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.

When alignment is not agreement: Construct-specific validation of large language models for CAF measurement in L2 writing

Amin Roshanaie
University of Maragheh
Asghar Salimi Amirghayeb
asgharsalimi356@gmail.com

University of Maragheh

Article

Keywords: large language models, L2 writing, writing assessment, CAF measurement, human-AI agreement, construct validation

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11235965/v1

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

Additional Declarations: No competing interests reported.

Abstract

Although current research shows that large language models (LLMs) have become more popular in second-language (L2) writing assessment, we must note that strong correlation with human raters does not imply interchangeable measurement. This study assessed the ability of six public facing LLMs to replicate values for explicit complexity, accuracy and fluency (CAF) obtained from reviewed human consensus coding. There were 90 observations of Iranian EFL undergraduates' argumentative essay, which were produced by 45 students in class. Three measures of product-based CAF measures were used: words per T-unit for Fluency, S-nodes per T-unit for Complexity, and error-free T-units per T-unit for Accuracy. A single coding prompt and anonymized essays were each given to ChatGPT, Gemini, Grok, DeepSeek, Claude and Microsoft Copilot. Validation evidence included correlations, paired mean differences, bias, RMSE, normalized RMSE, Bland–Altman-type limits of agreement and within-writer change diagnostics. For all 18 human-LLM correlations, the results were positive and significant ($p < .001$); however, there was some variability across constructs and models. There was clear systematic underestimation with complexity, and the lowest participant-level change agreement, and accuracy was best reproduced. DeepSeek performed the best overall approximation using the combined criteria, and ChatGPT showed the best accuracy approximation. On tasks, human reference was more accurate but less fluent and complex in the essay two; models reproduced accuracy directions more consistently than the fluency and complexity directions. The results indicate that alignment is not interchangeable and suggests construct-specific validity, transparent prompting, and human verification for L2 writing assessment using LLMs.

1. Introduction

The use of large language models (LLMs) in writing assessment has transformed the context in which learner texts can be analyzed in practice. Generative systems are now able to segment text, detect linguistic features, apply scoring rubrics, and provide structured numerical outputs within seconds. These capabilities are appealing in second-language (L2) writing because detailed analysis is time consuming, and because researchers and teachers are increasingly dealing with larger sets of L2 texts. Meanwhile, the apparent fluency and accuracy of an LLM answer can mask a fundamental question: does the LLM actually replicate the construct-relevant decisions that a trained human analyst would make? This question is especially relevant when an LLM is not used to give general feedback or an overall score, but to generate explicit linguistic measures which can then be used in statistical analyses.

Complexity, accuracy and fluency (CAF) offers a helpful framework to consider that problem. The CAF framework has long been employed to describe partially distinct dimensions of L2 performance instead of reducing learner writing to a single global score (Housen & Kuiken, 2009; Norris & Ortega, 2009; Pallotti, 2009; Phuoc & Barrot, 2022; Wolfe-Quintero et al., 1998). However, the simplicity of the CAF formulas can be deceptive. A ratio like words per T-unit needs a defensible segmentation of T-units, an error-free T-unit ratio needs consistent decisions on grammatical errors, and a clause-based complexity measure needs consistent decisions on the analyst's part as to what is a finite clause, what is a non-

finite clause, what is a co-ordinated clause, what is a sub-ordinated clause, and what is a non-standard learner clause. These choices render CAF analysis as a suitable test case for assessing the ability of LLMs to serve as construct-sensitive linguistic measurement tools, not just as plausible text generators.

The majority of the research in AI-assisted writing assessment has been centered on holistic or rubric-based scoring, automated essay scoring and feedback. Previous automated scoring studies focused on feature engineering, psychometric validation and the difference between reliability and validity (Attali, 2013; Ramineni & Williamson, 2013; Shermis, 2014). Recent LLM research has demonstrated that in certain scenarios, generative models can approximate human scores, and can also be prompt sensitive, score compressed, systematically severe or lenient, and dimensionally disagreeable (Bui & Barrot, 2025; Chen, 2026; Huang et al., 2026; Lee & Hwang, 2026; Wang et al., 2025; Wu et al., 2026). These results suggest that the relationship between human and AI is not consistent and cannot be captured by a single correlation coefficient.

The difference between alignment and agreement is therefore key. Correlation asks the question: do two sources move in unison with each other, that is, do essays with relatively high values from human coders also have high values from a model? Agreement asks the question, do the numerical values fall within a reasonable distance from one another so that they can be understood on the same scale or can be interchanged? A model can have a high correlation with human values and a negative bias, or have a small mean difference with large errors at the essay level. Psychometric research has warned repeatedly against the use of association as equivalence in automated scoring (Powers et al., 2015; Wang et al., 2025). This is particularly true for explicit language measures, where a few choices of segmentation can affect the final ratio without affecting the relative ordering of texts.

The second is construct specificity. LLM-based assessment is frequently talked about as if the performance of the model were an intrinsic attribute of the model. But, various CAF dimensions need different analytic operations. Product-based fluency can be relatively straightforward after words and T-units have been identified. Accuracy relies on a rule-governed decision as to whether or not a T-unit is grammatically error free. In the current study, complexity involves the identification of S-nodes at the clause level as well as the segmentation of T-units. In learner writing, incomplete structures, coordination issues, unclear clause boundaries and non-standard grammar patterns are often found. In such cases, human reviewers can talk about them and come to a consensus; an LLM might use a different implicit parsing rule. Thus, it is reasonable to assume that human–LLM correspondence will differ depending on the construct as well as the model.

The present study addresses these issues in the context of a matched multi-model design. A total of 90 argumentative essays were analyzed by reviewed human consensus coding and six public LLM sources, with the same operational definitions and prompt. The study intentionally does not consider the human values as a gold standard, but as a reference that was reviewed, as the independent pre-consensus ratings were not saved. The validation evidence is similarly limited: the study tests convergence with the documented human reference, not the validity of LLM-based writing assessment. The analysis includes

correlation, mean-level bias, numerical error, normalized error, limits of agreement, model-level synthesis and participant-level change diagnostics. This design enables the central contribution to transcend the question of whether or not LLMs correlate with human judgments to the more impactful question of when they do and when they do not.

1.1. CAF measurement and construct specificity

The basis of CAF research is that L2 performance is multidimensional. The elaboration and sophistication of linguistic resources (complexity), conformity to target-language norms (accuracy), and amount or ease of production (fluency) are all matters of concern depending on modality and tradition of operation (Housen & Kuiken, 2009; Housen et al., 2012). These dimensions can vary in different ways as the learners spend finite attentional resources in conceptualization, formulation, monitoring, and textual production (Skehan, 2009, 2016). Therefore, an overall writing score may mask a writing profile with one dimension showing gains and another dimension showing no change or losses.

There are choices for how to measure within each CAF dimension. Syntactic complexity has been operationalized using length based, subordination based, coordination based and phrasal indices, and no single index reflects the entirety of the construct (Bulté & Housen, 2012; Norris & Ortega, 2009; Ortega, 2003). Multiple indexes can now be analyzed with automated tools, for example Lu's L2 Syntactic Complexity Analyzer (Lu, 2010), which makes the analysis of multiple indexes more scalable, but still counts of constructs need interpretation. Complexity in the present study is knowingly limited to S-node/T-unit. S-node counter for every independent finite clause and every dependent finite (or non-finite) clause identified by its verbal predicate. This is a clause-based definition but to be transparent also puts a high demand on consistent segmentation.

