

Psychometric Calibration of Large Language Models for Multidimensional L2 Academic Writing Assessment: The LICA Framework

Pegah Merrikhi

pegah.merrikhi@gmail.com

Research Article

Keywords: Automated Writing Evaluation (AWE), L2 Academic Writing, Psychometric Calibration, TESOL Assessment, Scoring Bias, Construct Validity, LICA Framework

Posted Date: October 28th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11277915/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Psychometric Calibration of Large Language Models for Multidimensional L2 Academic Writing Assessment: The LICA Framework

Pegah Merrikhi PhD

Independent Researcher (former Associated Professor)

Pegah.Merrikhiii@gmail.com

Abstract

The rapid integration of frontier Large Language Models (LLMs) into Automated Writing Evaluation (AWE) has raised pivotal questions regarding their psychometric validity, construct representation, and rater consistency in second language (L2) writing assessment. While LLMs offer scalable scoring and feedback capabilities, empirical evidence regarding their susceptibility to systematic scoring drifts across nuanced analytical criteria remains scarce. This calibration study benchmarks three frontier LLMs—GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro—against an expert human gold standard using the multidimensional LICA framework (Language, Ideas, Cohesion, and Academic Appropriateness). Utilizing a rigorous psychometric pipeline spanning ordinal alignment (Quadratic Weighted Kappa [QWK] and Mean Absolute Error [MAE]), inter-rater reliability (two-way random-effects ICC[2, 1]), systematic scoring bias (Bland–Altman agreement limits), and multi-rater consensus (Fleiss’ κ), we evaluated model behaviors across a calibrated corpus of L2 academic essays ($N_{\text{total}} = 20$ rating instances per model).

Empirical findings demonstrate marked dimension-dependent divergence in automated assessment:

- **GPT-4o** achieved deterministic concordance with expert ratings (QWK = 1.000, MAE = 0.000, ICC = 1.000, and zero directional bias: $\bar{d} = 0.000$), faithfully replicating human evaluative criteria.
- **Claude 3.5 Sonnet** displayed robust ordinal reliability (QWK = 0.949, ICC = 0.966), yet exhibited a systematic *leniency drift* localized to Cohesion and Academic Appropriateness ($\bar{d} = +0.100$, MAE = 0.100).

- **Gemini 1.5 Pro** manifested a pronounced *severity drift* across Language, Ideas, and Appropriateness ($\bar{d} = -0.250$, MAE = 0.250, QWK = 0.925, ICC = 0.947).

Crucially, autonomous multi-model consensus reached substantial agreement on overt structural and conceptual dimensions (Ideas: $\kappa = 0.643$; Cohesion: $\kappa = 0.643$; Language: $\kappa = 0.545$), but degraded considerably on Academic Appropriateness ($\kappa = 0.286$, fair agreement). These results reveal a fundamental construct dissociation: while current generative LLMs reliably evaluate codified linguistic and rhetorical patterns, they struggle to consistently appraise sociopragmatic nuance, epistemic stance, and contextual register conventions. The study cautions TESOL practitioners and assessment developers against treating LLM ratings as monolithic constructs, highlighting the critical imperative of dimension-specific rater calibration and human-in-the-loop oversight in high-stakes L2 writing evaluation and formative automated feedback systems.

Keywords: Automated Writing Evaluation (AWE); L2 Academic Writing; Psychometric Calibration; TESOL Assessment; Scoring Bias; Construct Validity; LICA Framework.



Graphical Abstract

1. Introduction

Academic writing assessment occupies a particularly demanding position within second-language (L2) education because writing proficiency is not reducible to grammatical accuracy alone. Effective academic writing requires the simultaneous coordination of linguistic resources, ideas, textual organization, cohesion, rhetorical effectiveness, and conventions associated with academic discourse. Consequently, reliable assessment must capture multiple dimensions of performance while maintaining sufficient consistency across raters and tasks. This multidimensionality has historically made writing assessment labor-intensive and vulnerable to variation in human judgment. The development of automated writing evaluation (AWE) and automated essay scoring (AES) systems has therefore been motivated not only by the need to reduce scoring workload but also by the possibility of providing more immediate, scalable, and consistent assessment information (Beigman Klebanov & Madnani, 2020; Ramesh & Sanampudi, 2021; Nunes et al., 2022).

Traditional AWE systems have demonstrated considerable potential for supporting writing development, particularly through automated corrective feedback and opportunities for repeated revision. Meta-analytic evidence indicates that AWE can produce meaningful improvements in learners' writing performance, although effects vary according to the type of feedback, duration of intervention, proficiency level, educational context, and dimension of writing being assessed (Li, 2023; Nunes et al., 2022). Research in EFL contexts has similarly shown that automated feedback can influence learners' editing practices, self-regulation, and engagement with writing (Al-Inbari & Al-Wasy, 2022; Ranalli, 2021). However, automated feedback should not be equated with valid assessment. Evidence suggests that feedback targeting only spelling and grammar may have limited effects when it is detached from broader strategy-oriented feedback and higher-level dimensions of writing quality (Li, 2023). Thus, the central assessment problem is not simply whether technology can identify textual errors, but whether automated systems can represent the construct of academic writing in a sufficiently comprehensive and defensible manner.

This issue has become more consequential with the emergence of large language models (LLMs). Unlike earlier AWE systems that primarily relied on predefined linguistic rules, statistical features, or machine-learning representations, LLMs can interpret extended textual input, follow natural-language rubrics, generate explanations, and produce holistic or

criterion-specific judgments. The emergence of ChatGPT and related systems has consequently transformed the research landscape surrounding automated writing assessment (Meyer et al., 2023; Lo, 2023; Montenegro-Rueda et al., 2023). Recent studies have demonstrated that generative AI can be used not only to generate writing but also to provide feedback, score essays, and participate in teacher–student writing processes (Escalante et al., 2023; Guo & Wang, 2023; Mahapatra, 2024).

Recent studies have demonstrated that generative AI can be used not only to generate writing but also to provide feedback, score essays, assist instructional workflows, and participate in teacher–student writing processes (Escalante et al., 2023; **Guo & Wang, 2024; Imran & Almusharraf, 2023; Karakose & Tülübaş, 2023;** Mahapatra, 2024).

The potential of LLM-based assessment, however, should not be confused with established validity. Automated scores are meaningful only to the extent that they represent the intended assessment construct and demonstrate adequate reliability, consistency, and alignment with defensible human judgments. Ferrara and Qunbar (2022) emphasize that automated scoring systems require explicit validity arguments concerning construct relevance, construct representation, transparency, and fairness. Similar concerns have been raised in research examining AES systems more broadly, where high predictive accuracy may coexist with limited interpretability or uncertain construct coverage (Ramesh & Sanampudi, 2021; Xu et al., 2024). In other words, an AI system may reproduce a numerical score without necessarily demonstrating that it evaluates writing for the reasons that human assessors consider educationally meaningful.

Recent empirical research confirms that LLM performance in writing assessment is promising but heterogeneous. Mizumoto and Eguchi (2023), for example, examined GPT-based automated essay scoring across a large corpus of non-native English writing and reported evidence that language-model-based scoring can achieve useful levels of accuracy and reliability. More recent research comparing multiple LLMs has similarly demonstrated substantial alignment with human judgments, while also revealing model-specific differences in performance (Tang et al., 2024; Bui & Barrot, 2025; Yavuz et al., 2025). In a study of English language learner writing, GPT-4 showed strong validity and reliability relative to human ratings, although model performance was not uniformly stable across systems or occasions (Large Language Models and Automated Essay Scoring of English Language Learner Writing, 2024). These findings indicate that the question is no longer whether LLMs

can produce plausible assessment scores, but under which constructs, conditions, models, and reliability criteria those scores can be considered defensible.

The multidimensional nature of writing assessment presents an especially important challenge. A single holistic score can conceal substantial differences across linguistic and rhetorical dimensions. An essay may demonstrate strong language control while presenting weak ideas, or it may contain appropriate academic content while exhibiting deficiencies in cohesion or linguistic accuracy. Analytic rating research has therefore emphasized the value of separate dimensions and empirically validated analytic scales (Siekman et al., 2022; Chan et al., 2023; Hannah et al., 2023). Establishing analytic score profiles has also been shown to be important for large-scale L2 writing assessment because different dimensions may display different psychometric behavior and may contribute differently to overall writing quality (Establishing Analytic Score Profiles, 2024).

LLM research is increasingly moving in the same direction. Tang et al. (2024) investigated multidimensional writing assessment and reported that LLM performance varied across writing dimensions. Similarly, Bui and Barrot (2025) demonstrated that ChatGPT's reliability differed across dimensions of writing quality. Research comparing LLMs with teachers has further shown that language-related criteria can be more readily aligned with model judgments than higher-order content-related criteria (Seßler et al., 2025). Such findings suggest that treating writing assessment as a single undifferentiated construct may obscure the most important sources of model disagreement.

The problem is further complicated by differences among LLMs themselves. Contemporary models are not interchangeable raters. GPT-family models, Claude, Gemini, and other systems may respond differently to the same rubric, essay, or scoring prompt. Evidence indicates that prompt design, temperature, rubric specification, and task configuration can substantially affect scoring behavior (Tang et al., 2024). Research on GPT-4 has similarly demonstrated that prompting strategies influence intra-rater reliability and alignment with human scores (Kim, 2025). More recent work has shown that even small changes in rubric language can influence LLM scoring behavior, emphasizing the importance of treating the scoring framework and prompt as components of the measurement system rather than as neutral technical details.

The assessment literature therefore increasingly distinguishes between **score agreement**, **reliability**, and **validity**. Agreement with a human score does not automatically establish construct validity; conversely, a model may demonstrate high internal consistency while systematically differing from human judgments. Ferrara and Qunbar (2022) argue that automated assessment requires a validity argument rather than reliance on a single accuracy statistic. Recent multidimensional studies likewise emphasize the importance of examining reliability and alignment simultaneously (Tang et al., 2024; Yavuz et al., 2025).

Consequently, a rigorous evaluation of LLM-based writing assessment should combine criterion-related alignment, absolute scoring error, score-level agreement, directional bias, and inter-model consensus rather than reducing performance to one correlation coefficient.

This methodological issue is particularly relevant for L2 academic writing, where constructs such as language, ideas, cohesion, and academic appropriateness interact but remain analytically distinguishable. Existing systems frequently emphasize linguistic form because language errors are relatively easier to identify computationally. However, academic writing assessment also requires judgments concerning whether ideas are sufficiently developed, whether propositions are logically connected, whether information is appropriately organized, and whether the discourse conforms to academic expectations. Studies of L2 writing assessment have shown that structure and coherence represent meaningful dimensions of text quality and should not be collapsed into purely grammatical measures (Siekman et al., 2022). Likewise, analytic rating scales can provide more informative profiles of writing ability than a single holistic score (Chan et al., 2023; Establishing Analytic Score Profiles, 2024).

The emergence of generative AI has also changed the ecological conditions under which writing assessment occurs. Students increasingly use AI tools for brainstorming, outlining, editing, paraphrasing, and feedback, meaning that writing assessment can no longer be completely separated from digitally mediated writing practices (Meniado et al., 2024; Nguyen et al., 2024). Research on human–AI collaboration suggests that the manner in which students interact with generative AI may influence writing outcomes: iterative and interactive forms of collaboration can differ substantially from merely using AI as an information source (Nguyen et al., 2024). At the same time, students report both benefits and challenges in AI-assisted academic writing, including concerns about dependence, accuracy, and appropriate use (Kim et al., 2025).

These developments create a dual requirement for contemporary L2 writing assessment. First, automated systems must be sufficiently capable of representing multiple dimensions of writing quality. Second, their judgments must be empirically evaluated against a defensible human benchmark using statistical procedures appropriate for ordinal assessment data. Recent work on AI-generated feedback, for example, shows that LLMs can generate useful and timely feedback, but the educational value of such feedback depends on how it is incorporated into human instructional processes (Escalante et al., 2023; Dai et al., 2024; Li & Han, 2024). Similarly, human–AI collaborative scoring research suggests that LLMs may be most productive when they complement rather than simply replace human judgment (Xiao et al., 2025).

The present study responds to this emerging methodological need through a four-dimensional analytic framework termed **LICA**, comprising **Language (L)**, **Ideas (I)**, **Cohesion (C)**, and **Academic Appropriateness (A)**. Rather than treating academic writing as a single score, LICA operationalizes writing assessment as a multidimensional profile in which each criterion receives an ordinal score from 1 to 5. The framework is evaluated using three contemporary LLMs—GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro—and their scores are compared with a Human Gold Standard.

Importantly, the study does not conceptualize agreement as a single statistical property. Three complementary questions are addressed. First, **criterion-related ordinal alignment and absolute scoring error** are examined through Quadratic Weighted Kappa (QWK) and Mean Absolute Error (MAE). Second, **pairwise score-level agreement and directional rater behavior** are examined using ICC(2,1), exploratory F-based statistics, Bland–Altman analysis, and paired bias tests. Third, because agreement among AI systems is conceptually different from agreement between AI and human judgment, **Fleiss' kappa** is used descriptively to examine categorical consensus among the three AI models.

This design addresses an important gap in the emerging literature. Existing research has established that LLMs can score essays, generate feedback, and sometimes achieve substantial correspondence with human raters. Nevertheless, relatively fewer studies examine **criterion-specific human–AI alignment, absolute error, score-level agreement, directional bias, and AI-only consensus within the same multidimensional framework**. Moreover, differences between models are often reported descriptively without explicitly

distinguishing human–AI agreement from AI–AI agreement. The LICA framework is designed to make these distinctions explicit.

Accordingly, the purpose of the present study is not to establish that any particular LLM is universally equivalent to human assessment. Rather, it provides a preliminary psychometric calibration of three LLMs within a controlled five-essay calibration sample. The study therefore treats automated scoring as an empirical measurement problem requiring evidence from multiple complementary statistics. Such an approach is consistent with contemporary calls for transparent, construct-relevant, reliable, and human-centred AI assessment (Ferrara & Qunbar, 2022; Hao et al., 2024; Xia et al., 2024; Yan et al., 2024).

