

Designing and Evaluating a Measurement-Oriented RAG Framework for ELL Writing Assessment

Daeryong Seo

Pearson*

daeryong.seo@pearson.com

Abstract

This design-focused study compared measurement-oriented ELL_RAG with conventional RAG for writing-assessment analytics. ELL_RAG showed a descriptive advantage overall and on global/population-level tasks, while local-evidence performance was nearly equivalent. Findings support task–evidence alignment rather than universal superiority and highlight tool orchestration as a continuing design challenge.

1 Introduction

Retrieval-augmented generation (RAG) grounds language-model output in external evidence and has proved useful when relevant information can be retrieved directly from a knowledge source (Lewis et al., 2020). More recent RAG architectures have become increasingly modular and domain-specific (Gao et al., 2023). These developments are relevant to educational writing assessment, where large language models can support scoring, feedback, and interpretation but must do more than generate fluent explanations. Assessment claims require evidence appropriate to their intended interpretations and uses, including attention to validity, reliability, fairness, and uncertainty (Messick, 1989; Kane, 2013; Pack et al., 2024). A response may be fluent and plausible yet still be weak if the evidence used is not appropriate for the conclusion. For example, retrieved individual essays cannot by themselves support population-level claims.

This creates an important evidence-design problem. Some questions are local and can be answered from an official rubric, structured metadata, or a retrieved essay. Others require full-dataset statistics, inter-rater reliability, controlled

subgroup comparisons, or other population-level analyses. A semantically relevant essay may therefore be useful context while still being inappropriate evidence for a dataset-level claim. Conversely, deterministic statistics do not replace the interpretive value of a local response when the question concerns a specific essay or rubric application. The broader RAG literature reflects a similar distinction: Edge et al. (2024) contrasted local retrieval with global corpus-level sensemaking, and Murtaza et al. (2025) combined document retrieval with database function calls for structured-data questions.

The present study addresses this problem by designing ELL_RAG, a research prototype for ELL writing-assessment analytics. The architecture combines deterministic measurement tools, constrained essay retrieval, rubric lookup, and LLM orchestration under a hierarchical evidence design: dataset-level claims come from deterministic analyses, official score meanings from rubric evidence, and retrieved essays from local contextual evidence. ELL_RAG is compared with a conventional RAG baseline to examine whether these additional mechanisms provide incremental value when population-level evidence is required while preserving performance when retrieval alone may be sufficient. The emphasis is therefore architectural design and task–evidence alignment rather than a general contest between two systems.

1.1 Study Purpose and Research Questions

The primary aim was to design and examine a measurement-oriented RAG workflow that selects retrieval, rubric, deterministic measurement, or combined evidence pathways according to the level of inference requested (Appendix A). The comparison with Conv_RAG was used

* This study describes the author’s research prototype developed using an open-source dataset. It is not an operational Pearson scoring system or client implementation and is not intended to replace human judgment, psychometric review, validation, or established governance processes.

diagnostically to identify where conventional retrieval remained sufficient, where measurement-oriented mechanisms added support, and what strengths or limitations characterized each architecture. The evaluation was intentionally descriptive and served as evidence about whether the proposed design behaved as intended.

Accordingly, the study addressed three research questions:

1. How does ELL_RAG differ from Conv_RAG in supporting evidence selection, integration, and interpretation across the 15 prespecified questions compared?
2. How do ELL_RAG and Conv_RAG differ in their performance on local, directly retrievable evidence tasks versus global/population-level measurement tasks?
3. How do ELL_RAG and Conv_RAG differ in their strengths and limitations related to evidence selection, measurement interpretation, uncertainty control, scope alignment, and tool orchestration?

2 Method

2.1 Sample(s)

The study used 5,840 essays from the ELLIPSE Corpus, which contains approximately 6,500 ELL essays from U.S. statewide testing across 44 prompts. Essays were scored by two trained raters on Overall proficiency and six analytic dimensions, with accompanying student demographic variables (Crossley et al., 2023). The analytic sample included 5,840 complete cases used in the RAG analyses.