Operationalization also has an impact on accuracy. The error-free-unit ratios are very commonly used since they are a clear and comprehensible proportion of grammatical control, however if one local error occurs then the whole T-unit is counted as being a non-error-free unit. Additionally, rater choices may differ in the treatment of sentence boundaries, lexical choices, punctuation, and malformed L2 structures. The present study considers accuracy only in terms of grammatical errors and does not include spelling, punctuation, capitalization, stylistic awkwardness, coherence, cohesion, or lexical choice, unless they cause a clear problem in grammar. The definition of fluency used here is words per T-unit, which is a measure of productivity rather than of the time taken or pauses in writing. The narrow operational definitions are essential because the study does not test for the theoretical range of each of the constructs within the CAF, but rather if the models apply the same rules.

1.2. LLMs as Writing evaluators and Linguistic analyzers.

The broad writing-assessment traditions that focus on construct definition and interpretation of scores include a longer history of automated essay scoring (AES) in which LLM-based writing assessment fits in place. In traditional AES systems, validation was primarily done by assessing human-score correspondence, score reliability, fairness and construct coverage (Ramesh & Sanampudi, 2022;

Ramineni & Williamson, 2013). LLMs alter this architecture as they can process a natural-language instruction, analyse the text, explain the decisions and return a score or feature set without the need for a task-specific feature pipeline. This flexibility is helpful but it also moves some of the measurement tool to the prompt and the model configuration.

Empirical studies have yielded conflicting but gradually more coherent evidence. Mizumoto & Eguchi (2023) showed the potential of using an AI language model for automated essay scoring, and Barrot (2023) showed opportunities and risks for L2 writing. The hybrid approach of LLM-based AES has also incorporated model-based representations along with linguistic elements to enhance the performance of scoring (Atkinson & Palma, 2025). Evaluations of the quality of ChatGPT or other LLMs against human raters have found that while there are significant relationships, there is also variation between rounds, criteria, and prompt designs (Bui & Barrot, 2025; Kim et al., 2024; Tang et al., 2024; Yavuz et al., 2025). More recent studies have focused on reliability, bias, fairness, and criterion-specific performance, rather than just the correlation of scores (Chen, 2026; Han & Zheng, 2026; Huang et al., 2026; Lee & Hwang, 2026; Wang et al., 2025; Wu et al., 2026). Prompt specification is included in the reproducibility record, and prompt engineering can change the stability of the scoring and the human alignment with it (An & Lee, 2025).

Another, albeit less extensive, area of research focuses on particular language features rather than overall scores. Recent studies on L2 writing (Carlo & Takeuchi, 2026) have been examining whether LLM representations can capture characteristics of L2 writing proficiency; Jiang et al. (2023) examined L2 writing accuracy at T-unit and sentence levels with the aid of AI. This transition is significant from a methodological point of view. If the desired result is the number of T-units, clauses, errors or words, then a fluent explanation will not replace accurate measurement. The model should reflect the operational decision sequence that is dependent on the final value.

1.3. From alignment to agreement

Pearson or Spearman correlations are often reported in human–LLM studies as they offer an intuitive measure of correspondence. But there is a possibility of a high correlation and systematic bias. Rank ordering may not change much when a fixed value is subtracted from human complexity values across the board by a model, even though the measurements are not interchangeable. On the other hand, a small average difference may hide large positive and negative errors that average out to a small difference. Complementary evidence of association, bias, and dispersion is needed for validation, however.

Several agreement oriented indices are used in the present analysis to tackle these differences. Bias is the average of the paired difference between the LLM and human value. RMSE provides an overall indication of the size of the numerical differences and places more emphasis on the larger differences. The three dimensions of the CAF operate on different scales, so RMSE are divided by the human range of observations of the construct, or normalized RMSE (nRMSE). Bland-Altman type limits of agreements (Bland & Altman, 1986) are approximate limits that contain most paired differences. These measures are

not designed to provide universal thresholds of interchangeability, but to make apparent the forms in which disagreements occur that correlation does not detect.

Practical implications of construct-specific validation. A model good for accuracy screening of the students' product in the preliminary stage might not be suitable for syntactic-complexity measuring. Likewise, the model that has the lowest average means might not have the best essay level alignment. The question is therefore not just which LLM is 'best' but what is an acceptable model–construct combination for the purpose of the intended analysis. This view provides the motivation to carry out the multi-indicator comparison below.

1.4. Study purpose and research questions

Given this backdrop, the current study explored the possibility of six public LLM models as reproducible tools for explicit CAF measurement in L2 argumentative writing. Instead of correlation being used as evidence of validity, association, systematic error, numerical error and construct-specific correspondence are checked against a reviewed human consensus reference. The analysis was guided by three research questions:

RQ1. How well do LLM's agree with the human consensus reference on product-based fluency, syntactic complexity and accuracy?

RQ2. What are the differences between human-model agreement by model by CAF construct and which model gives the best overall approximation to the human reference with respect to the combined criteria of the study?

RQ3. How well do the models reflect the CAF change profile found by the human reference within the writer and between the two writing occasions?

RQ3 is not interpreted as a test of language development as it is secondary validation evidence as the same topic is repeated without an instructional intervention.

2. Methods

2.1. Design, setting, and participants

The study adopted a descriptive quantitative design with a comparative and correlational analysis. For fluency, complexity, and accuracy, each essay observation included one human consensus value and a corresponding value from each LLM. Thus, human-LLM analyzes were matched at the level of essay-observation. Each learner also produced two essays, permitting secondary within-participant comparisons across writing occasions.

Participants were Iranian EFL undergraduate students from the English Department of a university in Iran. Eighty students completed an online placement screening based on KET, consisting of grammar

and vocabulary items similar to KET. The instrument was used for sample homogeneity and was not an official Cambridge qualification. SIXTY-FIVE students achieved the desired intermediate to upper intermediate criterion. Forty-five students completed both writing occasions. The final analytic sample consisted of 45 learners (aged 18–26) and 90 essay observations. Convenience sampling was used. Participation was voluntary. Demographic variables were recorded in a descriptive manner and were not analyzed as predictors.

2.2. Writing task and dataset

Each participant wrote two argumentative essays in classroom conditions with a two-week interval. Argumentative writing is a long-standing area of L2 writing research (Amini Farsani et al., 2025). On both occasions the same prompt was used: 'Do social media platforms help or harm real-life communication skills?' Between Essay One and Essay Two, no corrective feedback, instructional treatment, or explicit guidance on revision was provided. Using the same topic reduced variation in difficulty of the topic, but also introduced the possibility of familiarity or practice effects. Nevertheless, occasion differences are interpreted as changes in observed performance under similar task conditions, rather than as evidence of instruction-driven development (Li & Yu, 2024; Wu et al., 2023).

Raw counts and calculated CAF values were both in the dataset. Raw counts included total words, total T-units, error free T-units and total S-nodes. All essay records were anonymized and aggregated by participant and occasion before human and LLM analysis.

2.3. CAF measures and human reference coding

Table 1
Operationalization of the three CAF measures

Measure	Operationalization	Key coding decision
Product-based fluency	Total words / total T-units	T-unit segmentation; word count includes misspellings and contractions as one word.
Syntactic complexity	Total S-nodes / total T-units	One S-node per independent finite clause and each dependent finite or non-finite clause identified through its verbal predicate.
Accuracy	Error-free T-units / total T-units	A T-unit is error-free only when it contains no grammatical error; spelling, punctuation, capitalization, style, coherence, and lexical choice are excluded unless they create a clear grammatical problem.