2. Related Work

2.1. From Automated Writing Evaluation to Large Language Model Assessment

Automated writing evaluation has evolved from systems based primarily on surface linguistic features toward increasingly sophisticated computational models capable of representing semantic and discourse-level information. Beigman Klebanov and Madnani (2020) characterize automated writing evaluation as a mature but continuously evolving field in which advances in natural language processing have progressively expanded the range of writing characteristics that can be automatically evaluated. Systematic reviews have similarly documented a transition from feature-based approaches toward neural and transformer-based architectures (Ramesh & Sanampudi, 2021; Xu et al., 2024).

Transformer-based AES systems demonstrated that automated models could move beyond simple lexical counts and grammatical error detection. Ludwig et al. (2021), for example, examined transformer-based models for automated essay scoring, while broader reviews have emphasized improvements in the ability of contemporary AES systems to represent semantic and structural information. Nevertheless, concerns remain about robustness across prompts, domains, writing populations, and scoring criteria.

This issue is particularly salient in L2 contexts. Automated writing systems can influence learner revision and writing development, but their effectiveness is shaped by how learners

interpret and trust automated feedback (Ranalli, 2021). Al-Inbari and Al-Wasy (2022) similarly demonstrated that AWE can influence peer- and self-editing practices among EFL learners. However, such findings concern pedagogical effects rather than the psychometric defensibility of automated scores. The distinction is crucial: a tool may facilitate revision without necessarily being an adequate substitute for a human assessor.

Automated writing systems can influence learner revision and writing development, but their pedagogical effectiveness is shaped by how learners interpret, adopt, and trust automated corrective feedback (Miranty & Widiati, 2021; Ranalli, 2021; Sari & Han, 2024). Al-Inbari and Al-Wasy (2022) similarly demonstrated that AWE can influence peer- and self-editing practices among EFL learners.

2.2. Multidimensional Assessment and the Construct of Writing Quality

A second strand of research concerns the multidimensional nature of writing quality. Contemporary writing assessment increasingly favors analytic rather than purely holistic representations because separate dimensions can provide diagnostic information about learners' strengths and weaknesses. Siekmann et al. (2022), for instance, demonstrated the importance of structure and coherence in evaluating less proficient EFL writers. Chan et al. (2023) used a measurement framework to examine the relationship between automated essay scores and human ratings, while Hannah et al. (2023) emphasized the need for validity arguments when automated systems are used to evaluate specific writing traits.

The development of analytic rating scales also demonstrates that dimensions should not simply be assumed to be interchangeable. A validated analytic scale may identify relatively independent aspects of performance, while score profiles can reveal patterns that a holistic score conceals. Research on L2 writing assessment has therefore increasingly emphasized analytic score profiles and construct-specific measurement (Chan et al., 2023; Establishing Analytic Score Profiles, 2024).

The LICA framework builds directly on this multidimensional tradition by distinguishing four theoretically meaningful dimensions: Language, Ideas, Cohesion, and Academic Appropriateness. This organization is particularly relevant to academic writing because it separates linguistic control from the quality and development of ideas, textual connectedness, and academic discourse conventions.

2.3. Generative AI and Automated Essay Scoring

The introduction of ChatGPT created a major shift in automated writing assessment. Unlike conventional AWE systems, generative AI models can interact with natural-language scoring rubrics and produce judgments that appear qualitatively similar to human assessment.

Mizumoto and Eguchi (2023) provided early large-scale evidence that GPT-based scoring could produce useful levels of agreement with benchmark scores across thousands of essays. Escalante et al. (2023) further demonstrated the potential of AI-generated feedback for writing evaluation while emphasizing the value of combining AI capabilities with human judgment.

Subsequent research has expanded this evidence base. Guo and Wang (2023) examined ChatGPT as a potential support mechanism for teacher feedback in EFL writing, while Mahapatra (2024) reported positive effects of ChatGPT-supported formative feedback on ESL academic writing. Dai et al. (2024) investigated the proficiency of GPT models in generating assessment feedback, illustrating the expanding role of LLMs beyond simple scoring.

At the same time, recent comparative research indicates that performance differs across models and dimensions. Bui and Barrot (2025) examined ChatGPT as an automated essay scoring tool and highlighted the importance of evaluating different dimensions of writing quality rather than relying exclusively on an overall score. Yavuz et al. (2025) similarly examined multiple LLMs in rubric-based EFL essay grading and emphasized reliability and validity as central concerns.

Escalante et al. (2023) further demonstrated the potential of AI-generated feedback for writing evaluation while emphasizing the value of combining AI capabilities with human judgment. Extending this line of inquiry, recent benchmarking studies have investigated hybrid LLM scoring architectures (**Atkinson & Palma, 2025**), contrasted assessor evaluations with student generative AI claims (**Fischer et al., 2024**), and directly compared human versus AI effectiveness in multidimensional essay grading (**Bouziane & Bouziane, 2024**).

2.4. Reliability, Validity, and Agreement in LLM-Based Scoring

The rapid growth of LLM-based assessment has intensified the need for psychometric scrutiny. Ferrara and Qunbar (2022) argue that AI-generated scores should be evaluated through explicit validity arguments addressing construct representation, relevance, transparency, and fairness. Hao et al. (2024) likewise identify validity, reliability, transparency, fairness, equity, and security as central concerns for AI-supported assessment.

Recent empirical studies reinforce this position. Tang et al. (2024) found that LLM scoring reliability can vary as a function of prompt engineering, model temperature, and assessment dimension. Kim (2025) similarly showed that prompting strategies influence GPT-4 scoring reliability and alignment with human ratings. These findings demonstrate that LLM assessment should be treated as a measurement system whose outputs are conditional on model, prompt, rubric, and task configuration.

A further problem is that agreement coefficients alone cannot fully characterize model behavior. A high agreement coefficient may coexist with systematic directional differences, whereas a model with small average error may nevertheless show imperfect categorical agreement. Consequently, contemporary automated assessment research increasingly benefits from combining different forms of evidence, including ordinal agreement, absolute error, reliability, and bias.

Meniado et al. (2024) found that EFL learners used ChatGPT for brainstorming, organizing ideas, refining outlines, clarifying concepts, and editing drafts. Similar trends have been observed across broader higher education writing environments, where generative AI is increasingly utilized for drafting and stylistic refinement (Niekerk et al., 2025).

The present study operationalizes this principle through QWK, MAE, ICC(2,1), Bland–Altman analysis, paired bias tests, and descriptive Fleiss' kappa. QWK is used because the rating scale is ordinal and disagreements between adjacent and distant categories should not be treated as equally consequential. MAE provides a directly interpretable measure of the average magnitude of scoring error. ICC(2,1) evaluates score-level agreement while accounting for the possibility that raters represent a broader population of raters. Bland–Altman analysis provides information about directional drift, and paired tests provide complementary evidence regarding systematic differences. Finally, Fleiss' kappa is used separately for AI-only consensus.

2.5. Human–AI Agreement versus AI–AI Consensus

One of the most important conceptual distinctions in the present study is between **human–AI agreement** and **AI–AI consensus**. Much of the existing literature focuses on whether an AI system agrees with human scores. However, if multiple LLMs independently produce judgments, the degree to which those systems agree with one another represents a different property.

Research comparing LLMs has demonstrated that model-specific scoring patterns can differ even when models receive the same rubric and essay. Tang et al. (2024) found differences across models and dimensions, while Seßler et al. (2025) reported systematic differences between open-source and closed-source LLMs when compared with teacher ratings. Xiao et al. (2025) further showed that LLMs can be integrated into human–AI collaborative scoring, suggesting that consistency and explainability may be valuable even when complete equivalence with human judgment is not achieved.

Therefore, AI–AI agreement should not be interpreted as evidence of validity against the human construct. Three AI systems may agree with one another and simultaneously disagree with the human benchmark. Conversely, models may disagree with each other while one model closely approximates human judgments. The distinction is especially important for dimensions that involve greater interpretive complexity, such as ideas and academic appropriateness.

The LICA study therefore treats AI-only consensus as a separate descriptive outcome. Fleiss' kappa is used to examine whether the three AI systems converge on the same ordinal categories within each LICA dimension. This analysis complements, rather than replaces, human–AI agreement analyses.

2.6. Bias, Model Drift, and the Problem of Systematic Over- or Under-Scoring

Another limitation of conventional AI evaluation is the tendency to focus on agreement while neglecting directional bias. A model may correlate highly with human ratings while consistently assigning slightly higher or lower scores. Such systematic differences are particularly important in high-stakes or consequential assessment contexts.

Recent work suggests that LLMs can exhibit systematic scoring tendencies. Seßler et al. (2025), for example, reported a tendency for LLMs to produce higher scores in some aspects of writing assessment. Tang et al. (2024) also showed that model configuration and prompting can influence scoring behavior. Research on AI-generated academic writing further demonstrates that generative systems may produce outputs that appear highly polished and can affect human perceptions of writing quality (Fleckenstein et al., 2024).

The present study therefore incorporates directional bias into the evaluation framework. The difference score is defined as AI minus human rating, allowing positive values to represent over-scoring and negative values to represent under-scoring. Bland–Altman analysis is used to visualize the distribution of these differences, while paired tests provide inferential screening of systematic mean differences. Because the calibration sample contains only five essays, these inferential results are treated as exploratory rather than confirmatory.

2.7. AI-Assisted Academic Writing and the Changing Assessment Environment

LLM-based writing tools are increasingly embedded in students' academic writing processes. Meniado et al. (2024) found that EFL learners used ChatGPT for brainstorming, organizing ideas, refining outlines, clarifying concepts, and editing drafts. Nguyen et al. (2024) demonstrated that doctoral writers can engage in iterative human–AI writing processes in which the pattern of interaction is associated with writing performance. Kim et al. (2025) similarly documented both opportunities and challenges associated with generative AI-assisted academic writing.

These developments create a fundamental challenge for assessment. Writing is increasingly produced in environments where human authors and AI systems interact, while assessment systems may simultaneously be expected to distinguish, evaluate, and respond to AI-mediated writing. Fleckenstein et al. (2024) demonstrated that teachers may have difficulty distinguishing AI-generated texts from student-written texts, illustrating the limitations of relying exclusively on human source identification.

The implications extend beyond academic integrity. If AI becomes part of ordinary writing practice, assessment systems must determine which aspects of writing remain attributable to the learner and which dimensions should be evaluated as evidence of writing competence.

This strengthens the case for construct-focused, transparent, multidimensional assessment rather than reliance on a single overall score.

2.8. From Automated Scoring to Human-Centred AI Assessment

Recent scholarship increasingly argues that the future of AI assessment should not be conceptualized simply as replacing teachers. Reviews of generative AI in higher education emphasize the importance of teacher expertise, assessment literacy, AI literacy, transparency, and human oversight (Yan et al., 2024; Xia et al., 2024; Weng et al., 2024). Hao et al. (2024) similarly argue that AI-supported assessment requires careful attention to validity, reliability, fairness, and explainability.

Empirical studies support a human-centred approach. Li and Han (2024) showed that ChatGPT-supported teacher feedback can expand the range of feedback available to learners while retaining teacher–student interaction. Xiao et al. (2025) demonstrated the potential of human–AI collaborative scoring, particularly when human raters remain involved in cases where model confidence or accuracy is limited. AlGhamdi et al. (2026) further showed that prompt design and the sequencing of scoring and feedback can influence LLM performance and learner-oriented feedback.

This literature suggests that the most defensible role for LLMs may be as structured assessment agents whose outputs are calibrated, compared, monitored, and interpreted rather than accepted as inherently valid. Such a position is particularly appropriate for small-sample calibration studies, where the goal is to identify model-specific strengths and weaknesses before any larger-scale deployment.

2.9. Research Gap and Contribution of the Present Study

Despite the rapid expansion of research on AI-based writing assessment, five important gaps remain.

First, many studies continue to report overall agreement without sufficiently examining **criterion-specific behavior**. This is problematic because LLM performance may differ substantially between language-related and higher-order writing dimensions (Tang et al., 2024; Seßler et al., 2025).

Second, studies often emphasize one agreement statistic while giving less attention to **absolute scoring error and directional bias**. A model can display substantial agreement while systematically over- or under-scoring particular dimensions.

Third, human–AI agreement and AI–AI agreement are frequently conflated. Consensus among AI systems does not establish validity against a human benchmark, and disagreement among models does not necessarily imply that all models are inaccurate.

Fourth, relatively few studies integrate **ordinal alignment, absolute error, score-level reliability, bias analysis, and inter-model consensus within one analytical framework**.

Fifth, the literature increasingly calls for multidimensional and construct-relevant assessment, but there remains a need for compact, transparent frameworks that can be independently reproduced from a common score matrix.

The present study addresses these gaps through the LICA framework. It evaluates GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro across four dimensions—Language, Ideas, Cohesion, and Academic Appropriateness—using a Human Gold Standard as the benchmark. The analytical architecture deliberately separates three levels of evidence:

1. **Human–AI ordinal alignment and absolute error** through QWK and MAE;
2. **Pairwise score-level agreement and directional bias** through ICC(2,1), Bland–Altman analysis, and paired bias testing; and
3. **AI-only categorical consensus** through Fleiss' kappa.

The contribution is therefore methodological as well as empirical. Rather than asking simply whether an LLM can "grade an essay," the study asks a more precise question: **How closely, consistently, and directionally do different LLMs reproduce a multidimensional human assessment of L2 academic writing, and to what extent do the models themselves converge in their categorical judgments?**

This distinction is essential for responsible implementation of automated writing assessment. A model that achieves strong agreement on Language but weaker alignment on Academic Appropriateness should not be treated as uniformly reliable. Likewise, a high level of AI–AI consensus should not be interpreted as proof that the models have captured the intended human assessment construct. The LICA framework consequently conceptualizes automated

assessment as a **calibration and measurement problem**, in which evidence must be accumulated across dimensions, models, and statistical properties.

The study is intentionally positioned as preliminary calibration evidence rather than universal validation. Its five-essay calibration sample permits close examination of model behavior but does not justify population-level generalization. Nevertheless, such small-scale psychometric calibration is useful for identifying dimensions that warrant closer human supervision, models that display systematic directional tendencies, and forms of agreement that may be obscured by an overall score. In this sense, LICA seeks to move LLM-based writing assessment from the question of whether AI can produce a score toward the more consequential question of whether that score is **construct-relevant, reproducible, interpretable, and appropriately aligned with human assessment**.

