Research Square

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

Beyond Automated Scoring: An Evidence-Informed Human–GenAI Complementarity Framework for Formative Writing Assessment

Herford Rei Biscayno Guibangguibang

herfordguibang-guibang@jrmsu.edu.ph

Jose Rizal Memorial State University https://orcid.org/0000-0001-9950-1689

Kian C. Hesula

Jose Rizal Memorial State University https://orcid.org/0009-0001-5509-2390

Research Article

Keywords: formative writing assessment, generative artificial intelligence, human–artificial intelligence complementarity, formative feedback, English as a Second Language writing, higher education

Posted Date: August 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10784458/v1

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

Additional Declarations: The authors declare no competing interests.

Abstract

Generative artificial intelligence (GenAI) is increasingly used in formative writing assessment, yet research has focused largely on scoring accuracy and feedback quality rather than on how GenAI and teachers may contribute distinct assessment functions. This study examined teacher–GenAI rating consistency, agreement, and practical equivalence, as well as differences in their formative feedback characteristics. Using a convergent mixed-methods design, 100 ESL student essays were independently assessed by three teacher-raters and three GenAI systems using a common analytic writing rubric. Quantitative data were analyzed using descriptive statistics, intraclass correlation coefficients, Wilcoxon signed-rank tests, and Two One-Sided Tests (TOST) with a ± 0.30 equivalence margin. Formative feedback was examined through reflexive thematic analysis. Results showed good inter-rater reliability within teacher (ICC = .815) and GenAI (ICC = .879) groups but limited teacher–GenAI agreement (ICC = .353). No statistically significant differences were found between teacher- and GenAI-generated ratings across the assessed dimensions. Equivalence testing indicated practical equivalence for organization, language use, mechanics, and overall ratings, but not content. Qualitative findings showed that teachers emphasized language accuracy, organization, and specific corrective guidance, whereas GenAI provided broader, multidimensional feedback. Integrating these findings, the study develops an Evidence-Informed Human–GenAI Complementarity Framework for Formative Writing Assessment, positioning GenAI as a support for, rather than a replacement for, teacher judgment.

1. Introduction

Writing development is an iterative process in which formative assessment guides feedback, instruction, and revision. Effective writing assessment therefore requires timely and pedagogically meaningful feedback grounded in teachers' professional judgment (Van Der Steen et al., 2023). However, providing such feedback consistently across multiple dimensions of writing can place substantial demands on teachers, creating a need for technologies that enhance assessment efficiency without displacing pedagogical expertise.

Recent advances in generative artificial intelligence (GenAI), particularly large language models (LLMs), have expanded formative writing assessment through rubric-based evaluation, contextualized feedback, and adaptive learning support. Unlike traditional automated essay scoring systems that primarily assign scores, GenAI can explain evaluations, generate personalized feedback, and support iterative revision (Pan, 2025; Teng, 2026). Its potential to improve the timeliness, consistency, accessibility, and scalability of writing feedback has consequently attracted growing research attention (Ouhayou, 2026).

Evidence suggests that GenAI can produce consistent ratings and detailed formative feedback that may support writing revision (Pan, 2025; Teng, 2026). Yet formative assessment involves more than producing scores or feedback; it requires interpreting evidence of learning, identifying salient needs, and providing contextually responsive guidance (Sohrabi et al., 2022; Yu & Xu, 2024). Teachers draw on disciplinary knowledge, classroom experience, and contextual understanding when making these judgments (Lee,

2011; Xu & Brown, 2016). Thus, GenAI should be evaluated not only by how closely it reproduces teacher judgments but also by how its capabilities may complement them.

This distinction matters because rating agreement does not necessarily imply assessment equivalence or functional interchangeability. Two assessment agents may assign similar ratings while attending to different features of writing, prioritizing different problems, or providing different forms of formative guidance. Conversely, differences in assessment may reflect distinct capabilities rather than simple deficiencies. Examining both convergence and divergence can therefore provide a more informative basis for understanding the respective roles of teachers and GenAI in formative writing assessment.

Despite the rapid growth of research on GenAI-supported writing assessment, existing work has largely emphasized scoring accuracy, feedback quality, consistency, usability, and instructional efficiency. Studies have also commonly compared GenAI ratings or feedback with those of human teachers or evaluated AI-generated feedback independently. Less attention has been given to what patterns of agreement and divergence reveal about the distinct functions of human and GenAI assessment within formative writing practice.

Human–GenAI complementarity offers a basis for addressing this gap by conceptualizing teachers and GenAI as contributing distinct yet potentially complementary capabilities. GenAI offers efficiency, consistency, scalability, and immediate, multidimensional feedback, whereas teachers contribute contextual judgment, pedagogical reasoning, ethical decision-making, and responsiveness to learners' needs (Edwards et al., 2025). From this perspective, effective human–GenAI assessment does not require identical judgments; complementarity may instead emerge when their different assessment functions are understood and intentionally coordinated.

The present study examines this possibility through a mixed-methods analysis of 100 ESL student essays. The quantitative component examines teacher- and GenAI-generated ratings across content, organization, language use, and mechanics, while the qualitative component examines similarities and differences in their formative feedback. Integrating these strands allows the study to distinguish among within-group consistency, teacher–GenAI agreement, and functional differences in feedback.

The study contributes an evidence-informed account of human–GenAI complementarity in formative writing assessment. Rather than treating GenAI as a substitute for teacher judgment, it examines how convergence and divergence in ratings and feedback can reveal complementary assessment functions. These findings provide the empirical basis for an Evidence-Informed Human–GenAI Complementarity Framework for Formative Writing Assessment.

2. Literature Review

2.1 Formative Assessment as a Learning-Oriented Process

Formative assessment uses evidence of student learning to inform instructional decisions and improve subsequent learning rather than merely evaluate achievement (Lee, 2011; Van Der Steen et al., 2023). Grounded in Assessment for Learning, it enables teachers and learners to monitor progress, identify needs, and adapt instruction. Its value therefore lies not simply in producing assessment evidence but in transforming that evidence into actions that support learning.

Writing is particularly suited to formative assessment because development occurs through iterative cycles of planning, drafting, feedback, revision, and reflection (Burner, 2016). Formative writing assessment supports these cycles by identifying areas for improvement and directing subsequent revision. Its effectiveness therefore depends not only on accurate evaluation but also on feedback that is timely, actionable, and relevant to learners' needs.

Providing such feedback, however, requires substantial teacher expertise and time, particularly in contexts involving large classes and diverse writing needs. This challenge has increased interest in technologies that can improve the efficiency, consistency, and accessibility of formative feedback (Alemdag & Narciss, 2025; Van Schalkwyk et al., 2026). GenAI extends this possibility by enabling rapid, individualized feedback across multiple aspects of writing (Pan, 2025; Teng, 2026). The central issue is therefore how such technology can contribute to, rather than simply automate, a learning-oriented assessment process.

2.2 Feedback Literacy as the Mechanism of Learning

Feedback becomes formative when learners can interpret and use it to improve subsequent performance (Carless & Boud, 2018; Xie & Liu, 2024). Feedback literacy refers to learners' capacity to understand and act on feedback, making learner engagement central to formative assessment (Carless & Boud, 2018; Weidlich et al., 2025). Thus, the value of assessment feedback depends not only on what assessors provide but also on how that information supports learners' decisions about revision and learning.

For feedback to support these processes, it must be timely, actionable, and responsive to learners' needs (Alemdag & Narciss, 2025). GenAI can provide immediate and individualized responses at scale (Pan, 2025; Teng, 2026), potentially expanding access to feedback beyond conventional teacher-mediated cycles. However, the availability of feedback does not guarantee its pedagogical usefulness. Comparing the focus and instructional character of AI-generated and teacher feedback is therefore important for determining how each may contribute to formative writing assessment.

2.3 Teacher Judgment and Writing Assessment Literacy

Writing assessment literacy refers to teachers' capacity to use assessment to support learning rather than merely evaluate performance (Crusan et al., 2016; DeLuca et al., 2016). In formative writing assessment, this involves more than applying scoring criteria; it requires interpreting learners' needs and providing feedback responsive to instructional and contextual conditions. Effective assessment consequently depends on contextual judgment as well as technical accuracy.

Teachers draw on disciplinary knowledge, classroom experience, and knowledge of learners when interpreting student writing and determining appropriate feedback (Lee, 2011; Xu & Brown, 2016). Their

judgments involve decisions about which strengths or problems deserve attention, how feedback should be communicated, and what revisions are appropriate for particular learners. Formative feedback is therefore a pedagogical process rather than merely a technical scoring task.