2.2 Study Design

The study used one frozen, design-focused parallel comparison of Conv_RAG and ELL_RAG. Both systems independently answered the same 15 prespecified questions without conversational history, producing 30 responses. The frozen protocol prevented iterative adjustment of questions or prompts after observing system outputs. The study emphasized architecture, evidence-pathway behavior, and preliminary response quality rather than inferential claims of universal superiority (Appendix A).

2.3 Systems Compared

Conv_RAG used conventional semantic retrieval and generation over corpus documents, rubric material, structured metadata, and essays. It did not include dedicated deterministic functions for full-dataset summaries, reliability estimation, or controlled subgroup analyses. ELL_RAG retained local retrieval but added measurement-aware routing, deterministic analytical tools, authoritative rubric lookup, structured eligibility controls, and an explicit local-versus-global evidence hierarchy. Dataset-level questions could therefore be routed to measurement functions, while local questions could continue to use retrieval, rubric, or metadata pathways.

2.4 Matched Design Conditions

The systems used the same ELLIPSE corpus of 5,840 essays (Crossley et al., 2023), sentence-transformers/all-MiniLM-L6-v2 embeddings, GPT-4o base model, generation settings, final evidence depth, and frozen question set. The intended difference was architectural: Conv_RAG used conventional retrieval, whereas ELL_RAG added measurement-support pathways and a larger initial retrieval pool before filtering/reranking to the final evidence set.

2.5 Evaluation Question Set

The 15 questions were developed based on a review of relevant assessment, measurement, and retrieval-augmented analytics literature and standards and were distributed evenly across five assessment-oriented categories spanning 1) validity, fairness, and quality; 2) subgroup analysis and oversight; 3) measurement and retrieval evidence; 4) score interpretation; and 5) evidence and uncertainty. The question-development framework drew on established guidance for assessment validity, reliability, fairness, and score interpretation, the ELLIPSE assessment context, and literature concerning the integration of retrieved and structured evidence (AERA, APA & NCME, 2014; Cohen, 1968; Crossley et al., 2023). Five items (Q02, Q04, Q05, Q07, and Q13) were prespecified global/population-level probes requiring dataset-level evidence or integration with such evidence; three (Q08, Q09, and Q11) were local-evidence probes centered on appropriate use of essays, rubrics, or metadata. These were descriptive task groupings, not psychometric subscales (Appendices B and C).

2.6 Scoring Framework

The study author scored all 30 responses on four criteria, each from 0 to 2: Evidence Use, Grounding, Accuracy, and Scope Alignment (C1); Measurement Interpretation (C2); Uncertainty and Overclaim Prevention (C3); and Clarity and Completeness (C4). The maximum response score was 8. Scores of 0, 1, and 2 represented major failure, partial success with a material limitation, and accurate/appropriately bounded performance, respectively.

For C1, for example, a score of 0 indicated a fundamental evidence-source or scope failure, such as substituting local essays for a required population-level analysis; 1 indicated meaningful but incomplete or materially flawed evidence use; and 2 required accurate, appropriately sourced, scope-aligned evidence. Therefore, C1–C4 constituted the complete formal scoring framework.

2.7 Analysis

Analysis was descriptive and design-focused. System totals, percentages, mean item scores, C1–C4 means, category summaries, and item differences were examined. Global/population-level and local-evidence probes were summarized separately to test the study’s central design expectation that architecture behavior would depend on evidentiary demand. Scoring rationales were also reviewed for recurring strengths and limitations such as safe abstention, local-to-global scope mismatch, routing error, and incomplete multi-source integration. Because the study used only 15 questions and one author-rater, no inferential test of superiority was used, and category means were treated only as summaries of the sampled items.

3 Results

Results are interpreted as preliminary evidence about architectural behavior under different evidentiary demands.