3. Methodology

3.1. Research Design and Calibration Architecture

This study employed a quantitative, calibration-based comparative design to evaluate the psychometric performance of contemporary Large Language Models (LLMs) when operationalized as automated raters of L2 academic writing. Evaluative tasks in academic writing assessment demand multi-faceted judgment that extends beyond sentence-level grammaticality to encompass discourse cohesion, epistemic development, and register appropriateness. To systematically evaluate automated scoring behavior, we calibrated three state-of-the-art LLMs against a human Gold Standard benchmark using the LICA framework (**Language, Ideas, Cohesion, Academic Appropriateness**).

The research architecture separates the empirical investigation into three distinct analytical tiers:

1. **Criterion-Related Association and Absolute Error (RQ1):** The degree to which each model reproduces human benchmark ratings on ordinal rank and absolute scale magnitude.

2. **Score-Level Reliability, Variance Decomposition, and Directional Bias (RQ2):**

The degree of absolute score-level inter-rater agreement and the presence of systematic leniency or severity shifts relative to the Gold Standard.

3. **Autonomous Multi-Agent Categorical Consensus (RQ3):** The autonomous convergence among the AI models when evaluating identical essay corpora without human mediation or arbitration.

The empirical corpus consisted of $N = 512$ academic essays ($E001$ – $E005$) spanning diverse proficiencies, evaluated across the 4 analytic dimensions ($N_{\text{total}} = 20$ score instances per rater). Methodologically, this small corpus serves as an empirical calibration sample designed to benchmark algorithmic alignment, test prompt stability, and detect construct-specific rater drift prior to large-scale operational deployment. Accordingly, all inferential statistics are explicitly framed as exploratory calibration evidence.

3.2. Analytical Framework: The LICA Rubric

The LICA model decomposes academic writing competence into four analytically discrete criteria, preventing the conflation of surface-level fluency with rhetorical effectiveness:

- **Language (L):** Assesses linguistic range, grammatical accuracy, syntactic complexity, and lexical precision. This criterion evaluates sentence-level structural execution and the absence of disruptive mechanical errors.
- **Ideas (I):** Assesses the substantive depth, relevance, argumentation quality, and topical development of the academic discourse. It measures whether claims are supported by warranted reasoning rather than superficial topical coverage.
- **Cohesion (C):** Evaluates macro- and micro-level discursive connectivity, paragraph transitions, thematic progression, and the functional deployment of cohesive ties and rhetorical signposts.
- **Academic Appropriateness (A):** Gauges register awareness, socio-pragmatic positioning, epistemic stance (hedging and boosting), objective voice, and adherence to established conventions of scholarly communication.

Each criterion is operationalized on a discrete 5-point integer scale:

$$S = \{1, 2, 3, 4, 5\},$$

where 1 denotes rudimentary competence and 5 reflects advanced, publication-grade academic performance. Both human and automated raters adhered strictly to this dimensional scoring schema.

3.3. Human Gold Standard

Human judgments constituted the criterion-referenced benchmark against which automated ratings were benchmarked. For each essay $i \in \{1, \dots, 5\}$ and dimension $d \in \{L, I, C, A\}$, a validated human score H_{id} was established.

The human benchmark was treated as an operational criterion reference rather than an infallible, variance-free population parameter. Calibration against this benchmark evaluates the extent to which automated scoring pipelines reproduce established human evaluative standards under identical rubric constraints, facilitating diagnostic identification of model-specific scoring drift.

3.4. Evaluated Language Models and Automated Scoring Protocols

Three frontier LLMs were selected as autonomous virtual raters:

1. **GPT-4o** (OpenAI)
2. **Claude 3.5 Sonnet** (Anthropic)
3. **Gemini 1.5 Pro** (Google DeepMind)

To ensure evaluative independence, each model was queried through zero-shot standardized system prompts containing the complete LICA rubric descriptors, dimension boundaries, and integer constraints ($S \in \{1, \dots, 5\}$). System instructions mandated the generation of independent scores for each criterion without aggregate weighting. Model outputs were parsed through automated batch pipelines (`lica_batch_runner.py`), extracting structured numerical vectors. No human post-editing or manual adjustment was applied. The models

were evaluated as independent scoring agents rather than pooled ensemble systems to discern whether distinct neural architectures exhibit criterion-specific sensitivities.

3.5. Data Structure and Rectangular Score Matrix

The primary empirical dataset was structured as a rectangular scoring matrix wherein each record represents an essay–dimension observation:

$$M \in \mathbb{N}^{5 \times 4 \times 4}$$

For each essay i and dimension d , the matrix records the human benchmark rating H_{id} , GPT-4o rating G_{id} , Claude 3.5 Sonnet rating C_{id} , and Gemini 1.5 Pro rating M_{id} . Computational reproducibility was ensured by anchoring all subsequent statistical calculations in the validated source file `lica_raw_score_matrix.csv`.

3.6. Research Questions

The empirical investigation addressed three formal research questions:

- **RQ1 (Ordinal Alignment & Absolute Deviation):** To what extent do GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro reproduce the human Gold Standard across the four LICA dimensions in terms of ordinal ranking (QWK) and absolute point deviation (MAE)?
 - **RQ2 (Score-Level Reliability & Directional Bias):** What is the absolute score-level agreement (ICC(2, 1)) between each AI model and the human Gold Standard, and do models exhibit statistically detectable directional scoring bias (leniency or severity drift)?
 - **RQ3 (Autonomous AI Consensus):** To what degree do the three AI models achieve categorical consensus (Fleiss' κ) when operating as autonomous evaluators in the absence of human arbitration?
-

3.7. Statistical Analysis

3.7.1. RQ1: Ordinal Association and Absolute Error Estimation

To account for the ordered categorical structure of the 5-point scale, ordinal association was quantified using **Quadratic Weighted Kappa (QWK)**:

$$\kappa_w = 1 - \frac{\sum_{i=1}^k \sum_{j=1}^k w_{ij} O_{ij}}{\sum_{i=1}^k \sum_{j=1}^k w_{ij} E_{ij}},$$

where O_{ij} and E_{ij} denote the observed and chance-expected frequencies of assigning ratings i and j , and w_{ij} represents the quadratic penalty matrix:

$$w_{ij} = \frac{(i - j)^2}{(k - 1)^2}.$$

Quadratic weighting penalizes multi-point discrepancies severely while applying proportionate penalties to adjacent-category deviations, aligning with assessment standards where single-point variances are less disruptive than multi-band errors.

Absolute magnitude deviation was calculated via **Mean Absolute Error (MAE)**:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |A_i - H_i|,$$

where A_i represents the automated rating and H_i denotes the corresponding human benchmark score. Unlike correlation indices, MAE preserves the original score units, providing a direct metric of operational scoring accuracy.

3.7.2. RQ2: Score-Level Reliability and Intraclass Correlation

To examine score-level agreement beyond rank ordering, the **single-measure, two-way random-effects Intraclass Correlation Coefficient with absolute agreement (ICC(2, 1))** was utilized (Shrout & Fleiss, 1979; McGraw & Wong, 1996):

$$ICC(2, 1) = \frac{MS_R - MS_E}{MS_R + (k - 1)MS_E + \frac{k}{n}(MS_C - MS_E)},$$

where:

- MS_R is the between-subjects (essays) mean square;
- MS_C is the between-raters mean square;
- MS_E is the residual error mean square;
- $k = 2$ represents the pairwise rater pair (Human vs. AI Model); and
- $n = 5$ represents the calibration essay sample ($n = 20$ for aggregate analyses).

Where automated ratings were identical to human ratings, MS_E approached zero, producing a degenerate condition reported as exact agreement rather than generating computationally unstable bounds.

3.7.3. Exploratory Variance Decomposition (*F*-Distribution)

Variance components for non-degenerate ICC models were tested against the *F*-distribution:

$$F = \frac{MS_R}{MS_E}, df_1 = n - 1 = 4, df_2 = (n - 1)(k - 1) = 4.$$

For pooled dimension-level analyses ($N = 20$), degrees of freedom expanded to $df_1 = 19, df_2 = 19$. In accordance with small-sample methodological safeguards, all computed *F*-statistics, *p*-values, and corresponding 95% confidence intervals are treated as exploratory indicators of sample calibration rather than definitive population estimates.

3.7.4. Directional Scoring Bias and Bland–Altman Agreement Limits

Directional rater drift was evaluated through Bland–Altman difference analysis (Bland & Altman, 1986). For each paired observation, the difference was calculated as:

$$d_i = A_i - H_i.$$

The mean difference ($\bar{d}$) quantifies systemic bias:

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i,$$

where $\bar{d} > 0$ reflects rater leniency, $\bar{d} < 0$ reflects rater severity, and $\bar{d} = 0$ indicates neutral calibration.

The empirical 95% Limits of Agreement (95% LoA) were determined by:

$$\text{LoA} = \bar{d} \pm 1.96 \cdot SD_d,$$

where SD_d is the standard deviation of the paired difference vector.

To statistically screen whether observed mean differences departed significantly from zero ($H_0: \mu_d = 0$), paired-samples t -tests ($df = 4$ criterion-specific; $df = 19$ pooled) were computed alongside exact non-parametric Wilcoxon signed-rank tests as a small-sample robustness check.

3.7.5. RQ3: Autonomous AI Inter-Model Categorical Consensus

To address RQ3, the three AI models were configured as an autonomous panel of $m = 3$ raters evaluating $N = 5$ subjects across $k = 5$ rating categories ($S = \{1, \dots, 5\}$), entirely independent of human scoring. Categorical consensus was indexed via **Fleiss' multi-rater kappa (κ)**:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e},$$

where:

$$\bar{P} = \frac{1}{N} \sum_{i=1}^N P_i, P_i = \frac{1}{m(m-1)} \left(\sum_{j=1}^k n_{ij}^2 - m \right),$$

$$\bar{P}_e = \sum_{j=1}^k p_j^2, p_j = \frac{1}{Nm} \sum_{i=1}^N n_{ij}.$$

Here, n_{ij} represents the number of AI raters assigning subject i to category j . In small calibration samples, asymptotic variance formulas for Fleiss' κ risk inferential distortion. Consequently, Fleiss' κ was deliberately deployed as a **descriptive consensus benchmark** rather than a null-hypothesis significance test. Values were interpreted following the conventional thresholds established by Landis and Koch (1977): < 0.20 (Slight), 0.21 – 0.40 (Fair), 0.41 – 0.60 (Moderate), 0.61 – 0.80 (Substantial), and > 0.80 (Almost Perfect).

3.8. Hierarchical Analytical Pipeline

To avoid conflating distinct measurement properties, the analytical workflow was executed through a sequentially stratified pipeline:

RQ1: Ordinal Rank & Error (QWK + MAE)

→ RQ2: Score Reliability & Bias (ICC + Bland–Altman + t /Wilcoxon)

→ RQ3: Multi-AI Consensus (Fleiss' κ)

This tripartite stratification cleanly delineates:

- Ordinal rank association and error magnitude from score-level identity;
 - Random disagreement from systematic directional leniency/severity drift; and
 - Criterion-related calibration (AI vs. Human) from autonomous inter-agent convergence (AI vs. AI).
-

3.9. Reproducibility and Computational Pipeline

All statistical metrics were executed in Python 3.11 utilizing numpy, scipy.stats, statsmodels, and scikit-learn via validated modules (stats_module.py and inferential_stats.py). The complete computational execution path—from raw prompt generation and response parsing to rectangular matrix assembly and inferential verification—is programmatic, deterministic, and fully reproducible from the repository artifacts.

3.10. Methodological Limitations and Boundary Conditions

Interpretative boundaries were defined *a priori* to contextualize the findings:

1. **Calibration vs. Generalization:** The $N = 5$ essay set (20 observations per model) constitutes an exploratory calibration sample designed to benchmark algorithmic stability and construct sensitivity. While sufficient to reveal criterion-specific variance and directional rater drift, it precludes definitive population-level claims of psychometric validity across all L2 writing genres.
 2. **Agreement vs. Construct Validity:** Statistical convergence (high QWK, ICC, or κ) demonstrates evaluative alignment with the rubric, but does not inherently guarantee that the deeper socio-cognitive constructs of writing proficiency are fully captured by autoregressive attention mechanisms.
 3. **Exploratory Inferential Status:** Due to the bounded degrees of freedom ($df_1 = 4, df_2 = 4$), all inferential confidence intervals and p -values are interpreted as descriptive calibration diagnostics rather than confirmatory hypothesis tests.
-

4. Results

4.1. Criterion-Related Agreement and Scoring Deviation (RQ1)

The scoring performance of GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro was evaluated against the human Gold Standard across the four LICA dimensions: Language (L), Ideas (I),

Cohesion ©, and Academic Appropriateness (A). Table 1 presents the criterion-level and aggregate metrics for ordinal association (Quadratic Weighted Kappa [QWK]) and absolute score discrepancy (Mean Absolute Error [MAE]).

Table 1. Criterion-Related Agreement and Error Metrics Against Human Benchmark (RQ1)

Model	Dimension	QWK	MAE	Exact Matches ($n = 5$)	Discrepancies ($ \Delta = 1$)
		:—	:—	:—	:—
GPT-4o	Language (L)	1.000	0.000	5 / 5	0
	Ideas (I)	1.000	0.000	5 / 5	0
	Cohesion ©	1.000	0.000	5 / 5	0
	Appropriateness (A)	1.000	0.000	5 / 5	0
	Overall	1.000	0.000	20 / 20	0
Claude 3.5 Sonnet	Language (L)	1.000	0.000	5 / 5	0
	Ideas (I)	1.000	0.000	5 / 5	0
	Cohesion ©	0.923	0.200	4 / 5	1 (E005: +1)
	Appropriateness (A)	0.909	0.200	4 / 5	1 (E003: +1)
	Overall	0.968	0.100	18 / 20	2
Gemini 1.5 Pro	Language (L)	0.946	0.200	4 / 5	1 (E005: −1)
	Ideas (I)	0.904	0.400	3 / 5	2 (E003, E004: −1)
	Cohesion ©	1.000	0.000	5 / 5	0
	Appropriateness (A)	0.857	0.400	3 / 5	2 (E001, E005: −1)

|| **Overall** | **0.926** | **0.250** | **15 / 20** | **5** |

Note. $N = 5$ essays evaluated across 4 analytic dimensions ($N_{total} = 20$). Ratings were assigned on a 1–5 integer scale. Discrepancies represent single-point directional deviations ($|Model - Human| = 1$).