Note. The operational definitions were applied to both the human consensus coding and the LLM extraction. "Fluency" is a product-based productivity measure and should not be interpreted as temporal writing speed.

The three CAF measures are summarized in Table 1. The analysis was based on segmentation and classification decisions and therefore only the study-specific rules needed for replication are listed here. Independent clauses linked by coordinating conjunctions were counted as separate T-units if each had

its own subject and finite verb. S-nodes were counted at the level of clauses, including dependent finite and non-finite clauses. Auxiliaries and modals within a single verbal construction were not counted as separate S-nodes. Accuracy coding was restricted to grammatical errors. The same rules were predetermined and applied for both human coding and the standardized LLM prompt.

Human reference values were established by a consensus procedure that was reviewed. The first author coded all essays for words, T-units, error-free T-units, and S-nodes. The coded data set was checked by the second author and an MA level reviewer. Ambiguous cases were resolved by discussion. The reference for all human–LLM comparisons was the average consensus values. Since independent pre-consensus ratings were not retained, it is not possible to estimate an ICC retrospectively and these values are referred to as a human consensus reference rather than a reliability-validated gold standard.

2.4. LLM sources, prompt standardization, and extraction

The anonymized essay texts were submitted to six public-facing LLM sources: ChatGPT (GPT 5.5), Gemini 3.1 Pro, Grok 4.3, DeepSeek V4, Claude Sonnet 4.6 High and Microsoft Copilot Smart. The labels correspond to the model or mode that each public interface exhibited during the data-collection period and not to immutable backend endpoints. The research archive held contemporaneous screenshots, and a brief model/mode log is provided in Supplementary Appendix B.

For each essay-model pair, we ran one standardized extraction run using the same prompt for all six systems. The prompt did not ask each model to rewrite, correct, summarize, evaluate, score, or improve the essay; it only asked for the raw counts and the calculated CAF values. The full prompt including study-specific coding rules and fixed output format is reproduced in Supplementary Appendix A. The reproducibility checklist is provided in Supplementary Appendix C. LLM-based scoring and coding may be prompt sensitive (An & Lee, 2025; Lee & Hwang, 2026; Tang et al., 2024). Thus, the prompt should be considered as part of the measurement procedure.

Outputs were checked for missing or impossible values. error-free T-units could not be greater than total T-units, total T-units had to be greater than zero, and accuracy values had to be between zero and one. If the model produced both explanatory prose and numerical output, only the numerical values that were requested were entered into the analytic dataset. One run per essay and model provided procedural symmetry but no evidence of repeated-run stability.

2.5. Data analysis

2.5.1. RQ1: Human-model alignment and numerical agreement

To answer RQ1, we first calculated descriptive statistics for all sources and all CAF dimensions. To explore human-model alignment, we computed Pearson correlations of matched essay observations. Because correlation is a measure of association rather than numerical equivalence, paired-samples t tests were also used to test systematic mean differences between each model and the human

reference. Cohen's dz and 95% confidence intervals were reported for the paired comparisons. Then, the agreement was evaluated by calculating mean bias, root mean square error (RMSE), normalized RMSE (nRMSE) and Bland–Altman-type limits of agreement (Bland & Altman, 1986) from the paired differences (LLM value - human value).

2.5.2. RQ2: Construct- and model-specific agreement

For RQ2, model performance was assessed by multiple indicators, rather than a single ranking statistic, across constructs. The synthesis considered correlation, mean absolute error, number of non-significant human-model mean differences, nRMSE, and balance of fluency, complexity, and accuracy. In the study, several planned comparisons were involved and converging evidence was used to interpret the results. Findings with $p < .001$ were considered the strongest evidence of sensitivity; larger but conventionally significant p values were interpreted more cautiously.

2.5.3. RQ3: Within-writer change correspondence

To address RQ3, paired-samples tests were used to compare Essay One and Essay Two within each source. Direction agreement at the participant-level was computed as the proportion of learners for which a model and the human reference pointed in the same direction of change. These percentages are descriptive diagnostics, not reliability coefficients corrected for chance. Occasion effects are assumed to reflect changes in observed performance within writers, as no intervention occurred between occasions and they can be considered repetitions of the same topic rather than developmental instructional effects.

3. Results

3.1. Descriptive overview

For the 90 essay observations, the human mean was 16.161 for product-based fluency (SD = 3.107), 3.169 for complexity (SD = 0.557), and.480 for accuracy (SD = .198). LLM fluency means we are relatively close to the human mean, ranging from 15.710 to 16.416. Greater separation emerged for complexity: Grok 4.3 produced a mean of 3.163, nearly identical to the human mean, while Microsoft Copilot Smart produced 2.670. Range was .448 to .533. These descriptive patterns indicated that mean closeness differed across model-construct pairings and needed to be analyzed concurrently with essay-level correspondence.

3.2. RQ1: Human-model alignment and numerical agreement

The human–LLM alignment and agreement results are presented in Table 2.

[Insert Table 2 **here]**

Table 2

Human–LLM alignment and agreement-oriented summary

Model	CAF	r	Bias (p)	nRMSE
ChatGPT	Fluency	.832	−.328 (.117)	.150
ChatGPT	Complexity	.775	−.403 (< .001)	.217
ChatGPT	Accuracy	.900	+ .030 (.003)	.108
Gemini	Fluency	.810	+ .123 (.580)	.158
Gemini	Complexity	.717	−.290 (< .001)	.205
Gemini	Accuracy	.850	−.025 (.065)	.145
Grok	Fluency	.882	+ .254 (.181)	.136
Grok	Complexity	.739	−.005 (.915)	.165
Grok	Accuracy	.735	−.032 (.079)	.194
DeepSeek	Fluency	.918	+ .176 (.236)	.106
DeepSeek	Complexity	.847	−.150 (< .001)	.132
DeepSeek	Accuracy	.867	+ .012 (.290)	.120
Claude	Fluency	.861	−.452 (.029)	.149
Claude	Complexity	.739	−.256 (< .001)	.193
Claude	Accuracy	.789	+ .053 (< .001)	.171
Copilot	Fluency	.648	−.342 (.317)	.243
Copilot	Complexity	.624	−.499 (< .001)	.287
Copilot	Accuracy	.690	+ .053 (.015)	.235

Note. All Pearson correlations were significant at p < .001. Bias = model mean - human mean; the p value in parentheses is from the paired-samples test of the mean difference. nRMSE = normalized root mean square error. Model names are shortened here; full model or mode labels are reported in Section 2.4.

All 18 human-LLMs correlations were positive and statistically significant (p < .001). Correlations for fluency ranged from r = .648 for Microsoft Copilot Smart to r = .918 for DeepSeek V4. Complexity correlations ranged from r = .624 (Microsoft Copilot Smart) to r = .847 (DeepSeek V4). The correlations for accuracy ranged from r = .690 for Microsoft Copilot Smart to r = .900 for ChatGPT. The universal positive alignment thus suggests that all models captured some of the relative ordering present in human coding but the strength of that alignment was model and construct dependent. This construct-specific alignment profile is illustrated in Fig. 1.