GenAI introduces different assessment capabilities. It can identify textual patterns, apply explicit criteria consistently, and generate feedback rapidly across multiple dimensions of writing (Pan, 2025; Teng, 2026). However, its outputs do not necessarily incorporate the contextual awareness, ethical reasoning, and pedagogical judgment that teachers bring to assessment (Edwards et al., 2025; Liang et al., 2026). This difference suggests that GenAI should be examined not only according to whether it reproduces teacher judgments but also according to how its distinct capabilities may contribute alongside them.

2.4 Generative Artificial Intelligence in Formative Writing Assessment

GenAI has expanded technology-supported writing assessment by enabling systems to interpret textual inputs, apply assessment criteria, and generate natural-language feedback. Unlike earlier automated writing evaluation systems that primarily relied on predefined rules and scoring mechanisms, LLM-based systems can respond across multiple dimensions of writing, including content, organization, language use, and rhetorical effectiveness (Pan, 2025; Teng, 2026). This development shifts technology-enhanced assessment from automated scoring toward interactive formative support.

Research indicates that GenAI can generate timely, personalized, and multidimensional feedback that may support writing revision and e (Xie & Liu, 2024; Zhan & Yan, 2025).Its ability to process multiple aspects of writing within a single response makes it potentially useful for broad diagnostic coverage. However, consistency in applying criteria does not necessarily mean that GenAI and teachers attend to the same features or assign the same instructional priorities.

This limitation complicates comparison-based approaches to AI assessment. If GenAI is evaluated primarily by its similarity to teacher scores, differences in evaluative focus may be treated as errors rather than evidence of distinct assessment functions. Conversely, evaluating AI feedback independently overlooks how it might complement teachers' contextualized judgments. A more informative approach is therefore to examine where human and GenAI assessments converge, where they diverge, and what these patterns imply for formative assessment.

2.5 Human–GenAI Complementarity in Formative Writing Assessment

Human–GenAI complementarity conceptualizes teachers and GenAI as contributing distinct capabilities that can be coordinated within formative assessment. GenAI can provide efficiency, consistency, scalability, and immediate multidimensional feedback, while teachers contribute contextual interpretation, pedagogical reasoning, ethical decision-making, and responsiveness to learners' needs (Edwards et al., 2025). Complementarity therefore does not require identical outputs; it depends on whether distinct capabilities can be coordinated toward a shared instructional purpose.

Existing research has increasingly compared GenAI ratings or feedback with those of teachers. However, agreement alone provides limited insight into how the two sources of assessment evidence function within formative practice. Similarly, differences in ratings or feedback do not, by themselves, establish that AI assessment is inferior or that teacher judgment is always preferable. What remains insufficiently examined is how patterns of consistency, agreement, and functional divergence can be integrated to explain the respective contributions of human and GenAI assessment.

The present study addresses this issue by examining teacher- and GenAI-generated ratings and formative feedback across 100 ESL student essays. By integrating quantitative evidence of rating consistency and agreement with qualitative evidence of feedback characteristics, the study moves beyond asking whether GenAI can reproduce teacher assessment. It instead examines whether differences between human and GenAI assessment reveal complementary functions that can inform formative writing assessment.

2.6 Research Gap and Research Questions

The literature establishes that GenAI can expand the efficiency, consistency, and breadth of formative writing feedback, while teachers contribute contextual and pedagogical judgment. However, existing research has largely treated teacher assessment as a benchmark against which AI performance is compared or has evaluated AI-generated feedback independently. These approaches establish whether AI outputs are similar to or different from human outputs but provide limited explanation of what those patterns mean for the organization of formative assessment.

Three related gaps emerge. First, rating similarity does not necessarily establish assessment equivalence or functional interchangeability. Second, rating comparisons alone provide limited insight into differences in the focus and instructional character of teacher- and GenAI-generated feedback. Third, although human–AI collaboration has gained theoretical attention, empirical evidence remains limited on how actual patterns of convergence and divergence can be translated into complementary assessment functions.

These gaps call for an integrated examination of three assessment dimensions: consistency, agreement, and complementarity. Consistency concerns whether assessors within an assessment group apply the rubric similarly. Agreement concerns the extent to which teacher and GenAI ratings correspond. Complementarity concerns what the convergence and divergence of ratings and feedback reveal about the distinct functions each assessment agent can contribute.

Accordingly, this study uses a mixed-methods analysis of 100 ESL student essays to examine teacher- and GenAI-generated ratings and formative feedback. The quantitative strand establishes patterns of rating consistency and agreement across four writing dimensions, while the qualitative strand examines differences in feedback focus and function. Their integration provides the empirical basis for developing an Evidence-Informed Human–GenAI Complementarity Framework for Formative Writing Assessment.

The study therefore addresses the following research questions:

RQ1. To what extent do teacher- and GenAI-generated ratings agree and differ across the dimensions of formative writing assessment?

RQ2. What functional similarities and differences characterize teacher- and GenAI-generated formative feedback across content, organization, language use, and mechanics?

RQ3. How do the quantitative and qualitative findings jointly define complementary assessment functions for teachers and GenAI in formative writing assessment?

3. Methodology

3.1 Research Design

This study employed a convergent mixed-methods design (Creswell & Clark, 2017) to examine teacher- and GenAI-generated writing assessments. Quantitative and qualitative data were derived from the same corpus of 100 student essays and analyzed separately before integration. The quantitative strand examined rating patterns and agreement across four rubric dimensions, while the qualitative strand examined similarities and differences in formative feedback.

The convergent design was appropriate because the study addresses both assessment outcomes and assessment functions. Quantitative analysis establishes the extent to which teacher and GenAI ratings converge or diverge, while qualitative analysis examines how their feedback differs in focus and pedagogical function. Integrating the two strands enables the study to move beyond score comparison toward an evidence-based account of complementary assessment functions.

The study was guided by the Human–GenAI Complementarity perspective presented in Fig. 1. Rather than treating teachers and GenAI as competing assessors, the study examines their respective assessment patterns to determine where they converge, where they diverge, and what these patterns reveal about their potential complementary roles in formative writing assessment.

3.2 Participants and Context

The study was conducted in a higher education setting involving undergraduate students enrolled in a Technical Writing course. The analytical corpus comprised 100 anonymized student essays produced through authentic classroom writing activities. The essays represented varying levels of writing performance and served as the common assessment materials for the human and GenAI assessment conditions.

Three experienced English language teachers served as human raters. The teacher-raters held graduate qualifications in English or English Language Teaching and had at least five years of tertiary-level experience in teaching and assessing academic writing. Before the main assessment, they participated in a rubric orientation to establish a shared understanding of the assessment criteria and their application.

Three GenAI systems—ChatGPT, Microsoft Copilot, and CiCi AI—were included as AI assessment conditions. Each system received the same student essays, analytic writing rubric, and standardized assessment instructions. Standardizing the assessment conditions minimized procedural variation across systems and enabled direct comparison with the teacher-generated assessments.

3.3 Instruments

Three instruments supported data collection: (a) an analytic writing rubric, (b) the student essay corpus, and (c) standardized GenAI assessment prompts. Together, these instruments established a common assessment framework for comparing teacher- and GenAI-generated ratings and formative feedback.

3.3.1 Analytic Writing Rubric

Student essays were evaluated using the analytic writing rubric developed by Carroll and West (1989). The rubric assesses four dimensions of writing performance: content, organization, language use, and mechanics, with each dimension rated on a four-point scale from 1 (*Beginning*) to 4 (*Advanced*). Its defined performance descriptors enabled consistent application across teacher and GenAI assessments.

Before the main assessment, teacher-raters participated in a rubric orientation to establish consistent interpretation of the performance descriptors. The same rubric was incorporated into the standardized GenAI prompts, ensuring that both assessment conditions operated against the same stated criteria.

3.3.2 Student Essay Corpus

The primary dataset consisted of 100 anonymized undergraduate student essays from a Technical Writing course. The essays represented varying levels of writing performance and served as the common assessment corpus for both quantitative rating analysis and qualitative feedback analysis.

3.3.3 Standardized GenAI Assessment Prompts

Standardized prompts were developed for ChatGPT, Microsoft Copilot, and CiCi AI. Each system received the same essay, analytic writing rubric, and assessment instructions. The prompts required each system to assign rubric-based scores for content, organization, language use, and mechanics and to generate formative feedback explaining the ratings and recommending revisions. Standardization reduced procedural variation and strengthened comparability across GenAI systems.