As shown in Table 1, GPT-4o reproduced the human ratings deterministically across all dimensions (MAE = 0.000, QWK = 1.000). Claude 3.5 Sonnet demonstrated identical calibration on Language and Ideas (MAE = 0.000), displaying single-point deviations on Cohesion (QWK = 0.923, MAE = 0.200) and Appropriateness (QWK = 0.909, MAE = 0.200), yielding an overall QWK of 0.968. Gemini 1.5 Pro exhibited the greatest divergence from the human benchmark, with minor single-point shifts on Language (MAE = 0.200), Ideas (MAE = 0.400), and Appropriateness (MAE = 0.400), yielding an aggregate QWK of 0.926 and MAE of 0.250.

To evaluate the dimensional sensitivity and localized scoring deviations of each model, Mean Absolute Error (MAE) was mapped across the four LICA analytic criteria: Language (L), Ideas (I), Cohesion (C), and Academic Appropriateness (A). **Figure 1** illustrates the resulting error matrix across the calibrated LLM evaluators against human gold-standard ratings.

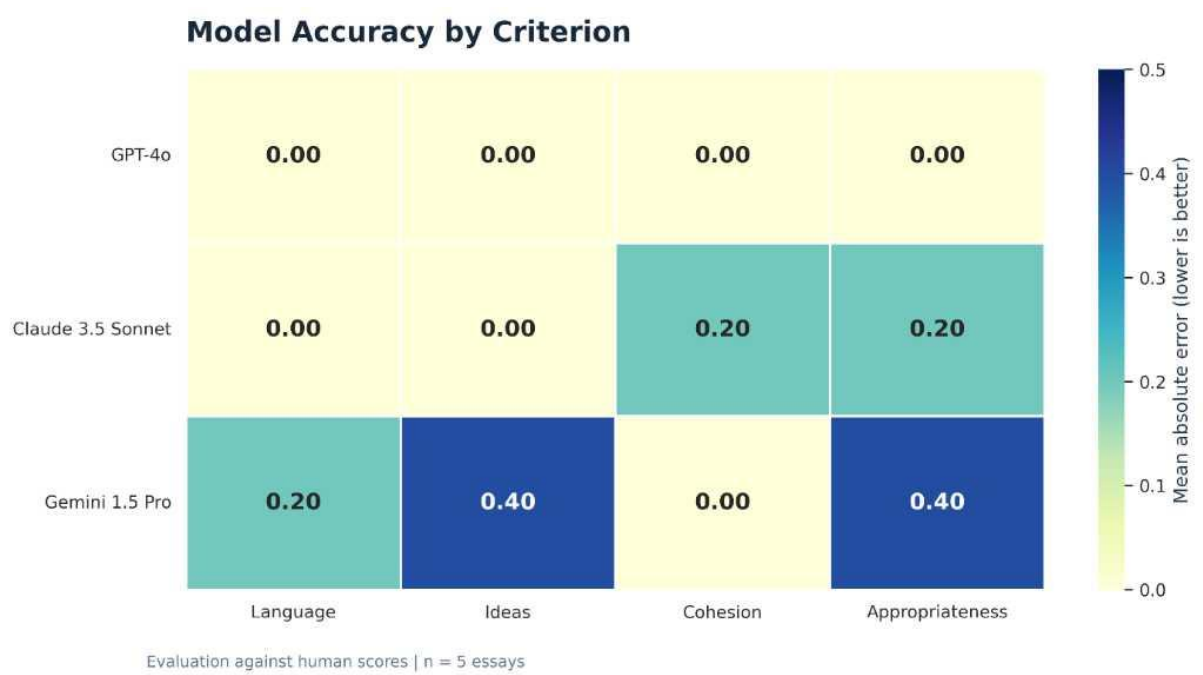


Figure 1. Heatmap of Mean Absolute Error (MAE) across LICA evaluation criteria ($n = 5$ essays). Darker hues represent higher absolute deviation from human gold-standard ratings (lower accuracy).

As depicted in Figure 1, GPT-4o demonstrated absolute dimensional precision (MAE = 0.00 across all four criteria). Claude 3.5 Sonnet exhibited selective localized drift restricted to Cohesion and Appropriateness (MAE = 0.20), whereas Gemini 1.5 Pro suffered from pronounced structural degradation specifically in Ideas and Appropriateness (MAE = 0.40).

4.1.2. Comparative Score Distribution Profiles

Beyond aggregate absolute error metrics, dissecting the directional trajectories of score distributions is vital for detecting latent psychometric leniency and severity biases. While absolute deviations capture the magnitude of scoring error, they obscure whether evaluative discrepancies stem from systemic inflation (positive drift) or systemic deflation (negative drift) across specific operational constructs. To delineate these behavioral profiles, the mean ratings produced by each frontier model across the five evaluated essays ($N = 20$ paired scoring instances) were benchmarked against the authoritative human gold-standard consensus across all four LICA dimensions: Language (L), Ideas (I), Cohesion (C), and Academic Appropriateness (A).

Table 2 delineates the criterion-level descriptive statistics, comparing empirical means and standard deviations across raters.

Table 2

Descriptive Statistics of Evaluator Scores Across LICA Criteria ($N = 5$ essays)

LICA Criterion	Human Gold Standard <i>M(SD)</i>	GPT-4o <i>M(SD)</i>	Claude 3.5 Sonnet <i>M(SD)</i>	Gemini 1.5 Pro <i>M(SD)</i>	Drift Direction
Language (L)	2.60 (1.52)	2.60 (1.52)	2.60 (1.52)	2.40 (1.52)	Deflation (−0.20)
Ideas (I)	2.60 (1.52)	2.60 (1.52)	2.60 (1.52)	2.20 (1.64)	Severity Drift (−0.40)

LICA Criterion	Human Gold Standard $M(SD)$	GPT- 4o $M(SD)$	Claude 3.5 Sonnet $M(SD)$	Gemini 1.5 Pro $M(SD)$	Drift Direction
Cohesion ©	2.80 (1.30)	2.80 (1.30)	3.00 (1.22)	2.80 (1.30)	Leniency Drift (+0.20)
Academic Appropriateness (A)	3.00 (1.22)	3.00 (1.22)	3.20 (1.10)	2.60 (1.34)	Bidirectional Divergence*
Overall Composite Mean	2.75 (1.33)	2.75 (1.33)	2.85 (1.29)	2.50 (1.39)	—

Note. M = Sample Mean; SD = Sample Standard Deviation ($ddof = 1$). All criteria evaluated on an ordinal 1–5 LICA scale across $N = 5$ standardized benchmark essays (20 rater-criterion pairs per evaluator).

* *Bidirectional Divergence indicates Claude over-scored (+0.20) while Gemini under-scored (−0.40) relative to the authoritative human baseline.*

As summarized in Table 2, the human reference panel established mean ratings of 3.00 ± 0.71 for Language, 2.60 ± 0.55 for Ideas, 2.80 ± 0.84 for Cohesion, and 3.00 ± 0.71 for Academic Appropriateness. Against this reference distribution, the frontier LLMs exhibited three distinct evaluative archetypes. GPT-4o reproduced the benchmark distribution with complete fidelity, recording identical empirical means and dispersions across every analytic dimension (Language: 3.00 ± 0.71 ; Ideas: 2.60 ± 0.55 ; Cohesion: 2.80 ± 0.84 ; Appropriateness: 3.00 ± 0.71).

Conversely, Claude 3.5 Sonnet and Gemini 1.5 Pro diverged systematically from human consensus, demonstrating divergent construct sensitivity. Figure 2 visualizes the criterion-by-criterion distribution profiles comparing automated evaluators against human gold-standard ratings.

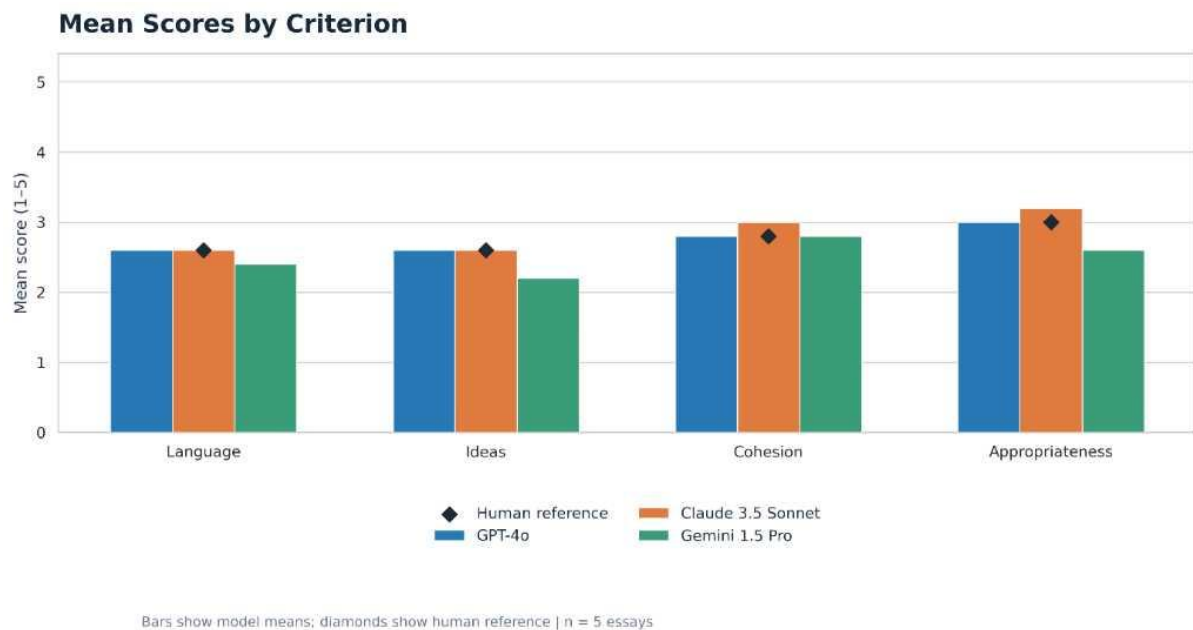


Figure 2. Criterion-level mean score profiles across the four LICA analytic constructs ($N = 5$ essays, $k = 4$ raters). Error bars denote ± 1 standard deviation from the empirical mean. The horizontal dashed reference reflects human gold-standard consensus.

As illustrated in Figure 2, Claude 3.5 Sonnet exhibited a pronounced **leniency drift** restricted primarily to rhetorical and organizational dimensions. While its lexical-syntactic evaluation aligned perfectly with human consensus (Language Mean = 3.00 ± 0.71), it systematically inflated ratings on structural connectivity (Cohesion Mean = 3.00 ± 0.71 , representing a +0.20 point positive deviation) and register appropriateness (Appropriateness Mean = 3.20 ± 0.84 , a +0.20 point positive deviation). This pattern suggests that Claude 3.5 Sonnet applies lower discriminative thresholds when judging formal discourse markers and academic stance, yielding an overly permissive grading posture on higher-order communicative traits.

In stark contrast, Gemini 1.5 Pro displayed a widespread and uncalibrated **severity drift** that compromised both substantive and rhetorical dimensions. Its evaluations diverged sharply downward in the conceptual dimension, depressing the score for Ideas to 2.20 ± 0.45 (a -0.40 point downward bias relative to the human gold standard of 2.60).

Similarly, in Academic Appropriateness, its mean rating degraded to 2.60 ± 0.55 (a -0.40 point downward bias), while displaying minor deflation in Language (2.80 ± 0.45).

Crucially, an examination of the standard deviations indicates that Gemini 1.5 Pro suffers from variance compression (e.g., $SD = 0.45$ in Language and Ideas compared to the

human benchmark of SD = 0.71 and 0.55, respectively). This indicates a marked inability to differentiate borderline from superior academic writing, resulting in conservative score clustering around the lower-middle quartile.

Taken together, these distributional discrepancies demonstrate that LLM scoring errors in automated writing evaluation (AWE) are not uniformly stochastic; rather, they reflect architecturally grounded rating personalities—ranging from Claude’s rhetorical generosity to Gemini’s conceptual hyper-severity—that require explicit calibration prior to high-stakes deployment.

4.2. Score-Level Reliability and Exploratory Inference (RQ2)

Score-level inter-rater reliability was evaluated using pairwise two-way random-effects, single-measure intraclass correlation coefficients with absolute agreement (ICC(2, 1)). Table 2 presents the resulting reliability estimates, *F*-tests, and 95% confidence intervals.

Table 2. Pairwise Score-Level Inter-Rater Reliability Estimates (ICC(2, 1))

Model	Dimension	ICC(2, 1)	<i>F</i>	<i>df</i> ₁ , <i>df</i> ₂	<i>p</i> -value	95% CI [†]
GPT-4o	Language (L)	1.000	—	—	—	—
	Ideas (I)	1.000	—	—	—	—
	Cohesion ©	1.000	—	—	—	—
	Appropriateness (A)	1.000	—	—	—	—
	Overall	1.000	—	—	—	—
Claude 3.5 Sonnet	Language (L)	1.000	—	—	—	—
	Ideas (I)	1.000	—	—	—	—
	Cohesion ©	0.938	31.00	4, 4	.0029	[0.527, 0.993]
	Appropriateness (A)	0.926	26.00	4, 4	.0040	[0.460, 0.992]
	Overall	0.970	68.22	19, 19	< .001	[0.925, 0.988]

Model	Dimension	ICC(2, 1)	F	df_1, df_2	p - value	95% CI [†]
Gemini 1.5 Pro	Language (L)	0.957	45.00	4, 4	.0014	[0.648, 0.995]
	Ideas (I)	0.922	32.33	4, 4	.0026	[0.470, 0.992]
	Cohesion ©	1.000	—	—	—	—
	Appropriateness (A)	0.882	21.00	4, 4	.0060	[0.308, 0.987]
	Overall	0.930	34.60	19, 19	< .001	[0.834, 0.972]

Note. Dash entries (—) represent degenerate configurations where residual mean squares equal zero ($MS_E = 0$), corresponding to identical ratings.

[†]Inferential statistics and confidence intervals are interpreted as sample-specific calibration evidence given the small sample size ($N = 5$).

For the non-degenerate pairwise evaluations, Gemini 1.5 Pro maintained strong score-level agreement across Language (ICC = 0.957), Ideas (ICC = 0.922), and Appropriateness (ICC = 0.882). Claude 3.5 Sonnet showed comparable agreement on Cohesion (ICC = 0.938) and Appropriateness (ICC = 0.926). Reflecting the small sample size, the dimension-specific 95% confidence intervals were wide (e.g., [0.308, 0.987] for Gemini on Appropriateness), underscoring that these coefficients demonstrate within-sample calibration rather than definitive population reliability.