Paired mean comparisons told a very different story. Reduced LLM complexity was the most consistent pattern. ChatGPT, Gemini, DeepSeek, Claude and Microsoft Copilot all generated significantly lower complexity means than the human reference (all $p < .001$) whereas Grok did not differ significantly in complexity ($p = .915$). We also found significant mean differences for accuracy of the ChatGPT ($p = .003$), fluency of the Claude ($p = .029$), accuracy of the Claude ($p < .001$), and accuracy of the Microsoft Copilot ($p = .015$). The clearest sign of systematic bias in the sensitivity reading of the study was the repeated differences in complexity.

Agreement-based error indices supported the same conclusion. DeepSeek achieved the lowest nRMSE for fluency (.106) and complexity (.132), whereas ChatGPT achieved the lowest nRMSE for accuracy (.108). Microsoft Copilot had the highest nRMSE in all three dimensions (.243, .287 and .235 respectively). Negative bias in Complexity was also found for five of the six models. Thus the most consistent cross-model calibration problem was underestimation of clause-based structural elaboration. Figure 2 shows the normalized error profile across models and constructs.

3.3. RQ2: Construct- and model-specific agreement

Table 3 summarizes the model-level results across the three CAF dimensions.

[Insert Table 3 **here]**

Table 3
Model-level synthesis across CAF dimensions

Rank	Model	Avg r	Avg MAE / nRMSE	NS diffs	Interpretation
1	DeepSeek	.878	.468 / .119	2	Best overall balance; top F/C alignment
2	ChatGPT	.836	.706 / .158	1	Best accuracy alignment
3	Claude	.796	.687 / .171	0	Strong fluency; bias in all three means
4	Gemini	.792	.734 / .169	2	Strong accuracy; complexity underestimation
5	Grok	.785	.513 / .165	3	Closest mean levels; weaker accuracy
6	Copilot	.654	1.096 / .255	1	Weakest alignment; largest normalized error

Note. Avg r = average Pearson correlation across the three CAF dimensions; Avg MAE = average mean absolute error on the original scales; nRMSE = normalized root mean square error; NS diffs = number of non-significant human-model mean differences. The ranking is based on converging evidence rather than any single statistic.

DeepSeek V4 was the first in the study's combined criteria, with the highest average correlation (r = .878), lowest average MAE in the original ranking, and lowest average nRMSE (.119). ChatGPT was ranked second and had the highest correlation for individual accuracy (r = .900). Grok illustrates the importance of multiple statistics to gage the quality of LLMs. It is the only model with no significant human–LLM mean difference in any CAF dimension but its average correlation is lower than DeepSeek's and its accuracy alignment is comparatively weaker. Microsoft Copilot showed the lowest average alignment and the highest normalized errors.

Thus, the pattern at the model level is less important than the stronger finding at the construct level. Complexity repeatedly led to less human–LLM agreement than the other dimensions. Five models significantly under-estimated complexity. Complexity also produced the weakest agreement in direction of change, as reported in participant-level diagnostics below. The same models could perform well on one dimension and less well on another, suggesting that 'LLM reliability' is not a unitary property independent of the construct being measured.

3.4. RQ3: Within-writer change correspondence

Table 4 presents the Essay Two minus Essay One changes for the human reference and each LLM source.

[Insert Table 4 **here]**

Table 4
Essay Two minus Essay One change by source

Source	Δ Fluency	Δ Complexity	Δ Accuracy
Human	−2.068***	−.304***	+ .236***
ChatGPT (GPT 5.5)	−2.272***	−.391***	+ .230***
Gemini 3.1 Pro	−1.944***	−.164 (p=.080)	+ .202***
Grok 4.3	−1.870***	−.116 (p=.254)	+ .209***
DeepSeek V4	−2.337***	−.268***	+ .224***
Claude Sonnet 4.6 High	−1.953***	−.234 (p=.023)	+ .209***
Microsoft Copilot Smart	−2.951***	−.226 (p=.078)	+ .246***

Note. Positive values indicate an increase from Essay One to Essay Two; negative values indicate a decrease. *** p < .001. Claude complexity was conventionally significant at p = .023 but did not meet the study's stricter p < .001 sensitivity benchmark.

The human reference showed a large drop in fluency from 17.195 to 15.127, a large drop in complexity from 3.320 to 3.017, and a large increase in accuracy from.362 to.598 (all p < .001). All six LLMs replicated the sharp drop in fluency and boost in accuracy. Complexity was less stable: ChatGPT and

DeepSeek showed $p < .001$ decreases, Claude showed a smaller conventional $p = .023$ decrease, and Gemini, Grok, and Microsoft Copilot did not show statistically significant complexity change.

Participant-level agreement of direction provided more local diagnosis. Accuracy agreement ranged from 77.8% to 93.3% where ChatGPT and DeepSeek agreed with the human direction for 42 of 45 participants (93.3%). Fluency agreement varied between 71.1% and 91.1%. Complexity was again the lowest, ranging from 51.1% for Microsoft Copilot to 73.3% for both ChatGPT and DeepSeek. Thus, both the average change profile and the participant level analysis came to the same conclusion: the LLMs were more consistent in replicating accuracy and fluency direction than complexity direction. Figure 3 shows the direction-agreement pattern at the subject level.

These occasion results are not to be seen as evidence of instructional development. And no intervention was done. Same topic was repeated after 2 weeks. The defensible conclusion is more limited: the human reference exhibited an accuracy-oriented shift with decreased product-based fluency and clause-based complexity under similar writing conditions, and the LLMs reproduced some aspects of that pattern with greater reliability than others.

4. Discussion

4.1. Strong alignment does not establish interchangeability

Our main result is that human-LLM alignment is consistently positive, but numerical agreement is inconsistent. All models were significantly correlated with the human reference on each CAF dimension. However, several model-construct combinations exhibited systematic mean bias and nontrivial normalized error. This pattern shows why correlation alone is not a sufficient validation statistic. One model can retain the relative ordering of essays, but operate on a shifted numerical scale. This change is relevant for research that employs LLM-derived CAF values as variables in subsequent analyzes, as it alters the interpretation of absolute values and may impact thresholds, group comparisons, or longitudinal interpretations.

This distinction is in line with recent psychometric research on LLM raters. Wang et al. (2025) emphasize the importance of assessment of reliability and validity with a range of types of evidence, and Wu et al. (2026) show that high internal consistency can exist with bias and criterion-specific effects. The present study expands that concern from holistic or rubric-based scoring to explicit linguistic measurement. Even when the output is a transparent ratio rather than a subjective writing score, the underlying segmentation and classification decisions create opportunities for systematic disagreement.

The results also temper claims about the possibility of replacing manual coding with LLMs, as they correlate highly with human judgments. Even strong correlation may be enough to make an LLM useful for initial pattern detection. An exact measurement still requires calibration or human verification. The amount of evidence of agreement is thus determined by the intended use. This nuanced meaning is better than a binary conclusion that LLMs can or cannot analyze writing by learners.

4.2. Complexity as the most vulnerable construct

The most consistent result across analyzes concerned complexity. Five models significantly underestimated complexity. Normalized error was generally higher for complexity than for accuracy in the better-performing models. Participant-level direction agreement was lowest for complexity. The explanation is probably the analytic operations required by the measure. S-nodes per T-unit comprises two segmentation decisions: the model has to identify T-unit boundaries and then identify independent and dependent finite or non-finite clauses. The structures in learner writing do not always map neatly to idealized grammar: subjects are omitted, coordination is faulty, syntax is run-on, clauses are incomplete, verb constructions are ambiguous. A small difference in the parsing of these cases affects both the numerator and denominator.

