical verification tasks — citation validation, fact verification, and hallucination detection — demonstrated strong and scalable automated performance. In contrast, metacognitive dimensions such as integrity disclosure and reflective reasoning remain constrained by rubric design and subjective interpretation. Different dimensions of AI literacy therefore require different validation strategies and should not be assumed equally amenable to automation.

5.4. Value and Limits of Internal Baseline Comparisons

The baseline analysis demonstrates that the complete six-checkpoint architecture outperforms all simplified alternatives, evidencing that effective AI literacy assessment requires integrating technical verification with higher-order evaluative checkpoints. However, because the framework and evaluation dataset were developed within the same research programme, the comparison should be interpreted as evidence of internal validity rather than an independent external benchmark. Validation using externally constructed datasets remains necessary.

5.5. Practical Implications for Higher Education

The findings have several practical implications. First, the checkpoint-based design provides transparent, actionable feedback, enabling students to identify specific areas requiring improvement. Second, the observed improvement across phases suggests that institutions can increase reliability through continued corpus expansion and local data curation. Third, the persistent false-positive case highlights the value of targeted human review for submissions with near-threshold integrity or reflection scores, offering a practical hybrid assessment model.

5.6. Relation to Prior Automated Assessment Systems

The proposed framework extends prior fact-verification and retrieval-based approaches by integrating source verification, hallucination detection, bias assessment, academic integrity, and reflective reasoning within a unified evaluation architecture. Unlike AI-text detection systems, it focuses on the quality, transparency, and critical use of AI-assisted work rather than attempting to determine whether text was AI-generated. The three-phase evaluation, combined with baseline and ablation analyses, provides a broader empirical foundation than is typically reported for comparable AI literacy assessment systems.

6. Limitations

This study has several limitations that should be considered when interpreting the findings.

- **Constructed Dataset:** All evaluation samples across the three phases were researcher-constructed rather than authentic student submissions. Although Phase 2 and Phase 3 introduced borderline and multi-error cases, real student work is likely to exhibit more complex and unpredictable error patterns. Validation using authentic, anonymized student submissions remains the highest priority.
- **Sample Size and Statistical Power:** Despite progressive expansion from $N = 10$ to $N = 50$, the Phase 3 Wilson 95% confidence interval remains broad. Results should therefore be interpreted as descriptive rather than inferential; formal significance testing was beyond the scope of this study.
- **Evidence Corpus Coverage:** The evidence corpus expanded to 50 passages across 21 domains; however, coverage remains limited relative to the diversity of academic disciplines and claim types encountered in practice. Performance may decline for claims outside the represented domains.
- **Rubric Validation:** CP5 and CP6 were developed without calibration against independent human assessors. Although inter-rater reliability infrastructure was implemented, empirical validation with genuine instructors has not yet been conducted.
- **Approval Strategy Evaluation:** While conjunction, majority, weighted, and confidence-based approval strategies were formally specified, only the conjunction strategy was empirically evaluated. The effectiveness of alternative strategies remains an open research question.
- **External API Dependence:** CP1 relies on external services, including Crossref and Semantic Scholar, whose coverage, availability, and rate limits may affect performance in operational settings.
- **Internal Baseline Circularity:** Baseline systems were evaluated on a dataset designed by the same research team that developed the framework, introducing potential construction-validity bias.
- **Lack of Independent Replication:** Framework design, corpus construction, and evaluation were conducted by a single research team without external replication or adversarial testing.

- **Limited Generalizability of Cross-Phase Trends:** The observed improvement from Phase 1 to Phase 3 is encouraging but is based on only three progressively expanded datasets. The CP5/CP6 plateau suggests that further gains are unlikely without rubric recalibration and authentic-data validation.

7. Future Work

The next phase of this research aims to transition ALCF from controlled evaluation using constructed datasets to large-scale validation in authentic educational settings. While the present study establishes the technical feasibility and methodological foundation of the framework, broader empirical validation is required to demonstrate its robustness, generalizability, and practical applicability across diverse higher education contexts. Several research directions emerge from the findings of this study:

- **Authentic Student Dataset:** Evaluate ALCF using a large corpus (100–500) of anonymized AI-assisted student submissions collected across multiple institutions and academic disciplines.
- **Inter-Rater Validation:** Conduct a multi-instructor annotation study to establish the reliability of CP5 and CP6, enabling empirical calibration of rubric thresholds using genuine expert judgements.
- **Approval Strategy Evaluation:** Empirically compare the conjunction, majority, weighted, and confidence-based approval strategies to determine their effectiveness and trade-offs under authentic assessment conditions.
- **Evidence Corpus Expansion:** Extend the evidence corpus to several hundred discipline-specific passages through collaboration with academic libraries and subject experts.
- **Independent and Adversarial Evaluation:** Validate the framework using independently constructed and adversarial benchmark datasets to assess robustness, reproducibility, and generalizability.
- **LLM-Assisted Rubric Assessment:** Investigate the use of large language models to evaluate academic integrity and reflective writing, enabling more context-aware assessment than fixed keyword- and threshold-based rules.
- **Longitudinal Educational Deployment:** Deploy ALCF as a formative assessment tool within undergraduate courses to examine its impact on students' AI literacy development, critical thinking, and responsible AI use over time.
- **Statistical Validation:** Apply inferential statistical analyses (e.g., McNemar's and Friedman tests) once sufficiently large authentic datasets become available.