Table 1

Research Instruments and Their Functions

Instrument	Description	Purpose
Analytic Writing Rubric	Carroll and West (1989) rubric assessing content, organization, language use, and mechanics using a four-point scale	Provide common assessment criteria for teacher- and GenAI-generated ratings
Student Essay Corpus	100 anonymized essays produced by undergraduate students enrolled in a Technical Writing course	Provide the common corpus for quantitative rating and qualitative feedback analyses
Standardized GenAI Assessment Prompts	Uniform prompts containing the essays, analytic rubric, and assessment instructions administered to ChatGPT, Microsoft Copilot, and CiCi AI	Standardize GenAI-generated ratings and formative feedback across platforms

3.4 Data Collection Procedures

Data collection followed a standardized sequence designed to ensure comparability across human and GenAI assessment conditions.

Phase 1: Essay Preparation.

The 100 student essays were collected from authentic Technical Writing activities and anonymized using identification codes before assessment.

Phase 2: Teacher Assessment.

The three teacher-raters independently evaluated the essays using the analytic writing rubric. Each rater assigned scores for content, organization, language use, and mechanics and provided written formative feedback.

Phase 3: GenAI Assessment.

The same essays were independently evaluated by ChatGPT, Microsoft Copilot, and CiCi AI using the standardized prompts and identical analytic writing rubric. Each system generated rubric-based ratings and corresponding formative feedback.

Phase 4: Data Preparation.

Teacher- and GenAI-generated ratings were encoded for quantitative analysis. The corresponding written feedback was compiled and organized for qualitative analysis.

Phase 5: Data Integration.

The quantitative and qualitative findings were analyzed independently and subsequently integrated during interpretation to examine how rating patterns and feedback characteristics jointly inform Human–GenAI

complementarity.

3.5 Data Analysis

The quantitative and qualitative datasets were analyzed separately and integrated during interpretation. Quantitative analysis examined rating distributions, within-group consistency, teacher–GenAI agreement, and rating differences, while qualitative analysis examined similarities and differences in the characteristics and functions of formative feedback. The findings were subsequently integrated to identify complementary assessment functions.

3.5.1 Quantitative Data Analysis

To address RQ1, quantitative analysis examined teacher- and GenAI-generated ratings across the four rubric dimensions: content, organization, language use, and mechanics. Means, standard deviations, medians, and interquartile ranges were calculated to summarize rating distributions.

Inter-rater consistency and agreement were assessed using the Intraclass Correlation Coefficient (ICC) with 95% confidence intervals. A two-way mixed-effects, absolute-agreement, mean-rating model ($k = 3$) was used to examine agreement among the three teacher-raters, among the three GenAI systems, and between teacher- and GenAI-generated ratings. ICC values were interpreted following Koo and Li (2016): values below .50 indicate poor agreement, .50–.75 moderate, .75–.90 good, and above .90 excellent agreement.

Teacher–GenAI rating differences were examined using the Wilcoxon signed-rank test because ratings were paired across the same essays and rubric dimensions. Statistical significance was evaluated at $\alpha = .05$. Because a nonsignificant difference test does not establish equivalence, practical equivalence was examined separately using the Two One-Sided Tests (TOST) procedure.

Because a nonsignificant difference test does not establish equivalence, equivalence testing was conducted using the Two One-Sided Tests (TOST) procedure. An equivalence margin of ± 0.30 rubric points was specified as the smallest practically meaningful difference on the four-point analytic writing rubric. Teacher–GenAI ratings were considered practically equivalent when the confidence interval for the paired mean difference fell entirely within the − 0.30 to + 0.30 equivalence interval (Lakens, 2017).

Quantitative findings were reported by rubric dimension and assessment comparison to distinguish rating consistency, teacher–GenAI agreement, statistical differences, effect magnitude, and practical equivalence. This approach allowed the study to determine whether similarities in ratings represented meaningful agreement or merely the absence of a statistically detectable difference.

All quantitative analyses were conducted using IBM SPSS Statistics Version 29.0, with TOST equivalence analyses conducted using an appropriate equivalence-testing procedure.

3.5.2 Qualitative Data Analysis

To address RQ2, teacher- and GenAI-generated formative feedback was analyzed using reflexive thematic analysis following Braun and Clarke (2006, 2021). The analysis involved familiarization with the feedback corpus, generation of initial codes, development and refinement of candidate themes, definition and naming of themes, and development of the analytical narrative.

Teacher and GenAI feedback were first examined separately and then compared to identify similarities and differences in focus, specificity, instructional emphasis, and formative function. Themes were developed inductively through iterative engagement with the feedback data and interpreted in relation to the study's focus on complementary assessment functions.

3.5.3 Integration of Quantitative and Qualitative Findings

To address RQ3, quantitative and qualitative findings were integrated through a joint interpretation of rating patterns and feedback characteristics. Convergent and divergent findings were examined together to determine whether differences in assessment outcomes and feedback reflected distinct assessment functions rather than simply differences in assessment quality.

The integrated findings provided the empirical basis for identifying complementary functions of teachers and GenAI and for developing the Evidence-Informed Human–GenAI Complementarity Framework for Formative Writing Assessment.

3.6 Ethical Considerations

The study was approved by the appropriate university authorities, and informed consent was obtained from all participants. Participation was voluntary, and students could withdraw at any time without academic consequences. All essays were anonymized using identification codes and securely stored for research purposes only. GenAI systems evaluated only anonymized essays using standardized prompts. No personally identifiable information was entered into any GenAI platform, and AI-generated outputs were used solely for research and did not contribute to students' academic grades.

4. Results

This section presents the quantitative and qualitative findings in relation to the three research questions. The quantitative results first examine rating patterns, within-group consistency, and teacher–GenAI agreement across the four writing dimensions. The qualitative results then describe differences in teacher- and GenAI-generated formative feedback, followed by an integrated interpretation of the two strands to identify complementary assessment functions.

4.1 Teacher–GenAI Rating Agreement and Differences

This subsection addresses RQ1 by examining teacher- and GenAI-generated ratings across content, organization, language use, and mechanics. The analysis first considers rating patterns and consistency within each assessment group before examining teacher–GenAI agreement and differences.

4.1.1 Descriptive Statistics of Teacher- and GenAI-Generated Ratings

The three teacher-raters assigned an overall mean rating of 2.84 (SD = 0.79), while the three GenAI systems produced an overall mean of 2.71 (SD = 0.75). Content received the highest mean ratings for both teachers (M = 3.07, SD = 0.73) and GenAI (M = 2.78, SD = 0.80), while Mechanics received the lowest ratings from both groups.

Table 2

Descriptive Statistics of Teacher- and GenAI-Generated Ratings Across the Analytic Writing Rubric

Rubric Dimension	Teacher 1 M (SD)	Teacher 2 M (SD)	Teacher 3 M (SD)	Teacher Overall M (SD)	ChatGPT M (SD)	Copilot M (SD)	CiCi M (SD)	GenAI Overall M (SD)
Content	3.10 (0.64)	2.90 (0.85)	3.20 (0.69)	**3.07 (0.73)**	3.25 (0.63)	3.00 (0.64)	2.10 (0.64)	**2.78 (0.80)**
Organization	3.40 (0.68)	2.20 (0.69)	2.90 (0.96)	**2.83 (0.92)**	3.15 (0.58)	3.05 (0.60)	2.10 (0.64)	**2.77 (0.76)**
Language Use	3.10 (0.44)	2.40 (0.68)	2.70 (0.65)	**2.73 (0.66)**	3.10 (0.55)	2.80 (0.61)	2.10 (0.64)	**2.67 (0.72)**
Mechanics	2.75 (0.63)	2.30 (0.73)	3.10 (0.85)	**2.72 (0.80)**	3.15 (0.58)	2.60 (0.50)	2.10 (0.64)	**2.62 (0.71)**
Overall	**3.09 (0.60)**	**2.45 (0.74)**	**2.98 (0.79)**	**2.84 (0.79)**	**3.16 (0.59)**	**2.86 (0.59)**	**2.10 (0.64)**	**2.71 (0.75)**

Note. M = Mean; SD = Standard Deviation.

Among the teacher-raters, Teacher 1 assigned the highest overall mean (M = 3.09, SD = 0.60), whereas Teacher 2 assigned the lowest (M = 2.45, SD = 0.74). Among the GenAI systems, ChatGPT produced the highest overall mean (M = 3.16, SD = 0.59), followed by Microsoft Copilot (M = 2.86, SD = 0.59), while CiCi AI produced the lowest (M = 2.10, SD = 0.64). Despite these differences in absolute ratings, both groups showed a similar pattern across dimensions, with Content receiving the highest ratings and Mechanics the lowest.

4.1.2 Inter-Rater Reliability Within Assessment Groups

Having established the rating patterns, this subsection examines the consistency of judgments within the teacher and GenAI groups. The three teacher-raters demonstrated good overall inter-rater reliability (ICC = .815), while the three GenAI systems also demonstrated good overall reliability (ICC = .879).