Human consensus coding can stabilize such decisions as reviewers can compare interpretations and resolve ambiguous cases under a shared rule set. Public LLM interfaces, however, rely on latent internal representations and might not follow the syntactic definitions in the prompt as closely. The outcome is not necessarily random error: the negative complexity bias repeated across the models indicates a tendency for multiple models to find fewer S-nodes, more T-units, or a combination thereof when compared to the human reference. This is exactly the kind of construct-specific calibration problem that a global statement about "LLM accuracy" would miss.

The complexity result also concerns other matters than the specific S-node ratio used here. Syntactic complexity in L2 is multidimensional and more advanced writing may be based on phrasal elaboration or complex nominals rather than on increasing clausal subordination alone (Bulté & Housen, 2012; Kyle & Crossley, 2017; Norris & Ortega, 2009). We do not assess these broader indices in the present study. Instead, it demonstrates that even a narrowly defined clause-based index can be sensitive to model-specific interpretation. Thus, future work should validate each automated syntactic measure independently, rather than assuming that success on one index generalizes to others.

4.3. Accuracy and fluency: stronger but not error free

The most consistently reproduced dimension of change was accuracy. All models detected the large increase in accuracy from Essay One to Essay Two. Participant-level direction agreement was highest for accuracy. ChatGPT also had the best human correlation for accuracy (r = .900). One reason for this is that the study's error-free T-unit rule limited the decision space, such that a T-unit either contained a grammatical error under the stated rules or it did not. Grammatical-error classification is subjective by nature, but the prompt explicitly excluded punctuation, spelling, style, coherence, and lexical choice, which likely reduced variation in what models counted as an error.

Finally, we found strong alignment for several models with fluency, especially DeepSeek (r = .918) and Grok (r = .882). However, its apparent simplicity was still dependent on T-unit segmentation, as the measure was the number of words per T-unit. It should also be stressed that this is product-based fluency, rather than temporal fluency. The measure is not a measure of writing speed, pause, or online

processing. The current design is useful in that it provides a transparent productivity ratio that can be compared across human and LLM coding using the same formulas.

4.4. Model choice and the limits of rankings

DeepSeek V4 gave the best overall approximation under the combined criteria used in this dataset, while Grok was closest to the human mean level, and ChatGPT was closest in accuracy alignment. The results demonstrate why model rankings must be viewed as conditional rather than universal. A ranking is sensitive to the constructs, prompt, data, model configuration, error metric and purpose of analysis. Labels on public-facing models can also reflect backend configurations that change over time. The scientific contribution is not that one model consistently outperforms others, but that the model choice has a material effect on measurement and should be transparently reported.

This is particularly relevant for reproducibility. Store prompt text, interface displayed model or mode label, collection date, output format, and raw outputs. Where stable APIs exist, researchers should also report endpoint/version information, relevant decoding settings, and repeated-run stability. In the present study, one prompt was employed, eliminating one source of procedural variability but not stochasticity or hidden model updates.

4.5. Implications for L2 writing research and assessment

For L2 writing researchers, the findings indicate a supplementary rather than a substitutive role of LLM-based CAF analysis. Models can be useful for initial screening, finding patterns in large data sets, or flagging texts for closer human review. However, before using LLM-derived values as definitive research variables, especially if syntactic segmentation is the main concern, construct-specific validation should be conducted. A small, human-coded calibration subset could be utilized to estimate bias and error before applying a model on a larger corpus.

The study reinforces the distinction between analytic assistance and consequential scoring for assessment practice. The current measures were used for research-level linguistic analysis, not for high-stakes decisions about individual learners. Strong pattern-level correspondence does not justify substituting trained human judgment in contexts where exact values, fairness, or individual consequences matter. Check particularly for complexity and for essays with non-standard or ambiguous syntax.

The findings also support a multidimensional reading of learner writing. According to the human reference, Essay Two was more accurate, but less fluent and less syntactically complex. It would be misleading to combine those outcomes into a single ‘improvement’ claim. LLM-assisted CAF analysis can be pedagogically useful because it can make such profiles visible, but only if teachers and researchers understand the operational meaning and limitations of each index.

4.6. Limitations and methodological priorities

The findings are limited in their interpretation by several factors. First, the sample was only 45 learners from one Iranian university and one argumentative topic. Thus, the results should be interpreted as evidence specific to this context, not as a general ranking of LLMs or as a universal estimate of human–LLM agreement. Replication is needed across institutions, proficiency levels, L1 backgrounds, genres, and task types.

Second, the human reference was a reviewed consensus coding, not a set of maintained independent blind ratings. The consensus procedure provides a defensible common reference for matched comparison but does not allow for independent inter-rater reliability to be estimated. In future studies, coder scores prior to consensus should be retained and an appropriate agreement coefficient reported, with disagreements resolved by consensus. This would enable the model-human gap to be understood within the framework of the normal range of human-human disagreement.

Third, each essay-model combination was executed exactly once through a public web interface. The study does not provide estimates of stochastic variability within the model, prompt-run reliability, or effects of back-end model updates. Repeated runs in controlled API settings would provide more evidence of stability. A future design could separate between-model variance from within-model run variance and assess whether observed model differences are larger than repeated-run fluctuation.

Fourth, each participant wrote two essays, which resulted in nesting at the participant level. For the main human-LLM comparisons, matched essay observations were used; occasion comparisons were within-participants. Larger replications should more explicitly model the repeated essays using mixed effects approaches. Agreement inference would also benefit from formal false-discovery-rate or family-wise procedures, concordance correlation coefficients, and raw-data Bland-Altman plots.

Finally, the repeated writing part had the same topic and no instructional intervention. Familiarity or practice effects are indistinguishable from other occasion-related changes. Hence, the repeated-writing analysis here is not evidence for L2 development due to teaching or feedback, but a diagnostic test of whether LLMs replicate a human-observed pattern of change.

5. Conclusion

This study evaluates six LLMs as tools for explicit CAF measurement in L2 argumentative writing by comparing their outputs to reviewed human consensus coding across 90 essays. All models showed significant positive alignment with human values, suggesting LLMs can capture meaningful relative patterns in learner writing. But alignment was not about numerical agreement. Bias, normalized error and participant-level diagnostics revealed well-defined model- and construct-specific differences. Clause-based complexity was the least stable dimension. Accuracy was the most consistently reproduced change dimension.

The major methodological implication is straightforward: human–LLM correlation should not be interpreted as interchangeability. Validation of the LLM-based linguistic measurement needs to occur at

the level of the specific construct, prompt, model configuration and intended use. Under the conditions of this study, DeepSeek V4 was the strongest overall approximation and ChatGPT was the strongest accuracy alignment, but these model-specific results are secondary to the larger conclusion that construct matters. LLMs can assist in CAF analysis, especially in the initial screening and pattern recognition, but accurate quantification still depends on well-defined protocols, calibration, and human validation.