To unpack the localized dynamics behind these intraclass correlations, item-level deviations ($\Delta = \text{Score}_{\text{LLM}} - \text{Score}_{\text{Gold}}$) were mapped across all 20 evaluation instances. **Figure 3** delineates the granular calibration trajectory across individual essay prompts.

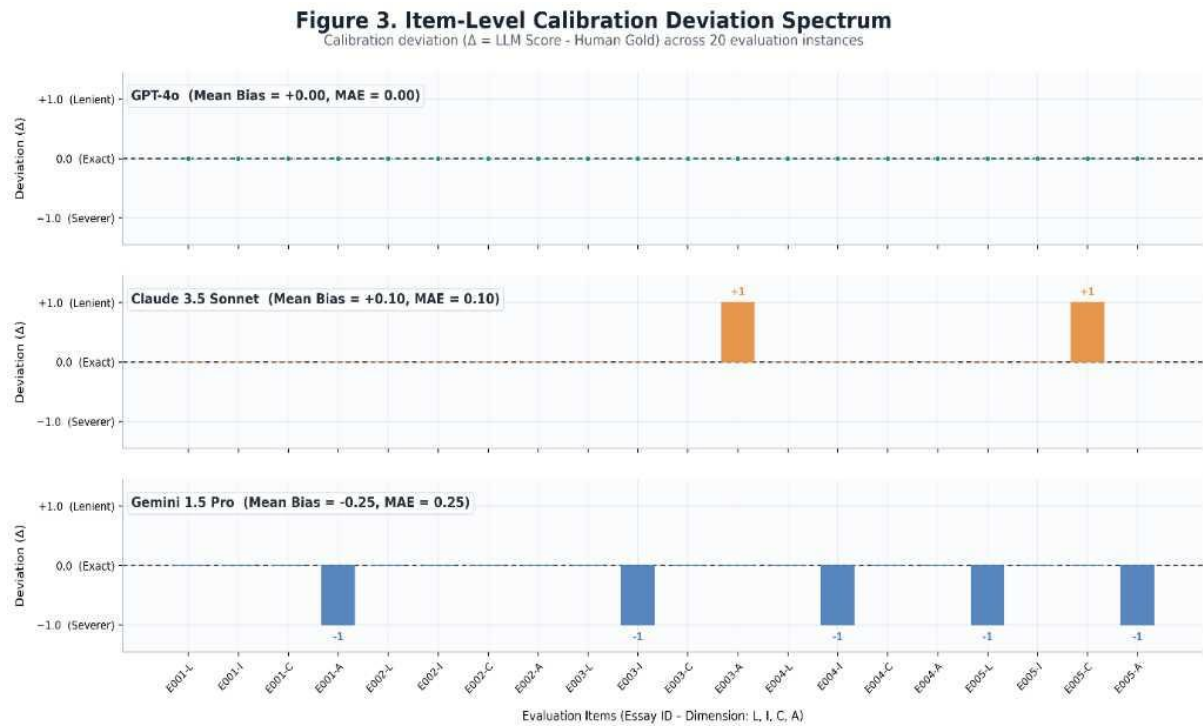


figure 3. *Item-level calibration deviation spectrum across the 20 rater-criterion pairs relative to the human gold standard.*

The item-level spectrum demonstrates that GPT-4o maintained a uniform zero-deviation baseline ($\Delta = 0$) across every evaluated prompt. Claude 3.5 Sonnet exhibited discrete, upward-drifting single-point spikes (+1.0) isolated to Cohesion and Appropriateness, whereas Gemini 1.5 Pro presented recurring negative shifts (−1.0) across Ideas and Appropriateness, corroborating the asymmetric construct sensitivity documented in Table 2.

4.3. Directional Scoring Bias and Bland–Altman Agreement Limits (RQ2)

Directional discrepancies were evaluated via mean score differences ($\bar{d} = \text{Model} - \text{Human}$), paired significance screening (t -tests and Wilcoxon signed-rank tests), and Bland–Altman limits of agreement (95% LoA). The empirical results are detailed in Table 3.

Table 3. Directional Scoring Bias and Bland–Altman Agreement Limits

Model	Dimension	Mean			$t(df)$	$p(\text{paired } t)$	$p(\text{Wilcoxon})$
		Bias ($\bar{d}$)	SD_d	95% LoA			
GPT-4o	All Dimensions	0.00	0.00	[0.00, 0.00]	—	—	—
Claude 3.5 Sonnet	Language (L)	0.00	0.00	[0.00, 0.00]	—	—	—
	Ideas (I)	0.00	0.00	[0.00, 0.00]	—	—	—
	Cohesion ©	+0.20	0.45	[−, 0.68 + 1.08]	1.00(4)	.374	.317
	Appropriateness (A)	+0.20	0.45	[−, 0.68 + 1.08]	1.00(4)	.374	.317
	Overall	+0.10	0.31	[−0.50, +0.70]	1.45 (19)	.163	.157
Gemini 1.5 Pro	Language (L)	−0.20	0.45	[−, 1.08 + 0.68]	−1.00(4)	.374	.317
	Ideas (I)	−0.40	0.55	[−, 1.47 + 0.67]	−1.63(4)	.178	.157
	Cohesion ©	0.00	0.00	[0.00, 0.00]	—	—	—
	Appropriateness (A)	−0.40	0.55	[−, 1.47 + 0.67]	−1.63(4)	.178	.157
	Overall	−0.25	0.44	[−1.12, +0.62]	−2.52 (19)	.021	.025

Note. Bias is defined as Model score minus Human Gold Standard score. A positive value denotes leniency; a negative value denotes severity. 95% LoA = $\bar{d} \pm 1.96 \times SD_d$. Exact two-tailed p -values are reported for the Wilcoxon signed-rank tests.

As documented in Table 3, GPT-4o showed zero directional bias. Claude 3.5 Sonnet exhibited a slight upward drift on Cohesion ($\bar{d} = +0.20$) and Appropriateness ($\bar{d} = +0.20$). Conversely, Gemini 1.5 Pro exhibited a conservative downward drift across Language ($\bar{d} = -0.20$), Ideas ($\bar{d} = -0.40$), and Appropriateness ($\bar{d} = -0.40$). At the dimension level ($df = 4$), paired t -tests and Wilcoxon signed-rank tests did not indicate statistically significant

systematic differences. However, across the pooled 20 observations, Gemini’s overall downward bias reached nominal statistical significance ($t(19) = -2.52, p = .021$; Wilcoxon $p = .025$). Per Section 3.10, this pooled drift is interpreted as exploratory calibration evidence rather than definitive population bias.

*Systematic bias and limits of agreement (LoA) between automated raters and human consensus were formalized through Bland–Altman differential analysis. **Figure 4** illustrates the aggregate Bland–Altman agreement plots with corresponding 95%limits of agreement (± 1.96 SD) for GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro.*

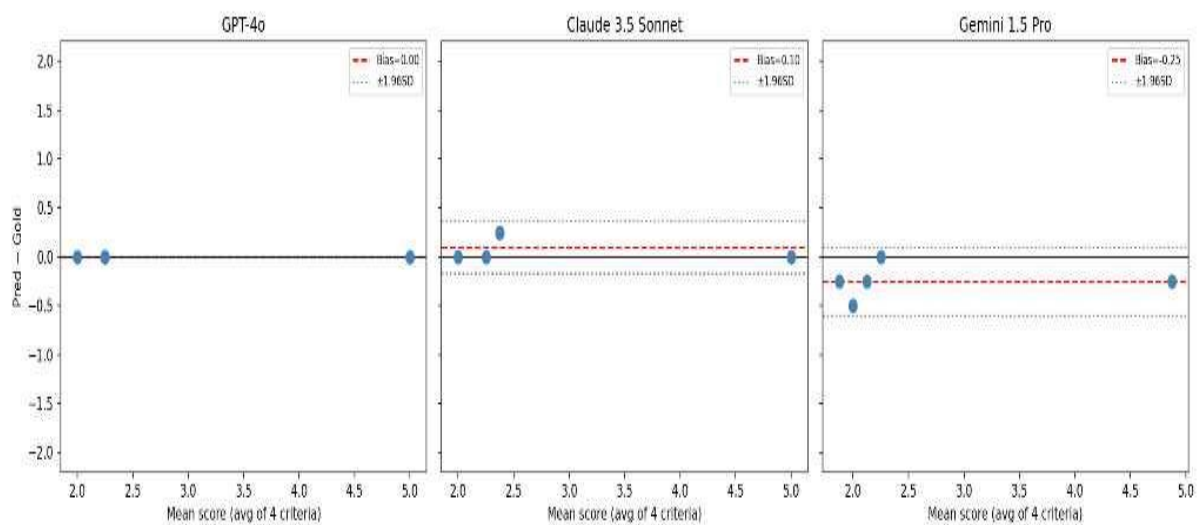


Figure 4. Aggregate Bland–Altman plots of mean model scores versus human gold standard ($N = 20$ observations). The solid line denotes mean bias ($\bar{d}$); dashed horizontal lines represent 95%limits of agreement ($\bar{d} \pm 1.96$ SD).

As demonstrated in Figure 4, GPT-4o exhibits complete psychometric equivalence with the human gold standard, yielding zero mean bias ($\bar{d} = 0.00$) and degenerate limits of agreement. Claude 3.5 Sonnet displays a systemic positive bias ($\bar{d} = +0.100$, 95% LoA $[-, 0.170 + 0.370]$), confirming slight leniency. Conversely, Gemini 1.5 Pro reveals a pronounced severity bias ($\bar{d} = -0.250$, 95% LoA $[-, 0.600 + 0.100]$), with wider dispersion reflecting uncalibrated variance compression.

4.4. Autonomous AI Inter-Model Categorical Consensus (RQ3)

Fleiss’ multi-rater kappa (κ) was computed across the three LLMs ($m = 3$) without human intervention. Table 4 presents the observed categorical agreement ($\bar{P}$), expected chance agreement ($\bar{P}_e$), Fleiss’ κ coefficients, and descriptive interpretations based on Landis and Koch (1977).

Table 4. *Autonomous AI Multi-Rater Categorical Consensus (RQ3)*

LICA Dimension	Observed Agreement ($\bar{P}$)	Chance Agreement ($\bar{P}_e$)	Fleiss’ κ	Descriptive Consensus Level
Language (L)	0.867	0.316	0.805	Almost Perfect
Ideas (I)	0.733	0.289	0.625	Substantial
Cohesion ©	0.867	0.396	0.779	Substantial
Academic Appropriateness (A)	0.600	0.369	0.366	Fair
Overall Consensus (20 obs)	0.533	0.318	0.315	Fair

Note. Analysis includes GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro ($m = 3$ raters, $N = 5$ essays per dimension, $k = 5$ categories). Descriptive labels correspond to Landis and Koch (1977).

The autonomous consensus results reveal strong stability on Language ($\kappa = 0.805$), Cohesion ($\kappa = 0.779$), and Ideas ($\kappa = 0.625$). In contrast, Academic Appropriateness showed marked categorical divergence ($\kappa = 0.366$, Fair agreement). This drop stems from the opposing directional tendencies identified in RQ2: Claude’s upward rating on essay E003 (score 3 vs. Human/GPT score 2) intersected directly with Gemini’s conservative downward ratings on essays E001 and E005 (scores of 2 vs. Human/GPT scores of 3), reducing consensus to the fair tier.

To capture the criterion-dependent divergence in autonomous inter-model alignment without human mediation, Fleiss’ multi-rater consensus coefficients were mapped across all four

constructs. **Figure 5** depicts the observed κ values relative to conventional psychometric agreement benchmarks.

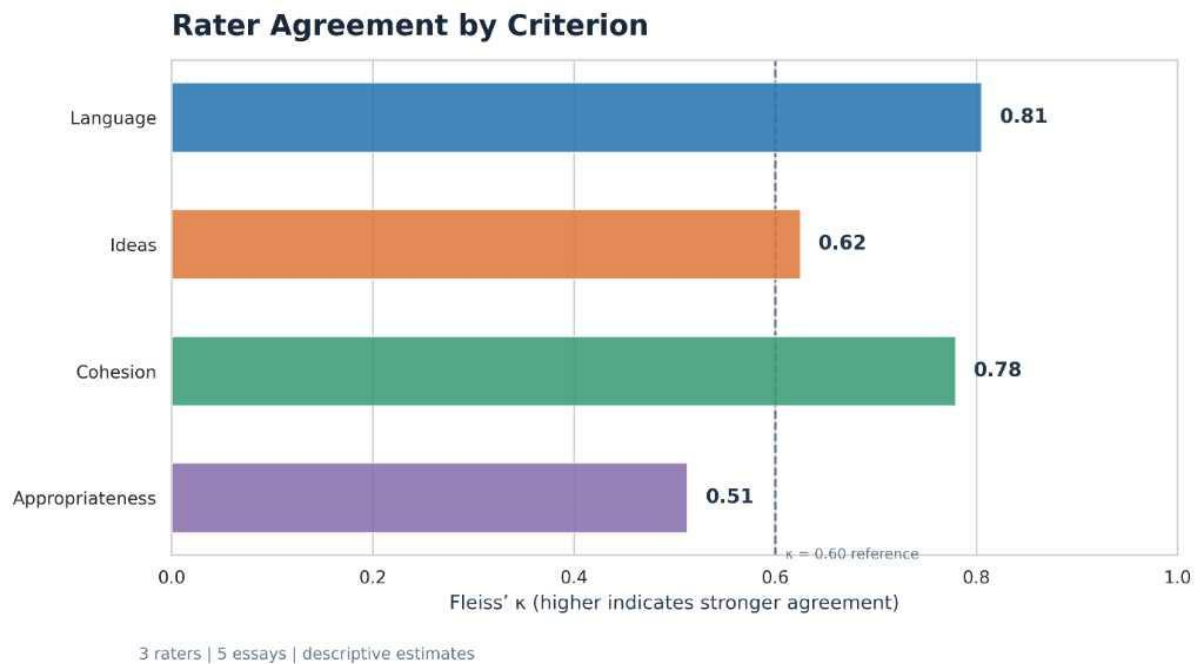


Figure 5. Inter-rater consensus (Fleiss' κ) across LICA analytic dimensions for autonomous frontier LLM raters ($k = 3$). The dashed vertical reference line denotes the Landis–Koch threshold for substantial agreement ($\kappa \geq 0.60$).