Table 3

Inter-Rater Reliability of Teacher-Raters and GenAI Systems Across the Analytic Writing Rubric

Rubric Dimension	Teacher-Raters ICC	Interpretation	GenAI Systems ICC	Interpretation
Content	.911	Excellent	.897	Good
Organization	.865	Good	.888	Good
Language Use	.765	Good	.876	Good
Mechanics	.715	Moderate	.855	Good
Overall	**.815**	**Good**	**.879**	**Good**

Note. ICC = Intraclass Correlation Coefficient. Reliability estimates were computed using a mean-rating (k = 3), absolute-agreement, two-way mixed-effects model. Interpretation follows Koo and Li (2016): poor < .50; moderate .50–.75; good .75–.90; excellent > .90.

Among the teacher-raters, Content demonstrated the highest agreement (ICC = .911), followed by Organization (ICC = .865) and Language Use (ICC = .765), while Mechanics showed moderate reliability (ICC = .715). The GenAI systems demonstrated good reliability across all four dimensions, with ICCs ranging from .855 to .897.

These findings indicate that both assessment groups applied the common rubric with substantial internal consistency. However, high reliability within each group does not necessarily indicate agreement between teachers and GenAI.

4.1.3 Agreement Between Teacher- and GenAI-Generated Ratings

Having established consistency within each assessment group, this subsection examines the extent to which teacher- and GenAI-generated ratings correspond across the rubric dimensions.

The overall agreement between the two groups was poor (ICC = .353). Content demonstrated the highest teacher–GenAI agreement (ICC = .440; fair), followed by Language Use (ICC = .379; poor) and Organization (ICC = .339; poor). Mechanics showed the lowest agreement (ICC = .185; poor).

Table 4

Agreement Between Teacher- and GenAI-Generated Ratings Across the Analytic Writing Rubric

Rubric Dimension	ICC	Interpretation
Content	.440	Fair
Organization	.339	Poor
Language Use	.379	Poor
Mechanics	.185	Poor
Overall Score	**.353**	**Poor**

Note. ICC = Intraclass Correlation Coefficient.

The results reveal a clear distinction between within-group consistency and between-group agreement. Teachers and GenAI each produced relatively consistent ratings, but their ratings showed limited correspondence across the four dimensions. Mechanics exhibited the greatest divergence, whereas Content showed comparatively stronger, though still limited, agreement.

4.1.4 Differences and Equivalence Between Teacher- and GenAI-Generated Ratings

Beyond agreement, the analysis examined whether teacher- and GenAI-generated ratings differed systematically using the Wilcoxon signed-rank test. The results showed no statistically significant differences between the two assessment groups across Content ($Z = -0.85$, $p = .309$), Organization ($Z = -0.87$, $p = .312$), Language Use ($Z = -1.07$, $p = .253$), or Mechanics ($Z = -1.38$, $p = .132$). The overall rating comparison was likewise nonsignificant ($Z = -1.16$, $p = .246$).

Table 5

Wilcoxon Signed-Rank Test Comparing Teacher- and GenAI-Generated Ratings

Rubric Dimension	Z	*p*	Decision
Content	−0.85	.309	Not Significant
Organization	−0.87	.312	Not Significant
Language Use	−1.07	.253	Not Significant
Mechanics	−1.38	.132	Not Significant
Overall Score	−1.16	.246	Not Significant

Note. The Wilcoxon signed-rank test compared paired teacher- and GenAI-generated ratings. Statistical significance was evaluated at α = .05.

Because a nonsignificant difference test does not establish practical equivalence, a Two One-Sided Tests (TOST) procedure was additionally used with a predefined equivalence margin of ± 0.30 rubric points. The equivalence analysis indicated that the observed differences for Organization, Language Use, Mechanics, and Overall Score fell within the predefined equivalence interval, whereas the difference for Content did not meet the equivalence criterion.

Table 6
TOST Equivalence Analysis of Teacher- and GenAI-Generated Ratings

Rubric Dimension	Mean Difference	90% CI	Equivalence Margin	TOST Result
Content	0.29	[0.12, 0.46]	± 0.30	Not equivalent
Organization	0.06	[− 0.09, 0.21]	± 0.30	Equivalent
Language Use	0.06	[− 0.08, 0.20]	± 0.30	Equivalent
Mechanics	0.10	[− 0.05, 0.25]	± 0.30	Equivalent
Overall Score	0.13	[− 0.02, 0.28]	± 0.30	Equivalent

4.2 Characteristics of Teacher- and GenAI-Generated Formative Feedback

This subsection addresses RQ2 by describing the similarities and differences in formative feedback generated by the teacher-raters and GenAI systems across content, organization, language use, and mechanics. The analysis focused on the focus, coverage, and specificity of the feedback produced by each assessment source.

The teacher-raters' feedback primarily addressed language accuracy, organization, and specific areas requiring correction. Comments identified particular weaknesses in student writing and, in several instances, provided direct suggestions for revision.The GenAI systems produced feedback addressing multiple dimensions of writing, including content, organization, language use, and mechanics. Their responses commonly included observations about strengths and weaknesses as well as suggestions for revision.

Table 7

Representative Remarks and Initial Coding Categories of Teacher- and GenAI-Generated Formative Feedback

Subject	Remarks	Manner
Teacher 1	"Lacks concluding statement."	Criticism
	"Subject agreement is erratic."	
Teacher 2	"Error in spelling ang capitalization."	
	"Your punctuations are not appropriate."	
Teacher 3	"It does not have a strong thesis statement."	
	"Faulty parallelism."	
Subject	**Remarks**	**Manner**
Teacher 1	*"To improve their teaching [insert comma]…"*	Suggestive (Specific)
	"students table… [add apostrophe]."	
Teacher 2	*"It make…[makes instead of make.]"*	
	"ChatGPT are one of the… [verb tense-'is' to be used]	
Teacher 3	*"Your going to…[you're instead of your]…"*	
	"spacing in the 2nd and 3rd paragraphs…"	
Subject	**Remarks**	**Manner**
ChatGPT	"*Clear and strong points but there is a minor language issue.*"	Praise (specific) + Criticism
	"Good insights but minor grammar issues."	
CiCi	"*The essay provides relevant points about AI in education but lacks depth in analysis."*	
	"The content is informative and detailed, with good organization. However, some grammatical and mechanical issues slightly detract from clarity."	
Subject	**Remarks**	**Manner**
ChatGPT	*"Insufficient Content; Vague arguments."*	Criticism
	"Off-topic; lacks educational focus'	
Microsoft Copilot	*"Needs more depth; lacks specific examples"*	
	"Very weak; one-sided argument; lacks details"	

The feedback from both assessment sources contained comments related to areas for improvement and recommendations for revision. However, differences were observed in the distribution and emphasis of the feedback. Teacher-rater comments were more frequently directed toward specific concerns, whereas GenAI responses covered a broader range of writing features within individual feedback outputs.

Across the four rubric dimensions, both assessment sources generated feedback related to **c**ontent, organization, language use, and mechanics. The frequency and emphasis of these categories varied between teacher- and GenAI-generated feedback.

Overall, the qualitative results show that both teacher-raters and GenAI systems generated formative feedback addressing multiple aspects of student writing, with observable differences in the focus, coverage, and specificity of the feedback.

4.2.1 Themes in Teacher-Generated Formative Feedback

Three themes emerged from the analysis of formative feedback provided by the teacher-raters. Figure 2 presents the coding network showing the primary categories identified across the three teacher-raters.

Theme 1. Prioritization of Language Accuracy and Organization

The teacher-raters' feedback primarily focused on language accuracy and organization. Comments addressed grammatical correctness, punctuation, capitalization, sentence construction, and the organization of ideas. Representative remarks included *“Lacks concluding statement,” “Subject agreement is erratic,” “Error in spelling and capitalization,” “Your punctuations are not appropriate,” “It does not have a strong thesis statement,”* and *“Faulty parallelism.”*

Theme 2. Criticism Accompanied by Specific Suggestions

The feedback frequently identified writing deficiencies and included specific suggestions for revision. Examples included *“Insert comma,” “students table… [add apostrophe],” “It make… [makes instead of make],” “ChatGPT are one of the… [verb tense—‘is’ to be used],” “Your going to… [you're instead of your],”* and *“spacing in the 2nd and 3rd paragraphs.”* These comments combined the identification of errors with explicit revision guidance.