Abbreviations

AES
automated essay scoring
AI
artificial intelligence
API
application programming interface
CAF
complexity, accuracy, and fluency
EFL
English as a foreign language
GPT
generative pre-trained transformer
ICC
intraclass correlation coefficient
KET
Key English Test
L1
first language
L2
second language
LLM
large language model
LLMs
large language models
MA
Master of Arts
MAE
mean absolute error
nRMSE
normalized root mean square error
NS

nonsignificant
RMSE
root mean square error
RQ
research question
SD
standard deviation
SLA
second language acquisition
Clinical trial number
not applicable

Declarations

Ethics Declarations

Institutional approval and ethical oversight

This study was conducted as part of a master's thesis at the University of Maragheh, Iran, under the university's institutional research approval procedures. The research proposal was reviewed by faculty members of the English Language Department and formally approved by the Faculty of Humanities Council on 7 October 2025, before data collection commenced. The approval was documented in the official University of Maragheh resolution dated 12 October 2025 (document no. 1404/100/4725, Item 12). The university's Research Department was identified by the thesis supervisor as the institutional authority responsible for the research approval procedure. The proposal was subsequently registered with the Iranian Research Institute for Information Science and Technology (IranDoc) on 27 October 2025 (registration no. 1998856).

Research conduct and confidentiality

The research involved 45 undergraduate students aged 18–26 years who voluntarily participated in classroom-based argumentative writing tasks. Participant identities were replaced with anonymized codes, and the collected data were used for academic research purposes. Data collection commenced only after institutional proposal approval had been obtained.

Informed consent

Informed consent was obtained from all 45 participants before data collection. Participants confirmed their voluntary participation and consent to the use of their responses for academic research purposes through a mandatory confirmation item in the study's Google Form before completing the writing tasks.

Author Contribution

A.R. contributed to the conceptualization and design of the study, data collection, data coding, analysis and interpretation of the results, and preparation of the original manuscript. A.S. contributed to data verification, interpretation of the findings, and critical review and revision of the manuscript. Both authors reviewed and approved the final manuscript.

Data Availability

The anonymized numerical dataset underlying this study, comprising 90 essay observations and human consensus and six-LLM measurements of complexity, accuracy, and fluency (CAF), is provided as Supplementary Data S1 (Excel). The statistical analysis workbook is provided as Supplementary Data S2 (Excel). Supplementary File S3 contains the participant instructions, Session A and Session B writing sheets, an anonymized example of a participant's original essay and corresponding human CAF coding, and an example of ChatGPT-generated CAF analysis. These supplementary materials are available alongside the article and support verification of the reported statistical analyses and measurement procedures.

References

1. Amini Farsani M, Stapleton P, Jamali HR (2025) Charting L2 argumentative writing: A systematic review. J Second Lang Writ 68 Article 101208. https://doi.org/10.1016/j.jslw.2025.101208
2. An Y, Lee S-K (2025) *The impact of prompt engineering on GPT-4o's scoring reliability in English writing assessment* [Preprint]. SSRN. https://doi.org/10.2139/ssrn.5929615
3. Attali Y (2013) Validity and reliability of automated essay scoring. In: Shermis MD, Burstein J (eds) Handbook of automated essay evaluation: Current applications and new directions. Routledge, pp 181–198
4. Atkinson J, Palma D (2025) An LLM-based hybrid approach for enhanced automated essay scoring. Sci Rep 15., Article 14551. https://doi.org/10.1038/s41598-025-87862-3
5. Barrot JS (2023) Using ChatGPT for second language writing: Pitfalls and potentials. Assess Writ 57., Article 100745. https://doi.org/10.1016/j.asw.2023.100745
6. Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 327(8476):307–310. https://doi.org/10.1016/S0140-6736(86)90837-8
7. Bui NM, Barrot JS (2025) ChatGPT as an automated essay scoring tool in the writing classrooms: How it compares with human scoring. Educ Inform Technol 30:2041–2058. https://doi.org/10.1007/s10639-024-12891-w
8. Bulté B, Housen A (2012) Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency*: *Complexity, accuracy and fluency in SLA* (pp. 21–46). John Benjamins. https://doi.org/10.1075/lllt.32.02bul

9. Carlo M, Takeuchi O (2026) Do large language models encode second-language writing proficiency? A CALF-based perspective. Digit Stud Lang Literature 3(1):97–142. https://doi.org/10.1515/dsll-2025-0019
10. Chen Y (2026) Evaluating GPT ratings of EFL writing: A scoping review. Assess Writ 68:101044. https://doi.org/10.1016/j.asw.2026.101044
11. Han T, Zheng Y (2026) AI and human scoring for postgraduate writing: Evaluating score reliability, variability, and rater behaviours. Stud Educational Evaluation 88 Article 101572. https://doi.org/10.1016/j.stueduc.2026.101572
12. Housen A, Kuiken F (2009) Complexity, accuracy and fluency in second language acquisition. Appl Linguist 30(4):461–473. https://doi.org/10.1093/applin/amp048
13. Housen A, Kuiken F, Vedder I (2012) Complexity, accuracy and fluency: Definitions, measurement and research. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency*: *Complexity, accuracy and fluency in SLA* (pp. 1–20). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.32.01hou
14. Huang Y, Palermo C, Wilson J (2026) Accuracy and fairness of generative AI in automated essay scoring: Comparing GPT-4o, feature-based models, and human raters. Assess Writ 69:101047. https://doi.org/10.1016/j.asw.2026.101047
15. Jiang Z, Xu Z, Pan Z, He J, Xie K (2023) Exploring the role of artificial intelligence in facilitating assessment of writing performance in second language learning. Languages 8(4):247. https://doi.org/10.3390/languages8040247
16. Kim H, Baghestani S, Yin S, Karatay Y, Kurt S, Beck J, Karatay L (2024) ChatGPT for writing evaluation: Examining the accuracy and reliability of AI-generated scores compared to human raters. In C. A. Chapelle, G. H. Beckett, & J. Ranalli (Eds.), *Exploring artificial intelligence in applied linguistics* (pp. 73–95). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2024.154.06
17. Kyle K, Crossley SA (2017) Assessing syntactic sophistication in L2 writing: A usage-based approach. Lang Test 34(4):513–535. https://doi.org/10.1177/0265532217712554
18. Lee Y, Hwang M (2026) ChatGPT for automated writing evaluation: Scoring and feedback across prompt conditions. *Language Testing*. Advance online publication. https://doi.org/10.1177/02655322261425759
19. Li S, Yu H (2024) Effects of topic familiarity on L2 writing processes and behaviors. Int J Appl Linguistics 34(1):348–366. https://doi.org/10.1111/ijal.12497
20. Lu X (2010) Automatic analysis of syntactic complexity in second language writing. Int J Corpus Linguistics 15(4):474–496. https://doi.org/10.1075/ijcl.15.4.02lu
21. Mizumoto A, Eguchi M (2023) Exploring the potential of using an AI language model for automated essay scoring. Res Methods Appl Linguistics 2(2) Article 100050. https://doi.org/10.1016/j.rmal.2023.100050
22. Norris JM, Ortega L (2009) Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. Appl Linguist 30(4):555–578. https://doi.org/10.1093/applin/amp044