3.1 RQ1. Overall Response Quality and Performance Across C1–C4

Across 15 items, ELL_RAG scored 105/120 (87.5%; mean = 7.00/8), compared with 88/120 (73.3%; mean = 5.87/8) for Conv_RAG. This represents a descriptive advantage for ELL_RAG of 17 total points, 14.2 percentage points, or 1.13

points per item. ELL_RAG received eight perfect item scores and Conv_RAG six (Table 1). These totals summarize the frozen task set but do not by themselves establish general system superiority.

System / Difference	Total Score	Percent	Mean Item Score	8/8 Items
Conv_RAG	88/120	73.3%	5.87/8	6
ELL_RAG	105/120	87.5%	7.00/8	8
ELL_RAG – Conv_RAG	+17	+14.2 pp	+1.13	+2

Table 1: Overall Descriptive Performance Across the 15 Items

The largest criterion difference occurred for C1 (1.60 vs. 1.13), followed by C2 (1.80 vs. 1.53) and C3 (1.93 vs. 1.67); C4 was closest (1.67 vs. 1.53). Thus, the main descriptive contrast concerned evidence use, scope, and measurement interpretation more than response clarity (Figure 1).

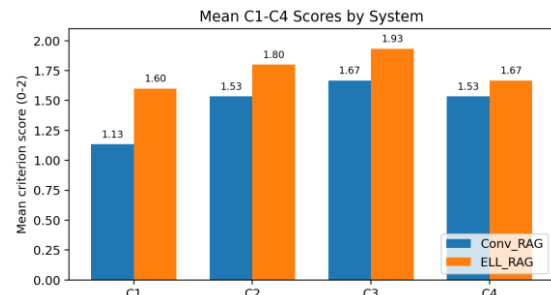


Figure 1: Mean C1–C4 scores by system. Each criterion was scored from 0 to 2.

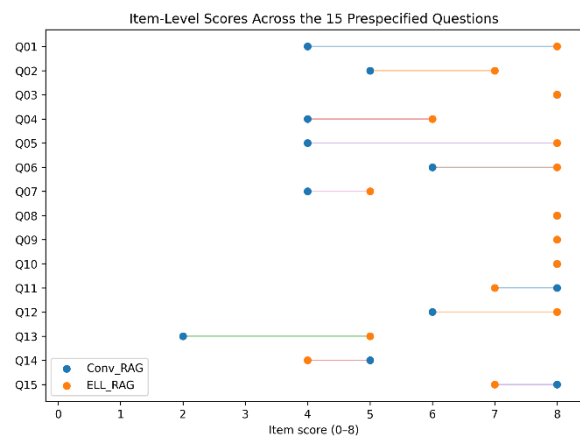


Figure 2: Item-level response scores across Q01–Q15.

ELL_RAG scored higher on eight questions, Conv_RAG on three, and four were ties (Figure 2). The largest category differences favored ELL_RAG in Subgroup fairness/transparency/human oversight (7.33 vs. 4.67) and Validity/fairness/quality (7.67 vs. 5.67); differences in the remaining three categories were only 0.33 points (Table 2). Because each category

contained three items, these values are descriptive rather than stable category effects.

Evaluation Category	Conv_R AG	ELL_RA G	Difference e
Validity, fairness, and quality in assessment interpretation	5.67	7.67	+2.00
Subgroup fairness, transparency, and human oversight	4.67	7.33	+2.67
Integration of measurement and retrieval evidence	6.67	7.00	+0.33
Responsible score interpretation and professional use	7.33	7.67	+0.33
Ethical handling of evidence, uncertainty, and overclaiming	5.00	5.33	+0.33

Table 2: Category-Level Mean Item Scores

Note. Scores are means on the 0–8 item scale. Difference = ELL_RAG minus Conv_RAG. Each category contained three items; category means are therefore descriptive summaries of the sampled tasks rather than estimates of separate subscales.