Overall, the primary objective of Phase 4 is to transform ALCF from a validated research prototype into a robust, institution-ready AI literacy assessment framework, strengthening both the scientific validity and practical adoption of the framework for explainable AI-assisted assessment in higher education.

8. Conclusion

This study introduced the AI Literacy Checkpoint Framework (ALCF), a six-checkpoint automated framework for evaluating AI-assisted academic submissions through source verification, fact verification, hallucination detection, bias and fairness analysis, academic integrity assessment, and reflective thinking. The framework integrates multi-source bibliographic verification, hybrid retrieval, evidence-based validation, rubric-driven assessment, and formal approval strategies into a transparent and explainable evaluation pipeline.

A three-phase process gradually increased the number of submissions from 10 to 50 and the number of evidence domains from 7 to 21. In Phase 3 ($N = 50$), ALCF achieved 98.0% accuracy, 96.4% precision, 100% recall, an F1-score of 98.2%, and an MCC of 0.960. Per-checkpoint analysis revealed high and stable accuracy for the technical verification checkpoints (CP1–CP4), whereas the rubric-based checkpoints (CP5–CP6) achieved approximately 0.72 accuracy, reflecting the different characteristics of automated technical verification and subjective metacognitive assessment.

Comparative analyses showed that the complete six-checkpoint model outperformed all simplified internal baselines, demonstrating the value of integrating technical verification with higher-order evaluation. Although the

retrieval ablation produced similar performance across configurations on the constructed dataset, this result is attributable to high lexical overlap rather than general equivalence among retrieval methods. The formal approval strategies and inter-rater reliability infrastructure further extend the framework and provide a foundation for future empirical validation.

The study's principal contributions are the integrated six-checkpoint architecture, the staged evaluation design, the comparison with simplified baseline configurations, the formalization of alternative approval strategies, the inter-rater reliability infrastructure for subjective checkpoints, and the explicit separation of demonstrated capabilities from those requiring validation with authentic student work.

In conclusion, this study provides an empirical, transparent, and modular framework for automated AI literacy assessment in higher education. Future research should transition from constructed benchmarks to authentic student submissions, supported by expert rubric calibration, expanded evidence corpora, and empirical testing of alternative approval strategies. These developments will strengthen the framework's validity and support its implementation as an explainable AI literacy assessment system.

Declarations

Funding: This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00256629) grant funded by the Korea government (MSIT).

Competing Interests: The authors have no relevant financial or non-financial interests to disclose.

Ethics approval: Not applicable. This study did not involve human participants or animals; all evaluation submissions were researcher-constructed.

Consent to participate: Not applicable.

Consent for publication: Not applicable.

Data, Materials and/or Code availability: The constructed evaluation dataset (`alcf_samples.jsonl`, $N = 50$), the trusted evidence corpus (`trusted_corpus.jsonl`, 50 passages across 21 domains), the expert-ratings template, and the supporting Phase 3 research outputs (figures, tables, and JSON reports) generated during this study are available from the corresponding author on reasonable request.

Authors' contributions: Muhammad Subhan conducted the study, implemented the ALCF framework, analyzed the results, and drafted the manuscript. Ali Raza contributed to the methodology and manuscript revision. Soonja Yeom supervised the research and contributed to its conceptualization and critical revision. Soo-Hyung Kim and Hyung-Jeong Yang contributed to the conceptual development, analysis, and critical review of the manuscript. All authors approved the final manuscript.

References

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: Complete edition*. Addison Wesley Longman.

Ansari, S. (2026). Compound deception in elite peer review: A failure mode taxonomy of 100 fabricated citations at NeurIPS 2025. *arXiv:2602.05930*.

Azevedo, L., Robles, P., Best, E., & Mallinson, D. J. (2025). Institutional policies on artificial intelligence in higher education: Frameworks and best practices for faculty. *New Directions for Adult and Continuing Education*, 2025(188), 70–78.

Bender, E. M., Gebru, T., McMillan-Major, A., & Mitchell, M. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., et al. (2021). On the opportunities and risks of foundation models. *arXiv:2108.07258*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.

Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading Wikipedia to answer open-domain questions. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 1870–1880.

Chen, X., Zou, D., Xie, H., Cheng, G., & Liu, C. (2022). Two decades of artificial intelligence in education. *Educational Technology & Society*, 25(1), 28–47.