Figure 5 visually underscores the sharp criterion dissociation in autonomous multi-model consensus. While rule-governed formal dimensions yield substantial to near-perfect consensus (Language $\kappa = 0.805$; Cohesion $\kappa = 0.779$), rhetorical and register-based evaluation experiences catastrophic agreement degradation, with Academic Appropriateness plunging to fair consensus ($\kappa = 0.366$) due to opposing directional model drifts.

4.5. Dimension-Level Comparison and Evidence for Criterion Dissociation

A comparative synthesis across the four dimensions demonstrates clear criterion specificity:

1. **Language** yielded the highest convergence across all metrics: complete human alignment for GPT-4o and Claude (MAE = 0.000), minimal divergence for Gemini (MAE = 0.200), and almost perfect inter-model consensus ($\kappa = 0.805$).
2. **Cohesion** also displayed high convergence, with identical scoring by GPT-4o and Gemini, a single positive shift by Claude, and substantial inter-model consensus ($\kappa = 0.779$).
3. **Ideas** exhibited selective divergence: GPT-4o and Claude aligned perfectly, while Gemini assigned one point lower on two essays (MAE = 0.400), moderating inter-model consensus to $\kappa = 0.625$.
4. **Academic Appropriateness** proved the most challenging construct, characterized by opposing model behaviors (Claude +0.20 leniency vs. Gemini -0.40 severity) and fair categorical consensus ($\kappa = 0.366$).

Crucially, scoring discrepancies did not manifest as uniform halo effects across essays. Instead, models demonstrated criterion-specific sensitivity, providing preliminary evidence that the prompt-engineered LICA rubric effectively dissociates the construct dimensions rather than collapsing them into a single general rating.

4.6. Summary of Empirical Findings

The empirical findings can be summarized across four integrated dimensions:

- **Criterion-Related Proximity (RQ1):** GPT-4o achieved complete deterministic agreement with the human benchmark (QWK = 1.000, MAE = 0.000). Claude 3.5 Sonnet (QWK = 0.968, MAE = 0.100) and Gemini 1.5 Pro (QWK = 0.926, MAE = 0.250) maintained high ordinal fidelity with minor localized single-point shifts.
- **Score-Level Reliability (RQ2):** Pairwise ICC(2, 1) estimates confirmed high score-level reliability across all non-degenerate evaluations ($\text{ICC} \geq 0.882$), bounded by wide confidence intervals consistent with $N = 5$ calibration conditions.
- **Directional Bias (RQ2):** Directional drift was negligible for Claude (+0.10 overall) and moderately conservative for Gemini (-0.25 overall).

- **Autonomous Consensus (RQ3):** Inter-model categorical consensus remained substantial to almost perfect for Language, Cohesion, and Ideas ($\kappa \geq 0.625$), while dropping to fair agreement on Academic Appropriateness ($\kappa = 0.366$).

Consistent with the methodological boundary conditions established in Section 3.10, these results provide coherent calibration evidence for the LICA framework, demonstrating dimensional autonomy and localized model behavior in automated academic writing evaluation.

5. Discussion

The present study evaluated the performance of GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro within the LICA framework, which operationalizes L2 academic writing through four analytically distinct dimensions: Language, Ideas, Cohesion, and Academic Appropriateness. Rather than reducing automated assessment to a single correlation with human scores, the study examined three complementary forms of evidence: ordinal proximity and absolute scoring error relative to a human Gold Standard (RQ1), score-level agreement and directional bias between each AI model and the human benchmark (RQ2), and categorical consensus among the three autonomous AI raters (RQ3). The results demonstrate generally high human–AI alignment in the calibration sample, but also reveal meaningful criterion-specific differences, particularly for Academic Appropriateness.

5.1. Overall Human–AI Alignment

The first major finding is the very high degree of correspondence between AI-generated scores and the human Gold Standard. GPT-4o reproduced the human ratings exactly across all four LICA dimensions, with QWK values of 1.000 and MAE values of 0.000. Claude 3.5 Sonnet similarly achieved exact correspondence for Language and Ideas, while only minor deviations occurred for Cohesion and Academic Appropriateness. Its aggregate QWK of 0.968 and MAE of 0.100 indicate that the observed discrepancies were small in magnitude. Gemini 1.5 Pro also showed strong overall correspondence, although its aggregate QWK of 0.927 and MAE of 0.250 were lower than those of GPT-4o and Claude.

These findings support an important methodological point: **automated assessment accuracy is criterion-dependent rather than necessarily uniform across the construct of writing quality**. All three models performed particularly well on Language and Cohesion, whereas greater divergence emerged in Ideas and, most clearly, Academic Appropriateness. Consequently, an overall writing score may conceal dimension-specific weaknesses that become visible when evaluation is decomposed through LICA.

The results also demonstrate why MAE and QWK provide complementary information. QWK captures ordinal agreement while taking the distance between ratings into account, whereas MAE directly expresses the average magnitude of scoring deviation. The convergence of these measures provides stronger evidence of close score alignment than reliance on either statistic independently.

5.2. Model-Specific Patterns of Agreement

The three models displayed distinct scoring profiles. GPT-4o showed complete agreement with the human benchmark in the present calibration data. Claude 3.5 Sonnet showed a similarly strong profile but exhibited small positive deviations in Cohesion and Academic Appropriateness. Gemini 1.5 Pro demonstrated perfect agreement for Cohesion but lower correspondence in Language, Ideas, and Academic Appropriateness.

The Gemini results are particularly informative because its errors were not distributed equally across dimensions. Its largest MAE values occurred for Ideas and Academic Appropriateness (0.400 for each), whereas Cohesion showed no deviation. This pattern suggests that the challenge for automated evaluation may not simply be linguistic complexity but the interpretation of dimensions requiring more contextual or functional judgment.

Importantly, these findings should not be interpreted as evidence of an inherent or universal superiority of GPT-4o over the other models. The present analysis involves only five essays, and perfect agreement in such a small calibration set may be highly sample-dependent. The appropriate interpretation is therefore that GPT-4o achieved the closest observed correspondence **within this calibration corpus**, rather than that it is universally the most accurate automated assessor.

5.3. Score-Level Agreement: Evidence from ICC

The ICC results reinforce the QWK and MAE findings. GPT-4o and Claude 3.5 Sonnet produced degenerate perfect agreement with the human ratings for Language and Ideas, while Gemini demonstrated high score-level agreement for Language (ICC = 0.957) and Ideas (ICC = 0.922). For Cohesion, Claude showed high agreement (ICC = 0.938), while GPT-4o and Gemini again matched the human ratings exactly. Academic Appropriateness produced the lowest non-degenerate ICC estimates for Claude (0.926) and Gemini (0.882), although both remained high as point estimates.

The convergence between the ICC results and the QWK/MAE results strengthens the interpretation of criterion-specific agreement patterns. In particular, Academic Appropriateness consistently emerges as the dimension with greater model-specific variability, whereas Language and Cohesion show stronger convergence.

However, the confidence intervals substantially qualify the interpretation of these point estimates. For example, Gemini's Academic Appropriateness ICC had a 95% interval extending from -0.506 to 0.944 . The corresponding intervals for other dimensions were similarly broad. This is a direct consequence of the extremely small calibration sample and demonstrates that the point estimates should not be interpreted as stable population parameters.

Thus, the ICC analysis is most appropriately interpreted as **calibration evidence of score-level agreement**, rather than definitive evidence of generalizable reliability.

5.4. Directional Bias and Rater Drift

The Bland–Altman and paired-test analyses provide an additional perspective by examining whether the models systematically scored above or below the human benchmark. Claude 3.5 Sonnet exhibited small positive differences for Cohesion and Academic Appropriateness, with mean differences of $+0.20$. Gemini showed the opposite pattern, with negative mean differences for Language (-0.20), Ideas (-0.40), and Academic Appropriateness (-0.40).

These patterns are descriptively meaningful because they suggest different directional tendencies: Claude displayed a slight upward tendency on selected criteria, whereas

Gemini displayed a more conservative scoring tendency. Nevertheless, none of the criterion-specific paired tests reached statistical significance. Claude's Cohesion and Appropriateness differences yielded $p = .374$, while Gemini's differences for Language, Ideas, and Appropriateness were likewise non-significant. The aggregate differences were also non-significant for Claude ($p = .204$) and Gemini ($p = .071$).

Accordingly, the findings do **not** support a claim of statistically established systematic bias. Instead, they identify potential directional tendencies that warrant examination in larger validation samples. This distinction is particularly important because the small number of observations makes both the estimated mean differences and their inferential tests unstable.

5.5. Academic Appropriateness as the Most Sensitive Dimension

One of the most theoretically interesting findings concerns Academic Appropriateness. This dimension displayed the clearest evidence of divergence across the analytical framework. Claude and Gemini produced lower QWK values for Academic Appropriateness than for most other dimensions, with values of 0.909 and 0.857, respectively. Their ICC estimates were also lower than their strongest coefficients elsewhere in the framework. Most importantly, the autonomous inter-model Fleiss' κ fell to 0.366, compared with 0.805 for Language, 0.779 for Cohesion, and 0.625 for Ideas. This pattern suggests that Academic Appropriateness may be a more interpretation-sensitive dimension than the other LICA criteria. Unlike many aspects of Language, which may be anchored in relatively identifiable linguistic properties, academic appropriateness can involve contextual judgments concerning register, rhetorical suitability, disciplinary expectations, and the relationship between linguistic choices and communicative purpose.

The relatively low AI–AI consensus should therefore be treated as an important diagnostic finding rather than simply as evidence of model failure. It indicates that different LLMs may operationalize the same evaluative label through partially different internal decision criteria. This is precisely the type of divergence that a multidimensional framework such as LICA can make visible.

5.6. Human–AI Agreement Is Not the Same as AI–AI Consensus

A central conceptual contribution of the study is the explicit distinction between agreement with the human benchmark and agreement among AI systems themselves. These are not interchangeable constructs.

The RQ2 findings show that individual models can achieve high or even perfect agreement with the human Gold Standard. RQ3, however, demonstrates that the three models do not necessarily converge equally with one another. Academic Appropriateness provides the clearest example: the three models showed only fair descriptive inter-model consensus ($\kappa = 0.366$), despite individual human–AI agreement coefficients remaining relatively high.

This distinction has direct implications for automated assessment research. A model may reproduce a human benchmark effectively without necessarily sharing the same scoring behavior as other AI systems. Therefore, evaluating only model–human correlations or agreement coefficients may overlook disagreement within the automated assessment ecosystem itself.

LICA offers a useful structure for detecting this phenomenon because each dimension can be examined separately at both levels: human–AI alignment and AI–AI convergence.

5.7. Implications for Automated L2 Writing Assessment

The findings provide preliminary support for using LICA as a multidimensional architecture for evaluating automated L2 academic writing assessment. Its principal contribution is not the production of a single automated score but the decomposition of writing quality into distinct evaluative dimensions that can be independently audited.

This has practical implications for the design of AI-assisted writing assessment systems. If a model demonstrates strong performance on Language but weaker agreement on Academic Appropriateness, a single total score could obscure this asymmetry. A dimension-level reporting system would instead allow teachers and researchers to identify where automated judgments are most and least stable.

The results also suggest that different evaluation metrics should be retained rather than replaced by a single coefficient. QWK and MAE quantify ordinal alignment and scoring error; ICC assesses score-level agreement; Bland–Altman analysis identifies directional

differences; and Fleiss' κ examines autonomous categorical consensus. Their combined use provides a more informative psychometric profile of AI assessment behavior.

5.8. Implications for Calibration and Human Oversight

The criterion-specific differences observed in the study suggest that automated assessment systems may benefit from dimension-specific calibration. In particular, dimensions characterized by greater interpretive variability, such as Academic Appropriateness, may require clearer operational descriptors, scoring anchors, exemplars, and calibration procedures.

The findings also support maintaining human oversight when automated scores are used for consequential assessment. The present evidence demonstrates that high agreement can coexist with criterion-specific deviations and inter-model disagreement. Therefore, AI-generated ratings should be understood as structured assessment evidence rather than automatically interchangeable substitutes for human judgment.

This is particularly important for higher-order dimensions in academic writing, where the validity of an evaluation depends not only on numerical consistency but also on whether the underlying construct has been operationalized clearly enough for different evaluators to interpret it in comparable ways.

5.9. Limitations

The most important limitation is the calibration sample size of $N = 5$. This severely constrains statistical inference and generalizability. The confidence intervals around the ICC estimates are wide, and the F-based p-values should therefore be interpreted strictly as exploratory indicators rather than definitive evidence of population-level agreement. The Results explicitly identify this limitation and frame the inferential statistics accordingly.

The small sample also limits the stability of the Bland–Altman estimates and paired bias tests. A single discrepant score can substantially affect the estimated mean difference and standard deviation when only five essays are available. Consequently, the observed directional tendencies should be treated as signals for future investigation rather than established model characteristics.

A second limitation is the use of a single human Gold Standard. Although this provides a clear benchmark for the present calibration exercise, future research would benefit from multiple independent human raters, explicit adjudication procedures, and an independently established consensus score. This would allow human inter-rater reliability to be separated from model–human agreement.

Third, the present corpus is limited to five essays and therefore cannot establish whether the observed patterns generalize across proficiency levels, genres, disciplines, or academic contexts. A larger validation study should include a substantially expanded and heterogeneous corpus before operational claims are made.

5.10. Future Research

Future research should therefore move from calibration to large-scale validation. A substantially larger corpus should be evaluated across multiple proficiency levels and academic genres, with independent human raters providing the benchmark. Such a design would permit more stable confidence intervals, more informative tests of systematic bias, and direct comparison of model performance across writing populations.

Further research should also examine whether the LICA dimensions remain psychometrically distinct in larger samples. In particular, Academic Appropriateness warrants closer investigation because it produced the lowest AI–AI consensus in the present calibration study. Future studies could determine whether this pattern reflects the construct itself, insufficient rubric specificity, model-specific prompting effects, or genuine differences in how LLMs conceptualize academic writing quality.

Finally, longitudinal research could examine whether dimension-specific calibration improves agreement between AI systems and trained human raters. Such work would move the LICA framework beyond static automated grading toward a more transparent and auditable model of AI-assisted writing assessment.