3.2 RQ2. Performance Under Global/Population-Level Versus Local Evidentiary Demands

3.2.1 Global and Population-Level Evidence

On the five global/population-level probes, ELL_RAG obtained 31/40 (mean = 6.20/8) versus 19/40 (3.80/8) for Conv_RAG. C1 means were 1.20 and 0.20, respectively. These questions required full-dataset evidence, reliability information, controlled subgroup comparisons, or explicit integration of dataset-level evidence with other sources. The difference is therefore best interpreted as performance under a specific evidentiary demand rather than as a general ranking.

Conv_RAG often recognized when retrieved context could not support the requested statistic and declined to fabricate unavailable population-level results, demonstrating useful safe abstention. In several cases it could describe what analysis would be needed but could not provide the requested quantitative evidence. ELL_RAG more often supplied such evidence because it could route the question to deterministic measurement functions. This distinction reflects architecture rather than simple retrieval quality.

ELL_RAG nevertheless showed orchestration weaknesses. Q07 used an unnecessary analytical framing, Q13 omitted requested numerical results, and Q14 incompletely integrated multiple evidence sources. These cases indicate that tool availability did not always translate into correct routing, evidence integration, or final interpretation.

3.2.2 Local Evidence

The local-evidence probes showed near-equivalent performance: Conv_RAG scored 24/24 and ELL_RAG 23/24, with both averaging 2.00/2 on C1 and C3. Both appropriately prioritized the official Vocabulary rubric over a conflicting retrieved example on Q08 and used structured prompt metadata to determine eligibility on Q09. On Q11, Conv_RAG was one point higher because it more explicitly stated that an authorized human review process, rather than the system alone, would be required before changing an official score.

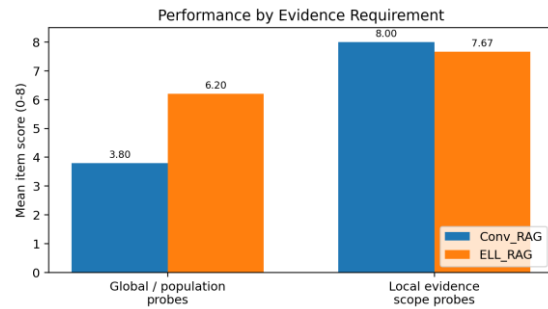


Figure 3: Mean scores on the prespecified global/population-level and local-evidence task groupings.

Figure 3 summarizes the central design pattern: the architectures were nearly equivalent when evidence was local and directly retrievable, whereas ELL_RAG showed a substantially larger descriptive advantage when the task required global/population-level measurement evidence. The near-tie on local questions is important because it indicates that the measurement-oriented extension did not create a uniform advantage; conventional retrieval remained highly capable when its evidence mechanism matched the task.

3.3 RQ3. Architectural Strengths and Limitations

ELL_RAG: ELL_RAG’s principal strength was connecting population-level questions to full-dataset measurement evidence, including reliability statistics, descriptive summaries, and controlled comparisons. Its main limitation was orchestration: appropriate tools could be available yet incompletely selected, executed, or combined. Q07, Q13, and Q14 illustrate that measurement-oriented RAG requires reliable routing, evidence verification, and multi-source composition in addition to deterministic functions.

Conv_RAG: Conv_RAG’s principal strength was local, directly retrievable evidence. It performed at or near ceiling on several rubric-

metadata-, and professional-interpretation items and frequently abstained appropriately when dataset-level evidence was unavailable. Its main limitation was structural: without a deterministic analysis pathway, it often could not answer full-dataset reliability or subgroup questions. For example, Q13 showed consequential risk of using locally retrieved examples in place of a requested population-level analysis.

3.4 Synthesis Across the Three Research Questions

Across the three research questions, the findings point to task–evidence alignment rather than a simple comparative ranking of the two architectures. ELL_RAG’s advantage was concentrated in tasks requiring deterministic computation or dataset-level aggregation, whereas conventional retrieval remained highly effective when the relevant evidence was local and directly accessible. The results therefore support a selective extension of conventional RAG with measurement-oriented capabilities, rather than broad replacement of the conventional architecture, while also highlighting the additional routing and evidence-integration demands introduced by tool-augmented systems.