23. Ortega L (2003) Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. Appl Linguist 24(4):492–518. https://doi.org/10.1093/applin/24.4.492
24. Pallotti G (2009) CAF: Defining, refining and differentiating constructs. Appl Linguist 30(4):590–601. https://doi.org/10.1093/applin/amp045
25. Phuoc VD, Barrot JS (2022) Complexity, accuracy, and fluency in L2 writing across proficiency levels: A matter of L1 background? Assess Writ 54 Article 100673. https://doi.org/10.1016/j.asw.2022.100673
26. Powers DE, Escoffery DS, Duchnowski MP (2015) Validating automated essay scoring: A (modest) refinement of the 'gold standard. ' Appl Meas Educ 28(2):130–142. https://doi.org/10.1080/08957347.2014.1002920
27. Ramesh D, Sanampudi SK (2022) An automated essay scoring systems: A systematic literature review. Artif Intell Rev 55(3):2495–2527. https://doi.org/10.1007/s10462-021-10068-2
28. Ramineni C, Williamson DM (2013) Automated essay scoring: Psychometric guidelines and practices. Assess Writ 18(1):25–39. https://doi.org/10.1016/j.asw.2012.10.004
29. Shermis MD (2014) State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. Assess Writ 20:53–76. https://doi.org/10.1016/j.asw.2013.04.001
30. Skehan P (2009) Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. Appl Linguist 30(4):510–532. https://doi.org/10.1093/applin/amp047
31. Skehan P (2016) Tasks versus conditions: Two perspectives on task research and their implications for pedagogy. Annu Rev Appl Linguist 36:34–49. https://doi.org/10.1017/S0267190515000100
32. Tang X, Chen H, Lin D, Li K (2024) Harnessing LLMs for multi-dimensional writing assessment: Reliability and alignment with human judgments. Heliyon 10(14) Article e34262. https://doi.org/10.1016/j.heliyon.2024.e34262
33. Wang Y, Huang J, Du L, Guo Y, Liu Y, Wang R (2025) Evaluating large language models as raters in large-scale writing assessments: A psychometric framework for reliability and validity. Computers Education: Artif Intell 9:100481. https://doi.org/10.1016/j.caeai.2025.100481
34. Wolfe-Quintero K, Inagaki S, Kim H-Y (1998) Second language development in writing: Measures of fluency, accuracy, & complexity. University of Hawai'i
35. Wu H-N, Chu M-N, Hsu J-L (2026) Comparing GPT and human raters in essay assessment: Variability, bias, and the potential of LLM-based scoring. Computers Educ Open 10:100341. https://doi.org/10.1016/j.caeo.2026.100341
36. Wu MY, Steinkrauss R, Lowie W (2023) The reliability of single task assessment in longitudinal L2 writing research. J Second Lang Writ 59 Article 100950. https://doi.org/10.1016/j.jslw.2022.100950
37. 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. Br J Edu Technol 56(1):150–166. https://doi.org/10.1111/bjet.13494

Figures

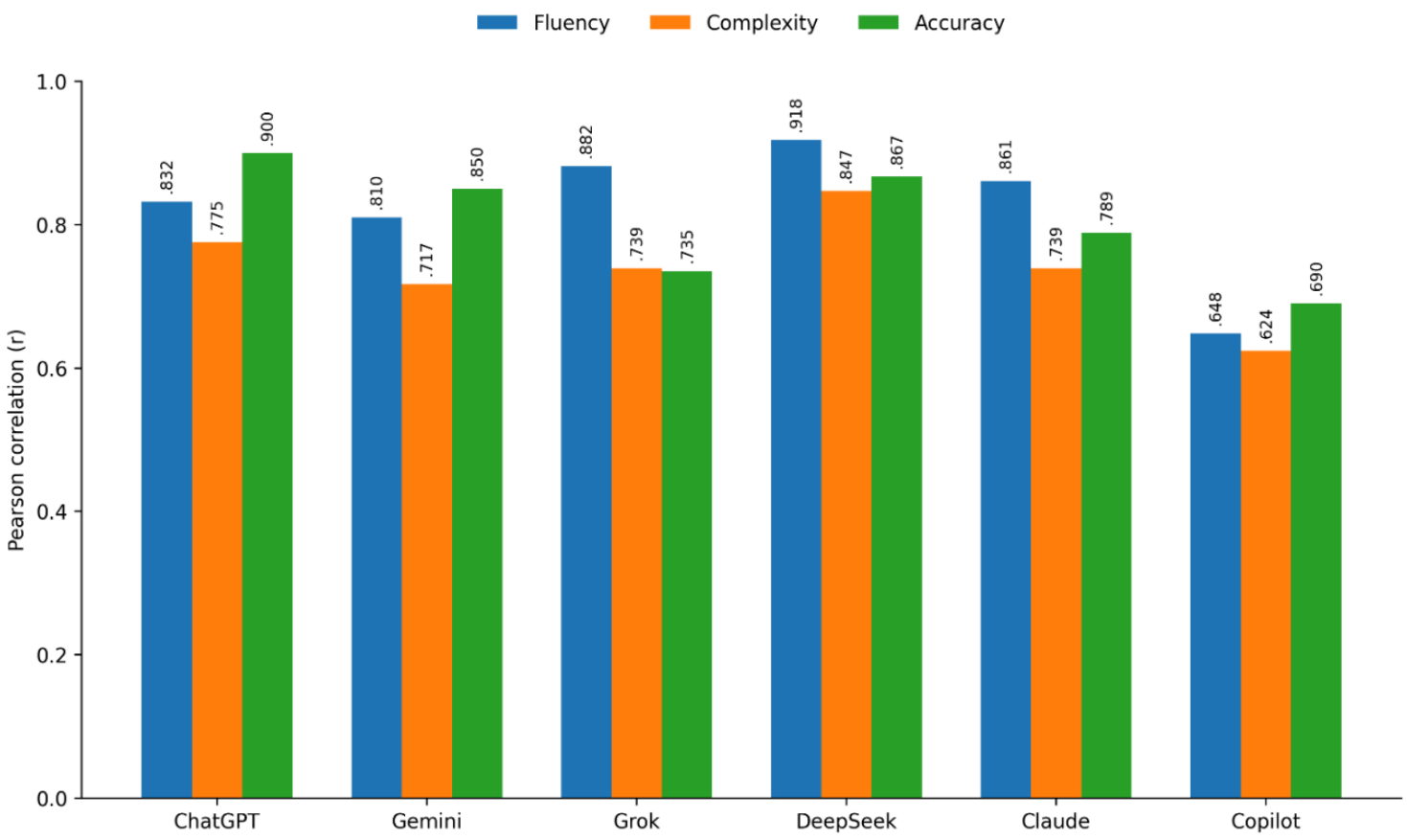


Figure 1

Human-to-LLM Pearson correlations across fluency, complexity, and accuracy. All correlations were $p < .001$. Higher values indicate stronger alignment with the reviewed human consensus reference.