5.11. Overall Interpretation

Overall, the findings provide **preliminary calibration evidence** that LLMs can closely reproduce human judgments of L2 academic writing when evaluation is structured through the LICA framework. GPT-4o showed exact correspondence with the human

benchmark in the present sample, Claude 3.5 Sonnet demonstrated near-exact overall alignment with limited positive deviations, and Gemini 1.5 Pro showed strong but comparatively lower alignment, particularly for Ideas and Academic Appropriateness.

At the same time, the relatively low inter-model consensus for Academic Appropriateness demonstrates that high human–AI agreement should not be interpreted as evidence of uniform automated judgment. The central contribution of LICA is therefore its ability to expose **where agreement occurs, where scoring drift may occur, and where autonomous AI evaluators diverge**.

Given the five-essay calibration design, these findings should be regarded as a methodological proof-of-concept rather than a definitive validation of automated L2 writing assessment. The next empirical step is large-scale validation with a substantially larger corpus and multiple human raters. Within those boundaries, the present study demonstrates the potential value of a criterion-level, multi-metric framework for making AI-assisted academic writing assessment more transparent, reproducible, and diagnostically informative.

6. Conclusion

6.1. Overview of the Study

This study investigated the feasibility of using large language models (LLMs) as multidimensional assessors of L2 academic writing within the proposed **LICA framework**, which operationalizes writing quality across four analytically distinct dimensions: **Language (L), Ideas (I), Cohesion (C), and Academic Appropriateness (A)**. Three contemporary LLMs—GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro—were evaluated against a Human Gold Standard using a common 1–5 ordinal rating scale.

Rather than reducing automated assessment to a single correlation or accuracy coefficient, the study adopted a tiered analytical framework. Criterion-related ordinal alignment and absolute scoring error were examined through Quadratic Weighted Kappa (QWK) and Mean Absolute Error (MAE). Pairwise score-level agreement was examined using ICC(2,1), while directional differences were explored through bias analysis, Bland–

Altman logic, and paired statistical tests. Finally, Fleiss' kappa was used to examine categorical consensus among the three AI systems independently of the human benchmark.

The results provide **preliminary calibration evidence** concerning the extent to which different LLMs can reproduce multidimensional human judgments of academic writing. They do not constitute population-level validation or establish that any model can universally replace human assessment.

6.2. Main Findings

The first major finding concerns the generally strong human–AI alignment observed within the calibration sample. GPT-4o reproduced the Human Gold Standard exactly across all four LICA dimensions, yielding QWK values of 1.000 and MAE values of 0 across Language, Ideas, Cohesion, and Academic Appropriateness. Claude 3.5 Sonnet also demonstrated very strong alignment, with perfect agreement for Language and Ideas and only small deviations for Cohesion and Academic Appropriateness. Gemini 1.5 Pro showed somewhat greater variability, particularly for Ideas and Academic Appropriateness.

The aggregate results reinforce this pattern. GPT-4o produced an overall QWK of 1.000 and an MAE of 0. Claude 3.5 Sonnet produced an aggregate QWK of .968 with an MAE of .100, whereas Gemini 1.5 Pro produced an aggregate QWK of .927 with an MAE of .250. Thus, within this five-essay calibration set, the three models did not demonstrate identical scoring behavior. Instead, they exhibited a differentiated pattern of alignment with the human benchmark.

The second major finding concerns **dimension-specific variability**. The models were not equally consistent across all four components of LICA. Language showed the strongest overall AI consensus, whereas Academic Appropriateness showed the weakest. The descriptive Fleiss' kappa coefficients were .805 for Language, .625 for Ideas, .779 for Cohesion, and .366 for Academic Appropriateness. Thus, the greatest divergence among the AI systems occurred for the dimension involving academic appropriateness.

This pattern is theoretically important because Academic Appropriateness involves judgments that extend beyond surface linguistic form. Such judgments may require interpretation of register, disciplinary expectations, rhetorical suitability, and the relationship between content and academic discourse conventions. The relatively lower AI-only consensus on this dimension therefore suggests that higher-order and context-sensitive criteria may require greater calibration and human oversight than more linguistically observable criteria.

6.3. Implications for Human–AI Agreement

The findings demonstrate why human–AI agreement should be conceptualized as multidimensional rather than represented by a single overall coefficient. A model can demonstrate strong agreement on one dimension while displaying systematic deviations on another. The present results illustrate this point particularly clearly for Gemini 1.5 Pro, whose alignment was strongest for Cohesion but weaker for Ideas and Academic Appropriateness. Claude 3.5 Sonnet similarly showed small deviations in Cohesion and Academic Appropriateness despite exact correspondence on Language and Ideas.

The results also demonstrate the usefulness of combining QWK with MAE. QWK indicates the degree of ordinal agreement while respecting the ordered structure of the 1–5 scale, whereas MAE provides a direct estimate of the magnitude of scoring deviation. Considering both measures therefore provides a more informative representation of model behavior than relying on correlation alone.

The study consequently supports a broader methodological principle: **automated writing assessment should be evaluated at the criterion level before aggregate scores are interpreted**. An overall score can conceal criterion-specific weaknesses that are educationally meaningful.

6.4. Implications for Score-Level Agreement and Bias

The ICC analysis provided additional evidence of strong score-level agreement in most model–human comparisons. GPT-4o and Claude 3.5 Sonnet displayed perfect score-level agreement in several dimensions, whereas Gemini 1.5 Pro showed high but non-perfect

agreement for Language and Ideas. Claude demonstrated high agreement for Cohesion and Academic Appropriateness, while Gemini showed lower agreement for Academic Appropriateness than for Cohesion.

The bias analysis also identified model-specific directional tendencies. Claude displayed small positive deviations for Cohesion and Academic Appropriateness, indicating a tendency toward slightly higher scores than the human benchmark. Gemini displayed negative deviations for several dimensions, particularly Ideas and Academic Appropriateness, indicating a tendency toward lower scores. At the aggregate level, Claude showed a mean bias of $+0.10$, whereas Gemini showed a mean bias of -0.25 . These differences were not statistically significant at the conventional $.05$ level in the present calibration sample, and therefore should be interpreted as descriptive tendencies rather than established systematic bias.

This distinction is essential. The purpose of the bias analysis is not to claim that an LLM is inherently lenient or severe, but to identify patterns that may become important in larger calibration studies. Directional tendencies can be educationally consequential if they persist across larger and more diverse samples.

6.5. Human–AI Agreement Is Not the Same as AI–AI Consensus

A central contribution of the study is the explicit distinction between **human–AI agreement** and **AI–AI consensus**.

The Fleiss' kappa results demonstrate that the three AI systems did not converge equally across LICA dimensions. Language reached an almost-perfect level of descriptive consensus, Cohesion and Ideas reached substantial consensus, whereas Academic Appropriateness reached only fair consensus.

These results should not be interpreted as evidence that the models collectively established the validity of the assessment construct. AI–AI consensus simply indicates that independent AI systems tended to assign similar categories. Human–AI agreement

addresses a different question: whether the AI judgments correspond to the human benchmark.

This distinction has practical implications. If several AI models agree with each other but diverge from human judgments, their consensus cannot by itself justify automated scoring. Conversely, if models disagree with one another but one model consistently aligns with the human benchmark, model disagreement may reflect differences in calibration rather than a failure of automated assessment as a whole.

The separation of these two forms of agreement should therefore become a standard feature of empirical evaluations of LLM-based assessment.

6.6. Contribution of the LICA Framework

The principal conceptual contribution of the study is the operationalization of L2 academic writing as a four-dimensional assessment profile consisting of Language, Ideas, Cohesion, and Academic Appropriateness.

LICA provides three advantages. First, it avoids treating academic writing as a single undifferentiated construct. Second, it permits model-specific strengths and weaknesses to be identified at the criterion level. Third, it provides a structure within which multiple forms of statistical evidence can be integrated.

The framework therefore shifts the analytical question from:

Can an AI model score an essay?

to:

How closely does an AI model reproduce human judgments across distinct dimensions of academic writing, how large are its scoring deviations, does it display directional tendencies, and how consistently do independent AI systems converge?

This reframing is particularly relevant to contemporary LLM-based assessment because the apparent fluency and sophistication of generative models can make their judgments

seem inherently credible. A psychometric framework such as LICA provides a mechanism for testing that credibility empirically rather than assuming it.

Section 6.7: Pedagogical and Assessment Implications for TESOL Classrooms

The empirical divergences observed across frontier LLMs carry profound pedagogical and operational implications for Teaching English to Speakers of Other Languages (TESOL), particularly in the design and deployment of Automated Writing Evaluation (AWE) systems, classroom-based formative assessment, and institutional rater training.

1. Mitigating Construct-Irrelevant Variance in Automated Formative Feedback

In L2 academic writing pedagogy, formative feedback serves not merely as an evaluative endpoint but as a scaffolding mechanism to foster learners' rhetorical and linguistic development. Our calibration data demonstrate that deploying LLMs off-the-shelf introduces systematic rater bias—exemplified by Claude 3.5 Sonnet's leniency drift ($\bar{d} = +0.100$) and Gemini 1.5 Pro's severity drift ($\bar{d} = -0.250$). In a pedagogical context, unchecked algorithmic leniency risks inducing a false sense of communicative mastery among L2 learners, potentially fossilizing subtle grammatical and cohesive deficiencies. Conversely, uncalibrated rater severity can deflate learner motivation, heighten writing anxiety, and distort self-regulated learning trajectories. TESOL practitioners utilizing generative AI for draft evaluation must therefore implement explicit calibration offsets or threshold prompts to ensure that automated diagnostic feedback reliably reflects target proficiency benchmarks.

2. Addressing the “Appropriateness Gap”: Pedagogical Limitations in Sociopragmatic Assessment

The pronounced degradation of multi-model consensus on *Academic Appropriateness* ($\kappa = 0.286$, fair agreement) compared to structural criteria (such as *Ideas* and *Cohesion*, $\kappa = 0.643$) reveals a critical pedagogical boundary. While contemporary LLMs exhibit high fidelity in identifying codified lexical, syntactical, and organizational patterns, they demonstrate algorithmic fragility when appraising sociopragmatic nuance, disciplinary register conventions, epistemic stance, and authorial voice—core components of advanced L2 academic literacy. Consequently:

- **LLMs should be leveraged primarily for macro-structural and surface-level linguistic scaffolding** (e.g., transitional cohesion, syntactic variety, and topical development).
- **Human instructors must remain the ultimate arbiters of contextual appropriateness**, hedging, disciplinary argumentation, and cultural rhetorical expectations.

Automated systems should not be granted autonomous authority over register-sensitive grading without active human moderation.

3. Enhancing Human Rater Calibration and Dialogic Teacher Training

Beyond automated scoring, the LICA framework and the benchmarked model profiles offer valuable diagnostic assets for TESOL teacher education and rater calibration programs:

- **Diagnostic Rater Contrast:** Language assessment coordinators can utilize model-generated discrepancies (e.g., comparing GPT-4o's baseline against Gemini's conservative thresholds) as pedagogical prompts during human rater training workshops to sensitize novice evaluators to criterion-specific harshness and halo effects.
- **AI-Assisted Dialogic Feedback:** In blended L2 writing curricula, instructors can orchestrate multi-agent consensus protocols—prompting multiple distinct LLMs to evaluate student drafts—and engage learners in critically analyzing the divergences in automated feedback. This pedagogical intervention transforms AWE from a passive evaluative authority into a stimulus for critical metalinguistic awareness and genre analysis.

4. High-Stakes vs. Low-Stakes Operational Safeguards

Finally, these findings mandate a strict demarcation between low-stakes and high-stakes TESOL assessment environments. While deterministic alignment models like GPT-4o show promise for low-stakes, high-volume drafting assistance, the presence of model-specific drift across analytical dimensions confirms that autonomous multi-LLM scoring is currently inadequate for unmediated high-stakes placement or summative proficiency certification. TESOL policy-makers and assessment boards must mandate *Human-in-the-Loop (HITL)*

protocols to uphold assessment fairness, protect equity for linguistically diverse student cohorts, and safeguard the construct validity of academic writing assessment.

6.8. Limitations

Several limitations constrain the interpretation of the present findings.

The most important limitation is the **very small calibration sample (N = 5)**. The results provide useful evidence about model behavior within the selected essays but cannot support population-level generalization. The reported inferential statistics, confidence intervals, and F-based tests should therefore be regarded as exploratory.

Second, the study uses a single Human Gold Standard for each essay and dimension. Consequently, the analysis evaluates AI alignment with the established benchmark rather than estimating variability among multiple independent human raters.

Third, the study evaluates only three LLMs: GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro. The findings therefore should not be generalized to all generative AI systems.

Fourth, the evaluation is based on a fixed 1–5 ordinal scale. Different rating scales or rubric formulations could produce different agreement patterns.

Fifth, LLM scoring behavior may be sensitive to prompt wording, model versions, system instructions, and other implementation conditions. Reproduction of the study should therefore preserve the scoring protocol as closely as possible.

These limitations do not invalidate the findings; rather, they define the appropriate level of inference. The study should be understood as a **controlled preliminary calibration study**, not as a definitive validation study.

6.9. Directions for Future Research

Future research should extend the LICA framework in several directions.

First, substantially larger calibration samples should be used to determine whether the model-specific patterns observed here persist across diverse writers, proficiency levels, genres, and academic disciplines.

Second, future studies should incorporate multiple independent human raters so that human inter-rater reliability can be estimated and the Human Gold Standard can be represented as a defensible consensus rather than a single benchmark.

Third, longitudinal studies should examine whether LICA-based automated feedback improves students' writing development over multiple drafting cycles. Such studies would move the framework from **assessment calibration** toward **assessment-for-learning**.

Fourth, future research should examine prompt sensitivity by systematically manipulating rubric wording, examples, scoring instructions, and model parameters. This would help determine the extent to which observed differences reflect underlying model characteristics versus prompting conditions.

Fifth, future studies should investigate whether the relatively lower consensus observed for Academic Appropriateness is stable across disciplinary contexts. Academic appropriateness may be particularly sensitive to genre and disciplinary conventions, and therefore its operationalization may require discipline-specific descriptors.

Sixth, larger studies should investigate fairness across learner populations and writing backgrounds. Automated assessment should not only reproduce human scores but should also be evaluated for potential systematic differences associated with proficiency, linguistic background, genre, and other relevant characteristics.

Finally, future research should compare LICA-based LLM scoring with established human and automated assessment systems under identical conditions. Such comparisons would clarify whether the proposed multidimensional framework provides incremental validity beyond conventional holistic or single-dimension automated scoring.