Theme 3. Variation in the Breadth of Feedback

The breadth of feedback varied across the teacher-raters. Teachers 1 and 3 addressed both language accuracy and organization, whereas Teacher 2 focused primarily on organization. The coding network also included instances categorized as No Remarks, indicating that some essays received little or no written feedback from particular raters.

Overall, the teacher-generated feedback was characterized by three observable features: emphasis on language accuracy and organization, criticism accompanied by specific revision suggestions, and variation in the breadth of written feedback across teacher-raters.

4.2.2 Themes in GenAI-Generated Formative Feedback

Three themes emerged from the analysis of formative feedback generated by ChatGPT, CiCi, and Microsoft Copilot. Figure 3 presents the coding network showing the primary coding categories identified across the three GenAI systems.

Theme 1. Comprehensive Coverage of Writing Dimensions

The GenAI systems consistently addressed multiple aspects of student writing. Across ChatGPT, CiCi, and Copilot, the coded areas included content, organization, language in general, and language accuracy. Representative remarks included *"The essay provides relevant points about AI in education but lacks depth in analysis"* and *"The content is informative and detailed, with good organization. However, some grammatical and mechanical issues slightly detract from clarity."*

Theme 2. Specific Praise Combined with Criticism

The GenAI-generated feedback frequently combined specific praise with criticism. Positive comments identifying strengths were followed by observations about areas requiring improvement. Representative remarks included *"Clear and strong points but there is a minor language issue"* and *"Good insights but minor grammar issues."*

Theme 3. Critical Evaluation of Writing Weaknesses

The analysis also identified feedback focused primarily on writing weaknesses. Representative comments included *"Insufficient content; vague arguments," "Off-topic; lacks educational focus," "Needs more depth; lacks specific examples,"* and *"Very weak; one-sided argument; lacks details."* These comments focused on deficiencies in content development and argumentation.

Overall, the GenAI-generated formative feedback was characterized by three observable features: comprehensive coverage of writing dimensions, the combination of specific praise with criticism, and critical evaluation of writing weaknesses.

4.2.3 Comparative Summary of Teacher- and GenAI-Generated Formative Feedback

This subsection summarizes the similarities and differences identified in the formative feedback generated by the teacher-raters and the three GenAI systems. The comparison considers the focus, manner, scope, breadth, and representative characteristics of the feedback.

Table 8

Comparative Summary of Teacher- and GenAI-Generated Formative Feedback

Category	Teacher-Generated Feedback	GenAI-Generated Feedback
Focus of feedback	Emphasized organization, language accuracy, and specific writing concerns	Addressed content, organization, language use, and language accuracy
Manner of feedback	Criticism accompanied by specific suggestions	Specific praise combined with criticism; criticism
Scope of feedback	Focused primarily on linguistic and organizational concerns	Addressed multiple writing dimensions simultaneously
Breadth of feedback	Varied across teacher-raters, including instances of no written remarks	Consistently covered multiple writing dimensions across the three GenAI systems
Representative characteristics	Explicit correction of language and mechanics	Broader evaluation integrating strengths and areas for improvement

The comparison shows that both teacher-raters and GenAI systems generated formative feedback addressing student writing. Teacher-generated feedback included comments on organization, language accuracy, and specific writing concerns, while GenAI-generated feedback addressed content, organization, language use, and language accuracy.

Differences were also observed in the manner and scope of the feedback. Teacher feedback included specific corrective suggestions, whereas GenAI feedback included both specific praise and criticism. Teacher feedback was primarily focused on linguistic and organizational concerns, while GenAI feedback addressed multiple writing dimensions within individual responses.

The breadth of feedback also varied across the two assessment sources. Teacher-rater feedback varied across raters, including instances in which no written remarks were provided. In contrast, the three GenAI systems consistently produced feedback covering multiple writing dimensions.

Overall, the comparative analysis identified differences between teacher- and GenAI-generated feedback in focus, manner, scope, breadth, and representative characteristics, while both assessment sources produced feedback intended to address aspects of student writing

5. Discussion

This study examined teacher–GenAI rating consistency and agreement, compared the characteristics of teacher- and GenAI-generated formative feedback, and examined how these findings inform a complementary approach to formative writing assessment. The findings show that teachers and GenAI systems produced relatively consistent ratings within their respective groups, but demonstrated limited agreement with one another. At the same time, their formative feedback differed in focus and breadth.

These findings provide a basis for understanding human and GenAI assessment as related but distinct contributions to formative writing assessment.

5.1 Rating Consistency Does Not Necessarily Imply Interchangeability

Both teacher-raters and GenAI systems demonstrated good inter-rater reliability, with ICCs of .815 and .879, respectively. This finding is consistent with research showing that large language models can apply analytic writing rubrics with substantial consistency under standardized assessment conditions (Mizumoto & Eguchi, 2023).

However, the overall teacher–GenAI agreement was substantially lower (ICC = .353), despite the absence of statistically significant differences in their ratings. This distinction is important because statistical similarity in rating levels does not necessarily indicate agreement in assessment judgment. Automated scoring research has similarly highlighted that rubric-based scores may conceal differences in the processes or criteria underlying those scores (Kumar & Boulanger, 2021; Mizumoto & Eguchi, 2023).

The finding suggests that teacher and GenAI ratings should not be treated as interchangeable simply because their overall rating tendencies are statistically similar. Teachers evaluate student writing within pedagogical and instructional contexts, whereas GenAI systems generate judgments through patterns learned from large-scale language data. Formative assessment, however, involves more than assigning scores; it also involves interpreting evidence of learning and determining what feedback is most useful for subsequent improvement (Black & Wiliam, 2009; Sadler, 1989).

The distinction between consistency, agreement, and interchangeability therefore provides an important qualification to research that evaluates GenAI primarily by its proximity to human scores. Rather than asking only whether GenAI can reproduce teacher ratings, the findings support examining where the two assessment sources converge, where they diverge, and how those differences may be used within a broader formative assessment process.

5.2 Distinct Feedback Characteristics and Complementary Assessment Functions

The qualitative findings showed that teacher- and GenAI-generated feedback differed in focus, breadth, and manner. Teachers emphasized language accuracy, organization, and specific corrective guidance, whereas GenAI systems addressed multiple writing dimensions and frequently combined praise with criticism.

The teacher emphasis on targeted correction is consistent with the established role of formative feedback in identifying learning needs and guiding subsequent revision (Black & Wiliam, 2009; Sadler, 1989). Effective feedback requires not only identifying weaknesses but also providing information that learners can use to improve their work. Research on teacher feedback in L2 writing likewise emphasizes the importance of selecting and prioritizing feedback according to instructional purposes (Lee, 2011).

GenAI's broader coverage is consistent with emerging evidence that large language models can generate multidimensional feedback across aspects of writing and provide such feedback rapidly (Kasneci et al., 2023). Recent work on GenAI-supported formative assessment further indicates the potential of rubric-indexed feedback to support writing development and provide learners with additional opportunities for revision (Pan, 2025).

The findings therefore suggest that the difference between teacher and GenAI feedback is not adequately described as a difference in quality alone. Rather, the two sources displayed different patterns of attention. Teachers more frequently focused on selected concerns requiring correction, whereas GenAI addressed a wider range of writing features within individual responses. This distinction provides a basis for considering how the two forms of feedback may be coordinated rather than evaluated solely in terms of which more closely resembles the other.

5.3 Refining the Proposed Conceptual Framework: An Evidence-Informed Human–GenAI Complementarity Framework

The integration of the quantitative and qualitative findings refines the proposed conceptual framework into an Evidence-Informed Human–GenAI Complementarity Framework for Formative Writing Assessment. The original framework was grounded in formative assessment, teacher expertise, and GenAI affordances; the empirical findings clarify how these elements can be connected through distinct assessment functions.

The findings indicate that teachers contribute pedagogical judgment, contextual interpretation, and targeted guidance, while GenAI contributes consistent rubric-based assessment, multidimensional feedback, and immediacy. These contributions should not be understood as fixed or exclusive capabilities, but as functions that can be distributed across the assessment process.

The framework therefore shifts the focus from technological equivalence to functional complementarity. Existing discussions of AI-supported assessment have often centered on whether automated systems can approximate or replace human judgment (Kasneci et al., 2023; Kumar & Boulanger, 2021). A complementary perspective instead positions GenAI within a human-in-the-loop process in which teachers retain responsibility for contextual interpretation and pedagogical decision-making (Holmes et al., 2022; Luckin, 2018).

This perspective is also consistent with broader research on human–AI collaboration, which emphasizes designing AI systems to support rather than displace human decision-making (Edwards et al., 2025). In formative writing assessment, such collaboration can involve GenAI generating broad initial feedback while teachers review, contextualize, prioritize, and refine that feedback in relation to learners' needs.