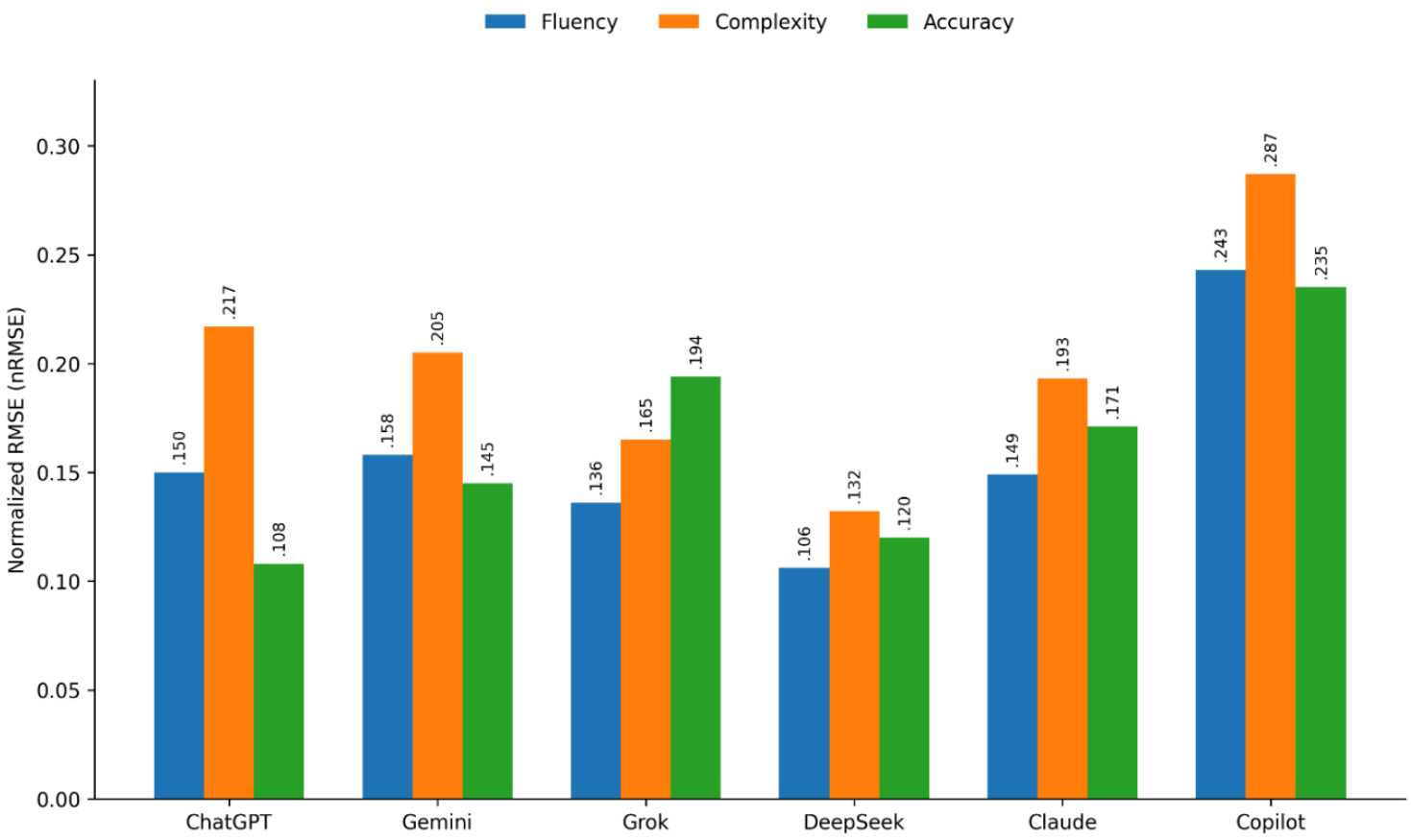


Figure 2

Normalized RMSE (nRMSE) for each model and CAF dimension. Lower values indicate closer numerical agreement with the reviewed human consensus reference after scale adjustment.

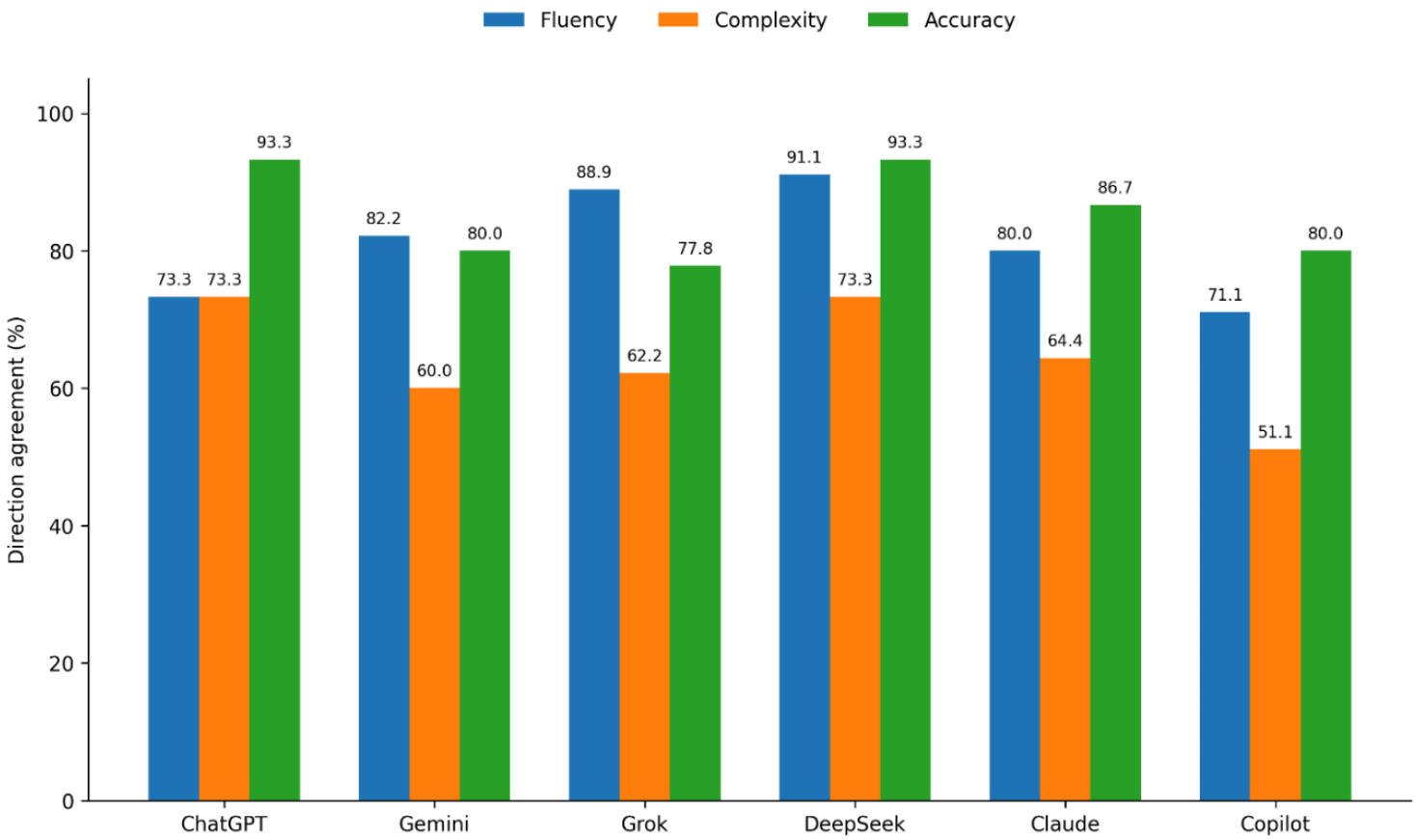


Figure 3

Participant-level agreement in the direction of Essay One-to-Essay Two change. Values show the percentage of learners for whom each LLM and the reviewed human consensus reference indicated the same direction of change.

Supplementary Files

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

- SupplementaryFileS3.pdf
- SupplementaryDataS2Analysis.xlsx
- SupplementaryDataS1CAF.xlsx