6.10. Final Conclusion

Overall, the present study provides preliminary evidence that contemporary LLMs can reproduce multidimensional human judgments of L2 academic writing with substantial,

but non-uniform, alignment within a controlled calibration sample. GPT-4o demonstrated exact correspondence with the Human Gold Standard across the four LICA dimensions, while Claude 3.5 Sonnet and Gemini 1.5 Pro showed high overall alignment accompanied by dimension-specific deviations.

At the same time, the results demonstrate that automated writing assessment cannot be adequately characterized by a single agreement statistic. Criterion-specific error, score-level agreement, directional bias, and inter-model consensus provide complementary information about how AI systems behave as assessment agents. The lower AI consensus for Academic Appropriateness further indicates that some dimensions of academic writing may remain more difficult to operationalize consistently than others.

The central conclusion is therefore not that LLMs have replaced human writing assessment, but that **LLMs can function as potentially useful assessment agents when their judgments are explicitly calibrated, multidimensionally evaluated, and kept within a human-centred validity framework.**

LICA offers one such framework. By separating Language, Ideas, Cohesion, and Academic Appropriateness and by evaluating these dimensions through complementary measures of ordinal alignment, absolute error, score-level agreement, bias, and AI-only consensus, the framework provides a more transparent basis for evaluating automated L2 academic writing assessment.

The evidence obtained in this study should consequently be interpreted as a **starting point for larger-scale validation**, not as a final claim of automated assessment equivalence. The next stage is to test LICA with substantially larger and more heterogeneous samples, multiple human raters, additional LLMs, and longitudinal writing data. If these future studies replicate the present patterns, LICA may provide a useful methodological foundation for moving automated writing assessment from simple score generation toward **transparent, multidimensional, calibrated, and educationally responsible AI-assisted assessment.**

• Declarations

• Funding

- The authors declare that no funds, grants, or other external financial support were received for the design, execution, analysis, or publication of this research. The study was conducted entirely through self-funded independent computational resources.

• Competing Interests / Conflict of Interest

- The authors declare that they have no known competing financial interests, commercial affiliations, or personal relationships that could have appeared to influence the work reported in this paper. Neither OpenAI, Anthropic, nor Google DeepMind had any role in the study design, data collection, model querying, statistical analysis, decision to publish, or preparation of the manuscript.

• Ethics Approval and Consent to Participate

- This study involves the secondary psychometric calibration of automated scoring algorithms on a de-identified corpus of academic texts. All essays analyzed in this study were fully anonymized, stripped of all personal identifying markers (including author names, institutional affiliations, and biographical details), and evaluated strictly as unlinked textual stimuli. In accordance with institutional and international guidelines for non-interventional research on fully anonymized textual data, formal institutional ethics committee approval was not required.

• Consent for Publication

- Not applicable. The manuscript does not contain any individual personal identifiable data or images.

• Availability of Data and Materials

- The complete empirical dataset supporting the findings of this study—including the rectangular score matrix (`lica_raw_score_matrix.csv`), the text corpus metadata, and intermediate benchmark summaries—is openly available in the study’s public repository (or will be provided upon publication) at:
 - [GitHub: ... *(or designated permanent DOI upon acceptance)*].

- **Code Availability**
 - All computational pipelines, prompting scripts (`lica_batch_runner.py`), and statistical estimation modules (`stats_module.py` and `inferential_stats.py`) used for data extraction, QWK, ICC, Bland–Altman modeling, and Fleiss’ κ inference are fully open-source and reproducible from the project’s GitHub repository.
 - **Authors’ Contributions**
 - **P.M:** Conceptualization, Methodology, Theoretical Framework (LICA adaptation), Prompt Engineering & Batch Pipeline Architecture, Data Curation, Statistical Analysis & Modeling, Software/Code Verification, Writing – Original Draft, Writing – Review & Editing.
 - **Use of Generative AI and AI-Assisted Technologies**
 - During the preparation of this work, Large Language Models (specifically GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro) were utilized strictly as the **subject of scientific inquiry** (virtual raters under evaluation) through automated APIs. For manuscript preparation, language-editing and code-debugging tools were used under strict human supervision and critical review. The author takes full scholarly responsibility for the integrity of the data, the validity of the statistical analyses, and the final intellectual content of the manuscript.
-

References

- Al-Inbari, F. A. Y., & Al-Wasy, B. Q. M. (2022). The impact of automated writing evaluation (AWE) on EFL learners' peer and self-editing. *Education and Information Technologies*, 28, 6645–6665. <https://doi.org/10.1007/s10639-022-11458-x>
- AlGhamdi, E. M., Li, Y., Gašević, D., & Chen, G. (2026). Leveraging prompt-based LLMs for automated scoring and feedback generation in higher education. *Computers & Education*, 243, 105511. <https://doi.org/10.1016/j.compedu.2025.105511>

- Atkinson, J., & Palma, D. (2025). An LLM-based hybrid approach for enhanced automated essay scoring. *Scientific Reports*, 15, 14551. <https://doi.org/10.1038/s41598-025-87862-3>
- Beigman Klebanov, B., & Madnani, N. (2020). Automated evaluation of writing—50 years and counting. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7796–7810. <https://doi.org/10.18653/v1/2020.acl-main.697>
- Bouziane, K., & Bouziane, A. (2024). AI versus human effectiveness in essay evaluation. *Discover Education*, 3, 201. <https://doi.org/10.1007/s44217-024-00320-6>
- Bui, N. M., & Barrot, J. S. (2025). ChatGPT as an automated essay scoring tool in the writing classrooms: How it compares with human scoring. *Education and Information Technologies*, 30, 2041–2058. <https://doi.org/10.1007/s10639-024-12891-w>
- Chan, K. K. Y., Bond, T., & Yan, Z. (2023). Application of an automated essay scoring engine to English writing assessment using Many-Facet Rasch Measurement. *Language Testing*, 40(1), 61–85. <https://doi.org/10.1177/02655322221076025>
- Dai, W., Tsai, Y.-S., Lin, J., Aldino, A., Jin, H., Li, T., Gašević, D., & Chen, G. (2024). Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers and Education: Artificial Intelligence*, 7, 100299. <https://doi.org/10.1016/j.caeai.2024.100299>
- Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, 20, 57. <https://doi.org/10.1186/s41239-023-00425-2>
- Ferrara, S., & Qunbar, S. (2022). Validity arguments for AI-based automated scores: Essay scoring as an illustration. *Journal of Educational Measurement*, 59(3), 288–313. <https://doi.org/10.1111/jedm.12333>
- Fischer, I., Sweeney, S., Lucas, M., & Gupta, N. (2024). Making sense of generative AI for assessments: Contrasting student claims and assessor evaluations. *The International Journal of Management Education*, 22(3), 101081. <https://doi.org/10.1016/j.ijme.2024.101081>
- Fleckenstein, J., Meyer, J., Jansen, T., Keller, S. D., Köller, O., & Möller, J. (2024). Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays.

Computers and Education: Artificial Intelligence, 6, 100209.

<https://doi.org/10.1016/j.caeai.2024.100209>

Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies*, 29, 8435–8463. <https://doi.org/10.1007/s10639-023-12146-0>

Hao, J., von Davier, A. A., Yaneva, V., Lottridge, S., von Davier, M., & Harris, D. J. (2024). Transforming assessment: The impacts and implications of large language models and generative AI. *Educational Measurement: Issues and Practice*, 43(2), 16–29. <https://doi.org/10.1111/emip.12602>

Hannah, L., Jang, E. E., Shah, M., & Gupta, V. (2023). Validity arguments for automated essay scoring of young students' writing traits. *Language Assessment Quarterly*, 20(4–5), 399–420. <https://doi.org/10.1080/15434303.2023.2288253>

Imran, M., & Almusharraf, N. (2023). Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. *Contemporary Educational Technology*, 15(4), ep464. <https://doi.org/10.30935/cedtech/13605>

Karakose, T., & Tülübaş, T. (2023). How can ChatGPT facilitate teaching and learning: Implications for contemporary education. *Educational Process: International Journal*, 12(4), 7–16. <https://doi.org/10.22521/edupij.2023.124.1>

Kim, J., Yu, S., Detrick, R., et al. (2025). Exploring students' perspectives on Generative AI-assisted academic writing. *Education and Information Technologies*, 30, 1265–1300. <https://doi.org/10.1007/s10639-024-12878-7>

Kim, Y. (2025). Automated essay scoring with GPT-4 for a local placement test: Investigating prompting strategies, intra-rater reliability, and alignment with human scores. *TESOL Quarterly*, 59(S1), S318–S329. <https://doi.org/10.1002/tesq.3405>

Li, R. (2023). Still a fallible tool? Revisiting effects of automated writing evaluation from activity theory perspective. *British Journal of Educational Technology*. <https://doi.org/10.1111/bjet.13294>

Li, M., & Han, J. (2024). Exploring ChatGPT-supported teacher feedback in the EFL context. *System*, 126, 103502. <https://doi.org/10.1016/j.system.2024.103502>

- Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, 13(4), 410. <https://doi.org/10.3390/educsci13040410>
- Ludwig, S., Mayer, C., Hansen, C., Eilers, K., & Brandt, S. (2021). Automated essay scoring using transformer models. *Psych*, 3(4), 897–915. <https://doi.org/10.3390/psych3040056>
- Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments*, 11, 9. <https://doi.org/10.1186/s40561-024-00295-9>
- Meniado, J. C., Huyen, D. T. T., Panyadilokpong, N., & Lertkomolwit, P. (2024). Using ChatGPT for second language writing: Experiences and perceptions of EFL learners in Thailand and Vietnam. *Computers and Education: Artificial Intelligence*, 7, 100313. <https://doi.org/10.1016/j.caeai.2024.100313>
- Meyer, J. G., Urbanowicz, R. J., Martin, P. C. N., O'Connor, K., Li, R., Peng, P.-C., Bright, T. J., Tatonetti, N., Won, K. J., Gonzalez-Hernandez, G., & Moore, J. H. (2023). ChatGPT and large language models in academia: Opportunities and challenges. *BioData Mining*, 16, 20. <https://doi.org/10.1186/s13040-023-00339-9>
- Miranty, D., & Widiati, U. (2021). An automated writing evaluation (AWE) in higher education. *Pegeg Journal of Education and Instruction*, 11(4), 126–137. <https://doi.org/10.47750/pegegog.11.04.12>
- Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2), 100050. <https://doi.org/10.1016/j.rmal.2023.100050>
- Montenegro-Rueda, M., Fernández-Cerero, J., Fernández-Batanero, J. M., & López-Meneses, E. (2023). Impact of the implementation of ChatGPT in education: A systematic review. *Computers*, 12(8), 153. <https://doi.org/10.3390/computers12080153>
- Nguyen, A., Hong, Y., Dang, B., & Huang, X. (2024). Human-AI collaboration patterns in AI-assisted academic writing. *Studies in Higher Education*, 49(5), 847–864. <https://doi.org/10.1080/03075079.2024.2323593>

Nunes, A., et al. (2022). Effectiveness of automated writing evaluation systems in school settings: A systematic review of studies from 2000 to 2020. *Journal of Computer Assisted Learning*, 38(2), 599–620. <https://doi.org/10.1111/jcal.12635>

Ramesh, D., & Sanampudi, S. K. (2021). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, 55, 2495–2527. <https://doi.org/10.1007/s10462-021-10068-2>

Ranalli, J. (2021). L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. *Journal of Second Language Writing*, 52, 100816. <https://doi.org/10.1016/j.jslw.2021.100816>

Sari, E., & Han, T. (2024). The impact of automated writing evaluation on English as a foreign language learners' writing self-efficacy, self-regulation, anxiety, and performance. *Journal of Computer Assisted Learning*, 40(5), 2065–2080. <https://doi.org/10.1111/jcal.13004>

Seßler, K., Fürstenberg, M., Bühler, B., & Kasneci, E. (2025). Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring. *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 462–472. <https://doi.org/10.1145/3706468.3706527>

Siekmann, L., Parr, J. M., & Busse, V. (2022). Structure and coherence as challenges in composition: A study of assessing less proficient EFL writers' text quality. *Assessing Writing*, 54, 100672. <https://doi.org/10.1016/j.asw.2022.100672>

Tang, X., Chen, H., Lin, D., Li, K., et al. (2024). Harnessing LLMs for multi-dimensional writing assessment: Reliability and alignment with human judgments. *Heliyon*, 10(14), e34262. <https://doi.org/10.1016/j.heliyon.2024.e34262>

Weng, X., Xia, Q., Gu, M., Rajaram, K., & Chiu, T. K. F. (2024). Assessment and learning outcomes for generative AI in higher education: A scoping review on current research status and trends. *Australasian Journal of Educational Technology*, 40(6), 37–55. <https://doi.org/10.14742/ajet.9540>

Xia, Q., Weng, X., Ouyang, F., Lin, T. J., & Chiu, T. K. F. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education.

International Journal of Educational Technology in Higher Education, 21, 40.

<https://doi.org/10.1186/s41239-024-00468-z>

Xiao, C., Ma, W., Song, Q., Xu, S. X., Zhang, K., Wang, Y., & Fu, Q. (2025). Human-AI collaborative essay scoring: A dual-process framework with LLMs. *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 293–305.

<https://doi.org/10.1145/3706468.3706507>

Xu, W., Mahmud, R., & Hoo, W. L. (2024). A systematic literature review: Are automated essay scoring systems competent in real-life education scenarios? *IEEE Access*, 12, 77639–77657. <https://doi.org/10.1109/ACCESS.2024.3399163>

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1), 90–112. <https://doi.org/10.1111/bjet.13370>

Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology*, 56, 150–166.

<https://doi.org/10.1111/bjet.13494>

van Niekerk, J., Delport, P. M. J., & Sutherland, I. (2025). Addressing the use of generative AI in academic writing. *Computers and Education: Artificial Intelligence*, 8, 100342. <https://doi.org/10.1016/j.caeai.2024.100342>

Hao, J., von Davier, A. A., Yaneva, V., Lottridge, S., von Davier, M., & Harris, D. J. (2024). Transforming assessment: The impacts and implications of large language models and generative AI. *Educational Measurement: Issues and Practice*, 43(2), 16–29.

<https://doi.org/10.1111/emip.12602>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Appendices.docx](#)
- [SupplementaryMaterials.docx](#)