The framework consequently conceptualizes formative assessment as a distributed process rather than a competition between human and artificial assessors. GenAI can extend the breadth and immediacy of feedback, while teachers provide contextual interpretation and instructional judgment. The value of the

combination therefore lies not in making the two assessment sources interchangeable, but in coordinating their distinct contributions toward revision and learning.

6. Conclusion

This study examined teacher- and GenAI-generated ratings and formative feedback in writing assessment, with particular attention to their consistency, agreement, and feedback characteristics. The quantitative findings showed good inter-rater reliability within both teacher and GenAI groups, but limited agreement between the two assessment sources. The qualitative findings showed differences in the focus, coverage, and specificity of their formative feedback: teacher-raters emphasized language accuracy, organization, and explicit corrective guidance, whereas GenAI systems addressed broader and more multidimensional aspects of writing.

Taken together, the findings indicate that teacher- and GenAI-generated assessment evidence contains both areas of convergence and divergence. The study therefore advances an Evidence-Informed Human–GenAI Complementarity Framework for Formative Writing Assessment, which conceptualizes teachers and GenAI as distinct but interconnected contributors to formative assessment rather than interchangeable sources of judgment.

The framework shifts the focus from whether GenAI can reproduce teacher assessment toward how its capabilities can be integrated with teacher expertise. Within this perspective, teachers retain responsibility for contextual interpretation, pedagogical judgment, and instructional decision-making, while GenAI can provide consistent rubric-based assessment and broad, timely feedback. This human-in-the-loop orientation positions GenAI as an instructional partner that augments rather than replaces teacher judgment.

The principal contribution of the study is therefore the evidence-informed framework, which provides a basis for understanding and designing human–GenAI collaboration in formative writing assessment. Rather than treating GenAI performance primarily in terms of replacement or score equivalence, the framework emphasizes the coordinated use of distinct assessment capabilities to support learning-oriented writing assessment.

As GenAI continues to evolve, future research should move beyond replacement-oriented comparisons and examine how complementary human–GenAI assessment practices can be designed, implemented, and evaluated across diverse educational contexts. Such work should also examine whether the proposed framework produces measurable changes in students' writing revision, feedback literacy, self-regulated learning, and writing achievement.

7. Limitations and Future Research

This study has three principal limitations. First, it was conducted within a single higher education ESL writing context, which limits the transferability of the findings. Second, the study examined one writing

task, one analytic rubric, and three GenAI systems, limiting the applicability of the findings across other assessment designs and evolving AI models. Third, although the study identified patterns of teacher–GenAI complementarity through comparative analysis, it did not evaluate the proposed framework through actual classroom implementation.

Future research should therefore validate the framework across diverse educational, disciplinary, linguistic, and cultural contexts, as well as across different writing tasks, assessment designs, and GenAI systems. Longitudinal, experimental, and design-based studies should further examine how human–GenAI assessment practices influence writing revision, feedback literacy, self-regulated learning, and writing achievement.

References

1. Alemdag E, Narciss S (2025) Promoting formative self-assessment through peer assessment: Peer work quality matters for writing performance and internal feedback generation. Int J Educational Technol High Educ 22(1):22. https://doi.org/10.1186/s41239-025-00522-4
2. Black P, Wiliam D (2009) Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5
3. Braun V, Clarke V (2006) Using thematic analysis in psychology. Qualitative Res Psychol 3(2):77–101. https://doi.org/10.1191/1478088706qp063oa
4. Braun V, Clarke V (2021) One size fits all? What counts as quality practice in (reflexive) thematic analysis? Qualitative Res Psychol 18(3):328–352. https://doi.org/10.1080/14780887.2020.1769238
5. Burner T (2016) Formative assessment of writing in English as a foreign language. Scandinavian J Educational Res 60(6):626–648. https://doi.org/10.1080/00313831.2015.1066430
6. Carless D, Boud D (2018) The development of student feedback literacy: Enabling uptake of feedback. Assess Evaluation High Educ 43(8):1315–1325. https://doi.org/10.1080/02602938.2018.1463354
7. Creswell JW, Clark VLP (2017) Designing and Conducting Mixed Methods Research. SAGE. https://books.google.com.ph/books?id=BXEzDwAAQBAJ
8. Crusan D, Plakans L, Gebril A (2016) Writing assessment literacy: Surveying second language teachers' knowledge, beliefs, and practices. Assess Writ 28:43–56. https://doi.org/10.1016/j.asw.2016.03.001
9. DeLuca C, LaPointe-McEwan D, Luhanga U (2016) Teacher assessment literacy: A review of international standards and measures. Educational Assess Evaluation Account 28(3):251–272. https://doi.org/10.1007/s11092-015-9233-6
10. Edwards J, Nguyen A, Lämsä J, Sobocinski M, Whitehead R, Dang B, Roberts A, Järvelä S (2025) Human- AI collaboration: Designing artificial agents to facilitate socially shared regulation among learners. Br J Edu Technol 56(2):712–733. https://doi.org/10.1111/bjet.13534
11. Holmes W, Porayska-Pomsta K, Holstein K, Sutherland E, Baker T, Shum SB, Santos OC, Rodrigo MT, Cukurova M, Bittencourt II, Koedinger KR (2022) Ethics of AI in Education: Towards a Community-

Wide Framework. Int J Artif Intell Educ 32(3):504–526. https://doi.org/10.1007/s40593-021-00239-1

12. Kasneci E, Sessler K, Küchemann S, Bannert M, Dementieva D, Fischer F, Gasser U, Groh G, Günnemann S, Hüllermeier E, Krusche S, Kutyniok G, Michaeli T, Nerdel C, Pfeffer J, Poquet O, Sailer M, Schmidt A, Seidel T, Kasneci G (2023) ChatGPT for good? On opportunities and challenges of large language models for education. Learn Individual Differences 103:102274. https://doi.org/10.1016/j.lindif.2023.102274
13. Kumar VS, Boulanger D (2021) Automated Essay Scoring and the Deep Learning Black Box: How Are Rubric Scores Determined? Int J Artif Intell Educ 31(3):538–584. https://doi.org/10.1007/s40593-020-00211-5
14. Lakens D (2017) Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses. Social Psychol Personality Sci 8(4):355–362. https://doi.org/10.1177/1948550617697177
15. Lee I (2011) Working Smarter, Not Working Harder: Revisiting Teacher Feedback in the L2 Writing Classroom. Can Mod Lang Rev 67(3):377–399. https://doi.org/10.3138/cmlr.67.3.377
16. Liang Z, Yang K, Sha L, Gašević D, Yan L, Chen G (2026) A systematic review of generative AI in education: Empirical insights from a human– AI interaction perspective. *British Journal of Educational Technology*, bjet.70055. https://doi.org/10.1111/bjet.70055
17. Luckin R (2018) Machine learning and human intelligence: The future of education for the 21st century. Institute of Education
18. Mizumoto A, Eguchi M (2023) Exploring the potential of using an AI language model for automated essay scoring. Res Methods Appl Linguistics 2(2):100050. https://doi.org/10.1016/j.rmal.2023.100050
19. Ouhayou C (2026) A Comparative Analysis of AI and Human Feedback on L2 Writing: Implications for Formative Feedback Integration. In A. Adoui (Ed.), *Artificial Intelligence, Academic Integrity, and the Internationalization of Higher Education* (pp. 337–361). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-19585-2_16
20. Pan Y (2025) Leveraging generative AI powered rubric-indexed feedback as a formative assessment strategy for enhancing medical English education. Discover Comput 28(1):284. https://doi.org/10.1007/s10791-025-09830-9
21. Sadler DR (1989) Formative assessment and the design of instructional systems. Instr Sci 18(2):119–144. https://doi.org/10.1007/BF00117714
22. Sohrabi Z, Ghanbari N, Abbasi A (2022) Iranian EFL teachers' perceptions of writing assessment literacy: A countrywide study. Lang Test Asia 12(1):23. https://doi.org/10.1186/s40468-022-00172-7
23. Teng MF (2026) Effectiveness of GenAI in enhancing writing performance: A meta-analysis. Asian-Pacific J Second Foreign Lang Educ 11(1):70. https://doi.org/10.1186/s40862-026-00444-2
24. Van Der Steen J, Van Schilt-Mol T, Van Der Vleuten C, Joosten-ten Brinke D (2023) Designing Formative Assessment That Improves Teaching and Learning: What Can Be Learned from the Design Stories of Experienced Teachers? J Formative Des Learn 7(2):182–194. https://doi.org/10.1007/s41686-023-00080-w