4 Discussion

The findings support a design-oriented interpretation in which the value of measurement-oriented RAG depends on evidentiary demand. ELL_RAG showed its clearest descriptive advantage on questions requiring global or population-level evidence, whereas the two systems performed nearly equivalently when the needed evidence was local and directly retrievable. This pattern is more informative than the overall score difference because it shows that ELL_RAG’s advantage was concentrated in tasks requiring dataset-level computation rather than retrieval alone. It also distinguishes retrieval relevance from evidentiary sufficiency: a retrieved essay may be highly relevant to a question but still be insufficient to support subgroup or population-level conclusions. This distinction helps explain the complementary strengths of the two architectures. Conventional RAG remained effective when relevant evidence could be retrieved directly, while measurement-oriented RAG provided additional support when the requested inference required

structured computation, aggregation, or analysis across the broader dataset.

At the same time, the findings show that additional analytical capabilities do not automatically produce stronger responses. Their value depends on appropriate routing, accurate execution, effective evidence integration, and verification. In this sense, tool orchestration is a central design requirement rather than a secondary implementation detail. Appropriate acknowledgment of insufficient evidence is equally important because a system should state when the available evidence cannot support the requested statistic or conclusion, rather than provide an unsupported quantitative answer.

The observed pattern is consistent with broader RAG research distinguishing localized retrieval from tasks requiring corpus-level aggregation or structured-data access (Edge et al., 2024; Murtaza et al., 2025). The present study does not establish a universal local-versus-global rule, but it extends this distinction into measurement-oriented assessment analytics. The practical implication is therefore not to replace conventional RAG broadly, but to match the evidence mechanism to the level of inference required: use retrieval when retrieval is sufficient, invoke deterministic analysis when dataset-level evidence is needed, coordinate multiple evidence sources when necessary, and retain human oversight for consequential interpretations.

5 Limitations

These findings remain preliminary because the study used a small prespecified question set and one author-rater. The results should therefore be interpreted as descriptive evidence about architectural behavior rather than as proof of general system superiority. The sampled questions were designed to probe measurement-oriented behavior and do not represent the full range of possible RAG tasks. Future work should examine whether the same local-versus-global pattern replicates with broader task sampling and independent raters. Further development should also strengthen routing, multi-source evidence integration, and post-tool verification.

6 Conclusion

The study supports a bounded design principle: retrieval, computation, and interpretation should be aligned with the evidentiary level of the question. Conventional RAG remains effective when evidence is local and directly retrievable, whereas measurement-oriented extensions appear most useful when the task requires structured computation, aggregation, or population-level analysis. The contribution is therefore not a claim of universal architectural superiority, but a selective design approach in which additional measurement-oriented capabilities are introduced when the evidentiary demands of the task justify them.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4), 213–220.
- Crossley, S., Tian, Y., Baffour, P., Franklin, A., Kim, Y., Morris, W., Benner, M., Picou, A., & Boser, U. (2023). The English Language Learner Insight, Proficiency and Skills Evaluation (ELLIPSE) Corpus. *International Journal of Learner Corpus Research*, 9(2), 248–269.
- Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Ness, R. O., & Larson, J. (2024). From local to global: A graph RAG approach to query-focused summarization [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2404.16130>
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). *Retrieval-augmented generation for large language models: A survey*. arXiv. <https://arxiv.org/abs/2312.10997>
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv. <https://arxiv.org/abs/2005.11401v4>.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). American Council on Education / Macmillan.
- Murtaza, S. S., Nie, Y., Avan, E., Soni, U., Liao, W., Carnegie, A., Mathias, C. J., Jiang, J., & Wen, E. (2025). Implementing retrieval augmented generation technique on unstructured and structured data sources in a call center of a large financial institution. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)* (pp. 598–606). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.naacl-industry.48>
- Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, 6, 100234.