Cignarella, A. T., Giachanou, A., & Lefever, E. (2025). A survey on stereotype detection in natural language processing. *ACM Computing Surveys*, 58(5), 1–33.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.

Cormack, G. V., Clarke, C. L., & Buettcher, S. (2009). Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 758–759.

Dai, Z., & Callan, J. (2020). Context-aware document term weighting for ad-hoc search. *Proceedings of The Web Conference 2020*, 1897–1907.

Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., et al. (2021). Documenting large webtext corpora: A case study on the Colossal Clean Crawled Corpus. *Proceedings of EMNLP 2021*, 1286–1305.

Falke, T., Ribeiro, L. F., Utama, P. A., Dagan, I., & Gurevych, I. (2019). Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. *Proceedings of ACL 2019*, 2214–2220.

Fan, J., Wang, H., & Gu, X. (2025). Development and validation of a framework and scale for primary and secondary school teachers' data-artificial intelligent competence. *Frontiers in Psychology*, 16, 1756893.

Ferrara, E. (2024). GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science*, 7(1), 549–569.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382.

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2021). How to design AI for social good: Seven essential factors. In *Ethics, Governance, and Policies in Artificial Intelligence* (pp. 125–151). Springer.

Franzoi, L., Cembrani, V., Mulatti, C., & Treccani, B. (2025). Retrieval practice enhances learning in real primary school settings, whether distributed or not. *Frontiers in Psychology*, 16, 1632206.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., et al. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3), 1097–1179.

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1), 6.

Gonsalves, C. (2026). Generative AI's impact on critical thinking: Revisiting Bloom's taxonomy. *Journal of Marketing Education*, 48(1), 4–19.

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., et al. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), 1–55.

Kaddoura, M. (2013). Think pair share: A teaching-learning strategy to enhance students' critical thinking. *Educational Research Quarterly*, 36(4), 3–24.

Kizilcec, R. F. (2016). How much information? Effects of transparency on trust in an algorithmic interface. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2390–2395.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.

López-Pernas, S., Oliveira, E., Song, Y., & Saqr, M. (2025). AI, explainable AI and evaluative AI: Informed data-driven decision-making in education. In *Advanced Learning Analytics Methods: AI, Precision and Complexity* (pp. 17–39). Springer Nature Switzerland.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta*, 405(2), 442–451.

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. *Proceedings of ACL 2020*, 1906–1919.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.

Perkins, M. (2023). Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching and Learning Practice*, 20(2), 1–24.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., & Miller, A. (2019). Language models as knowledge bases? *Proceedings of EMNLP-IJCNLP 2019*, 2463–2473.

Robertson, S., & Zaragoza, H. (2009). *The probabilistic relevance framework: BM25 and beyond* (Vol. 4). Now Publishers.

Sakai, Y., Kamigaito, H., & Watanabe, T. (2026). Hallucitation matters: Revealing the impact of hallucinated references with 300 hallucinated papers in ACL conferences. *arXiv:2601.18724*.

Salton, G., & McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill Education.

Shukor, N. A. F., & Osman, K. (2025). The role of artificial intelligence in developing critical thinking among science students: A systematic literature review. *International Journal of Academic Research in Progressive Education and Development*, 14(3), 164–176.

Singh, A., Ehtesham, A., Kumar, S., Khoei, T. T., & Vasilakos, A. V. (2025). Agentic retrieval-augmented generation: A survey on agentic RAG. *arXiv:2501.09136*.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285.

Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: A large-scale dataset for fact extraction and verification. *Proceedings of NAACL-HLT 2018*, 809–819.

Tomar, A., Murthy, R., & Bhattacharyya, P. (2025). Stereotype detection as a catalyst for enhanced bias detection: A multi-task learning approach. *Findings of the Association for Computational Linguistics: ACL 2025*, 17304–17317.

Touretzky, D., Gardner-McCune, C., Martin, F., & Seehorn, D. (2019). Envisioning AI for K–12: What should every child know about AI? *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1), 9795–9799.

UNESCO. (2019). *Beijing consensus on artificial intelligence and education*. UNESCO.

UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. UNESCO.

UNESCO. (2022). *K–12 AI curricula: A mapping of government-endorsed AI curricula*. UNESCO.

Vygotsky, L. S., & Cole, M. (1978). *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158), 209–212.

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in Educational Theory and Practice* (pp. 277–304). Lawrence Erlbaum.

Yang, Z., Feng, Z., Huo, R., Lin, H., Zheng, H., Nie, R., & Chen, H. (2025). The imitation game revisited: A comprehensive survey on recent advances in AI-generated text detection. *Expert Systems with Applications*, 272, 126694.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., et al. (2025). Siren's song in the AI ocean: A survey on hallucination in large language models. *Computational Linguistics*, 51(4), 1373–1418.

Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., et al. (2026). Retrieval-augmented generation for AI-generated content: A survey. *Data Science and Engineering*, 1–29.