25. Van Schalkwyk S, Kuper A, O'Sullivan P (2026) On the receiving end: Feedback literacy for mentees. Adv Health Sci Educ 31(1):7–10. https://doi.org/10.1007/s10459-026-10506-2
26. Weidlich J, Fink A, Frey A, Jivet I, Gombert S, Menzel L, Giorgashvili T, Yau J, Drachsler H (2025) Highly informative feedback using learning analytics: How feedback literacy moderates student perceptions of feedback. Int J Educational Technol High Educ 22(1):43. https://doi.org/10.1186/s41239-025-00539-9
27. Xie Z, Liu W (2024) What matters in the cultivation of student feedback literacy: Exploring university EFL teachers' perceptions and practices. Humanit Social Sci Commun 11(1):162. https://doi.org/10.1057/s41599-024-02648-8
28. Xu Y, Brown GTL (2016) Teacher assessment literacy in practice: A reconceptualization. Teach Teacher Educ 58:149–162. https://doi.org/10.1016/j.tate.2016.05.010
29. Yu C, Xu W (2024) Writing assessment literacy and its impact on the learning of writing: A netnography focusing on Duolingo English Test examinees. Lang Test Asia 14(1):24. https://doi.org/10.1186/s40468-024-00297-x
30. Zhan Y, Yan Z (2025) Students' engagement with ChatGPT feedback: Implications for student feedback literacy in the context of generative artificial intelligence. Assess Evaluation High Educ 1–14. https://doi.org/10.1080/02602938.2025.2471821

Figures

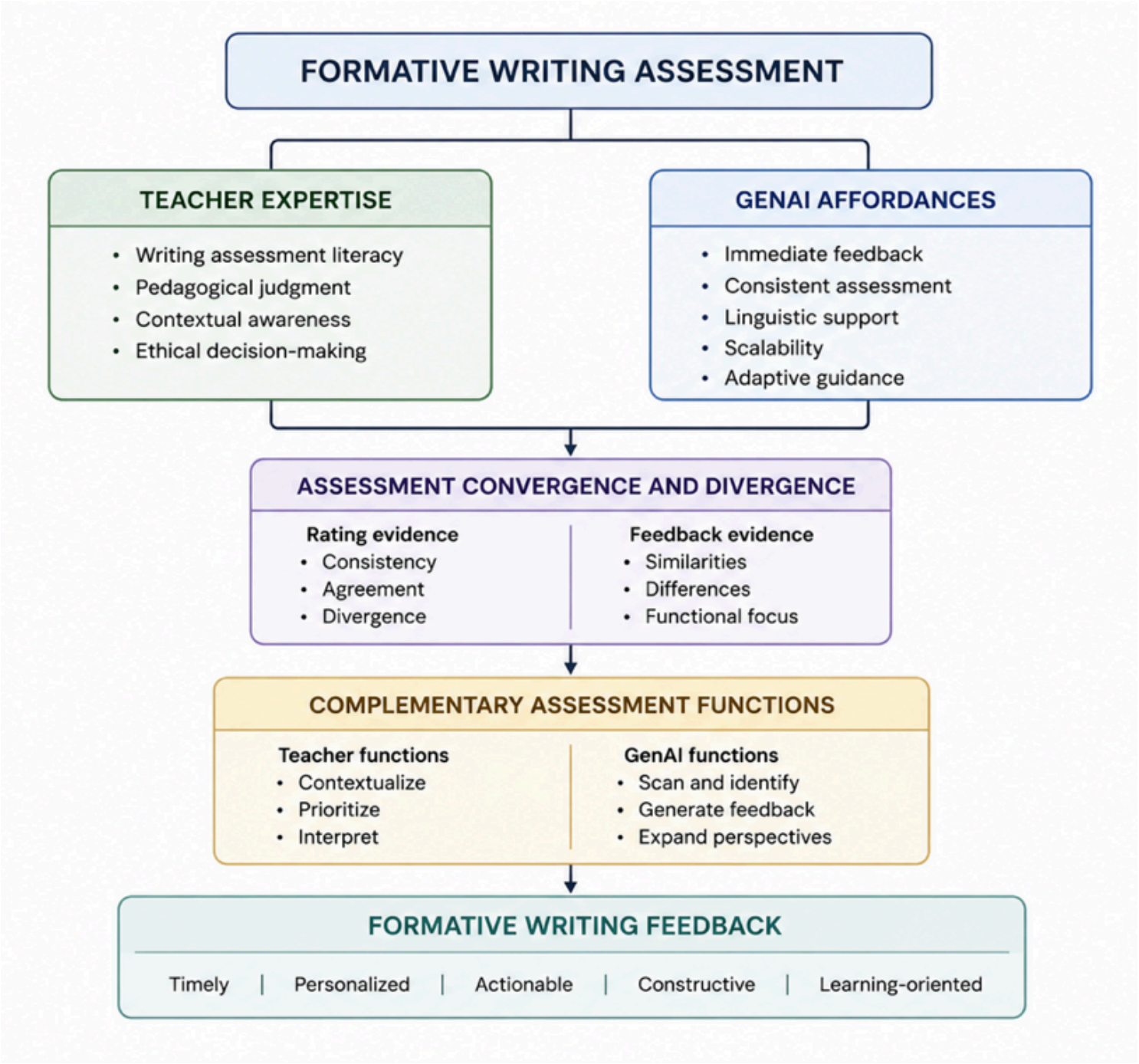


Figure 1

Proposed Conceptual Framework of Human–GenAI Complementarity in Formative Writing Assessment