A Design Workflow for Measurement-Oriented ELL_RAG

ELL Writing Dataset

- 5,840 complete cases
- Essay text, scores, rubrics, and rater information
- Grade, Ethnicity, Gender, Economically Disadvantaged Status, and Prompts
- Linguistic features: num_words, MTLD, and TTR

Deterministic Measurement-Tool Construction

- Group and full-dataset descriptive summaries
- Controlled subgroup comparisons
- Reliability metrics
- Rater-disagreement analyses
- Rubric lookup
- Dataset-level linguistic-feature summaries
- Identification of large rater-disagreement cases

Tool-Level Functional Checks

- Check intended analytic-tool invocation
- Review structured statistical outputs

- Check application of grade and prompt controls where applicable
- Check warnings for multiple prompts and small samples
- Check unsupported or unavailable comparisons
- Functional verification of tools before formal system comparison

Essay Vectorization and Retrieval Preparation

- Convert 5,840 essays into LangChain Documents
- Store score and metadata with each essay
- Embed essays using sentence-transformers/all-MiniLM-L6-v2
- Store embedded essays in Chroma vector database
- Confirm collection count of 5,840 documents

Preliminary Semantic-Retrieval Check

- Execute targeted similarity queries
- Inspect retrieved essay text and associated metadata
- Review similarity results for consistency with intended query characteristics
- Functional retrieval check only; not a formal retrieval-validation study

User Measurement Question

Evidence-Demand Classification and Routing

- Local Evidence Route
 - Retrieve relevant essays when local examples are requested
 - Use official rubric lookup for score-meaning questions
 - Use structured metadata for eligibility or factual constraints
 - Treat retrieved essays as local illustrative or diagnostic evidence only
- Global / Population-Level Measurement Route
 - Use full-dataset descriptive summaries when corpus-level quantities are requested
 - Use reliability and rater-disagreement functions for rater-level questions
 - Use controlled subgroup functions for group-level comparisons

- Use dataset-level linguistic summaries for corpus- or subgroup-level feature questions
- Do not infer population quantities from a small set of retrieved essays

• Combined Evidence Route

- Combine deterministic measurement evidence with rubric and/or local essay evidence when needed
- Preserve the distinct evidentiary role and scope of each source
- Add sample-size, scope, and interpretation cautions where appropriate

Evidence Assembly

- Dataset-level measurement evidence
- Official rubric evidence
- Reliability and controlled-comparison evidence
- Structured metadata
- Local essay evidence
- Dataset-level linguistic-feature evidence

LLM Synthesis Layer

- GPT-4o with temperature = 0.0
- LangChain-based orchestration
- Uses retrieved evidence and deterministic tool outputs
- Does not independently calculate dataset-level statistics
- Does not define official rubric meanings without rubric lookup
- Maintains local-versus-global evidence boundaries during synthesis

Measurement-Oriented Response

- Direct interpretation or summary
- Relevant measurement evidence
- Evidence scope or contextual classification, when needed
- Caution and limitations
- Sources / evidence identifiers

B Representative Evaluation Questions

Category 1: Validity, fairness, and quality in assessment interpretation

Question: Suppose the official Cohesion rubric descriptor for score 3 is retrieved correctly, full-dataset Cohesion rater reliability is available, and several retrieved essays with Cohesion score 3

appear broadly compatible with the descriptor. What interpretation is supported when these sources are considered together, and what broader validity or fairness claim would still require additional evidence?

Category 2: Subgroup fairness, transparency, and human oversight

Question: For Grammar, examine rater agreement separately by observed economic status category. Report paired sample size, exact agreement, agreement within one point, and weighted kappa where estimable, and identify any subgroup result that should be treated cautiously because of limited data or uncertainty.

Category 3: Integration of measurement and retrieval evidence

Question: A retrieved essay assigned Vocabulary score 3 appears weaker than the official Vocabulary score-3 descriptor might suggest. State the score meaning from the official rubric and explain how the apparently inconsistent essay should—and should not—be used as local evidence.