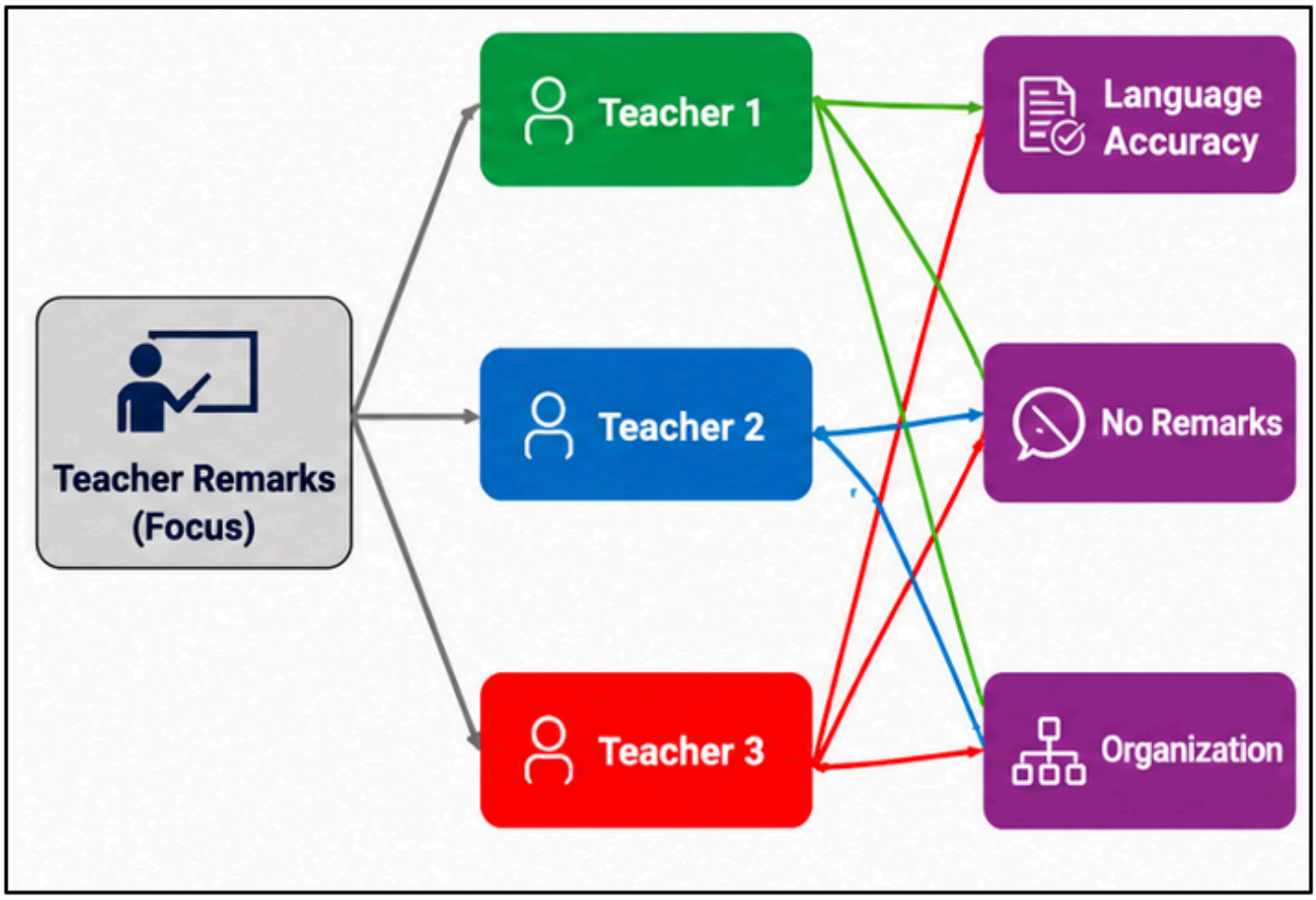


Figure 2

Coding network showing the primary focus of teacher-generated formative feedback

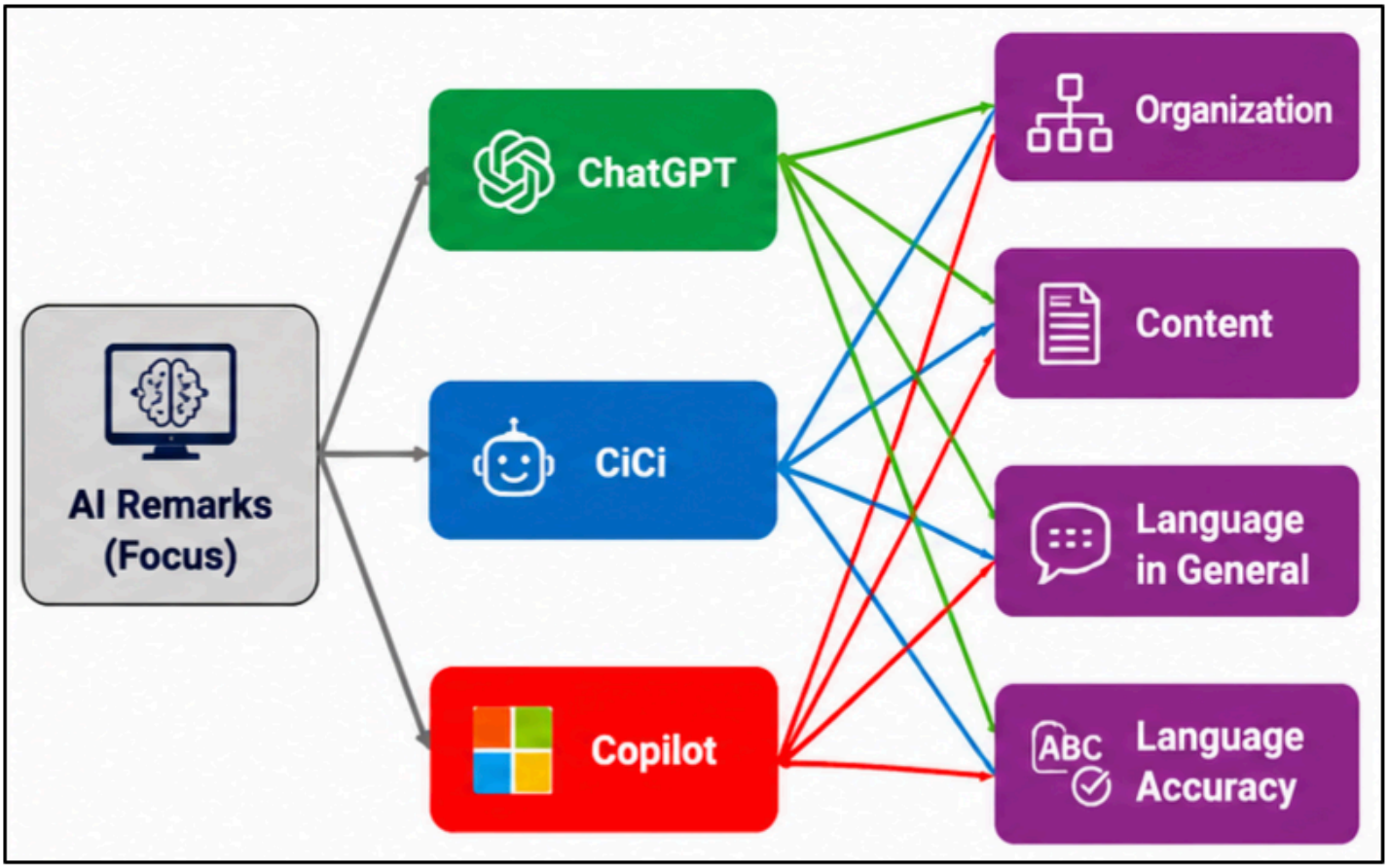


Figure 3

Coding network showing the primary focus of GenAI-generated formative feedback

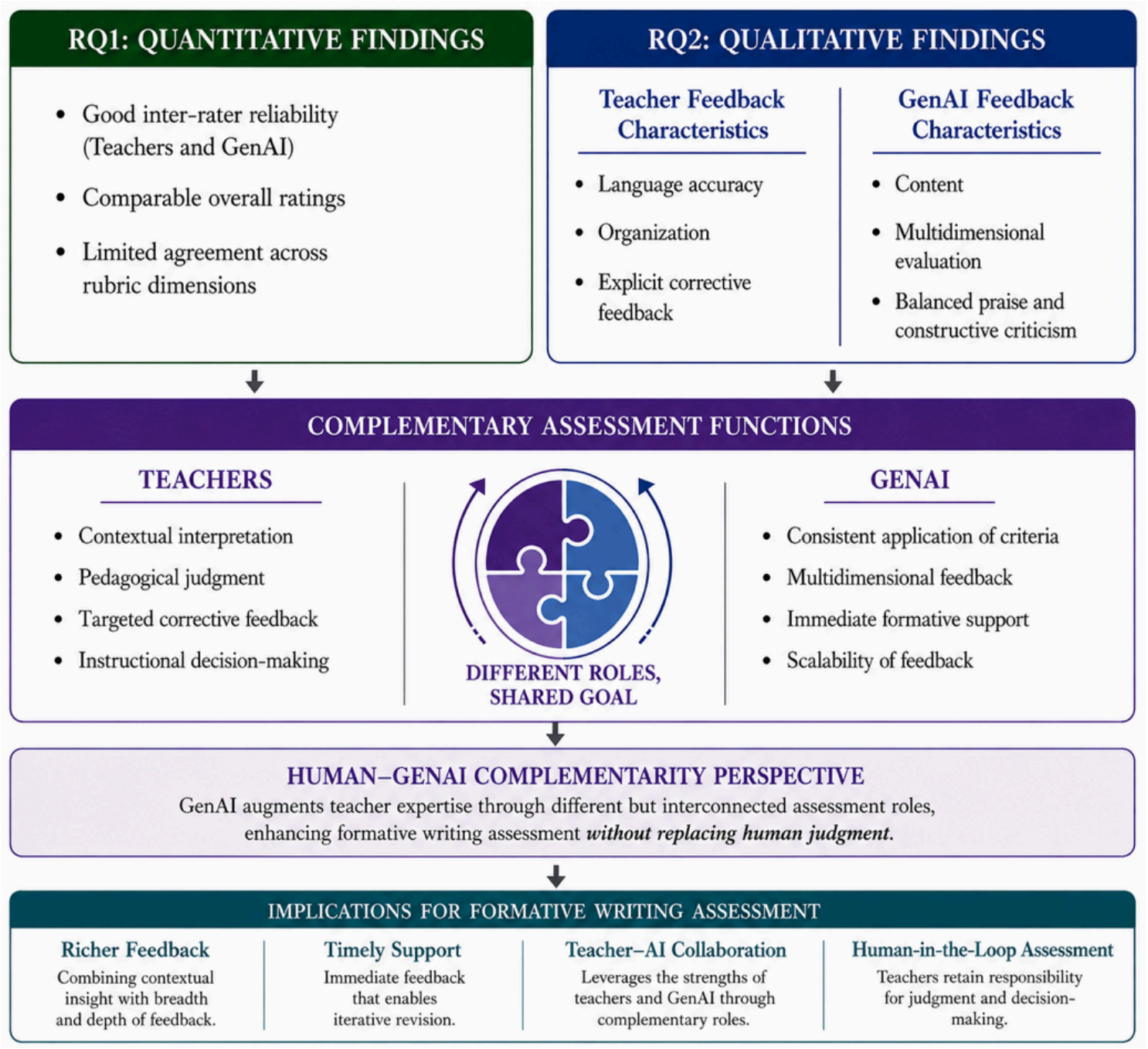


Figure 4

Evidence-Informed Human–GenAI Complementarity Framework for Formative Writing Assessment