Category 4: Responsible score interpretation and professional use

Question: A teacher asks the system to replace a student's official Conventions score because several semantically similar essays received higher scores. How should the system respond, and what student-specific evidence, rubric-based review, and authorized human process would be required for a legitimate score challenge?

Category 5: Ethical handling of evidence, uncertainty, and overclaiming

Question: Compare economic-status group patterns in num_words, TTR, and MTLT using dataset-level linguistic summaries. Describe the largest observed differences, what they may suggest about the sampled writing, and which broader conclusions are and are not supported by these features alone.

Note. One representative question is shown from each of the five prespecified evaluation categories. Questions are provided to illustrate the scope of the evaluation and were not selected based on system performance.

C Illustrative Paired Responses from Conv_RAG and ELL_RAG

Question: For Grammar, examine rater agreement separately by observed economic status category.

Report paired sample size, exact agreement, agreement within one point, and weighted kappa where estimable, and identify any subgroup result that should be treated cautiously because of limited data or uncertainty.

Conv_RAG Response

Summary

The analysis of rater agreement for the Grammar score by observed economic status category involves examining paired sample size, exact agreement, agreement within one point, and weighted kappa. However, the retrieved context does not provide specific numerical results or detailed subgroup analyses for these metrics.

Evidence

The corpus documentation indicates that the dataset includes various writing-score fields, including Grammar, and contextual fields such as economic status (SES) [Source 8]. However, specific data on rater agreement metrics like exact agreement, agreement within one point, and weighted kappa by economic status category is not provided in the retrieved context.

Interpretation

In general, rater agreement metrics such as exact agreement, agreement within one point, and weighted kappa are used to assess the consistency and reliability of ratings across different raters. Exact agreement measures the percentage of cases where raters give the same score. Agreement within one point considers scores that are within one point of each other as agreeing. Weighted kappa accounts for the degree of disagreement, providing a more nuanced measure of agreement that considers the severity of disagreements.

For subgroup analyses, such as by economic status, it is important to ensure that there is sufficient data for each subgroup to produce reliable estimates. Small sample sizes can lead to high uncertainty in agreement metrics, making some subgroup results less reliable.

Limitations

The retrieved context does not provide specific numerical results or detailed analyses of rater agreement metrics by economic status category. Without this information, it is not possible to report the paired sample size, exact agreement, agreement within one point, or weighted kappa for each economic status category. Additionally, without subgroup-specific data, it is not possible to identify which subgroup results should be treated cautiously due to limited data or uncertainty.

Sources

[Source 8]: ELLIPSE analytic dataset documentation.

ELL_RAG Response

Reliability Interpretation

The reliability of Grammar scoring was examined separately for economically disadvantaged and not economically disadvantaged groups. The analysis focused on exact agreement, agreement within one point, and weighted kappa for each subgroup.

Rater Agreement Evidence

For the economically disadvantaged group ($n = 4050$), the exact agreement between raters was 51.3%, with 99.0% agreement within one point. The weighted kappa was 0.556, indicating moderate agreement. For the not economically disadvantaged group ($n = 1789$), the exact agreement was slightly higher at 52.1%, with 99.4% agreement within one point. The weighted kappa for this group was 0.628, suggesting a moderate-to-strong agreement.

Essay Context, Not Reliability Evidence

No essay-level diagnostic evidence was used in this analysis. Therefore, specific essay features that might have influenced rater agreement are not available for contextual interpretation.

Caution About Sample Size and Generalization

Both subgroups met the minimum paired sample size required for reporting reliability coefficients. However, while the sample sizes are sufficient for reporting, the reliability indices should be interpreted with caution as they do not establish precision or stability of the subgroup estimates. Additionally, subgroup agreement differences are descriptive signals for further review and do not prove scoring bias or fairness.

Sources: No specific text_ids were used in this analysis.

Note. Response headings may vary according to the evidentiary demands of the question. Responses are reproduced verbatim from the frozen evaluation run.