

Semantic Similarity is Not Enough for Comparing LLM and Human Rater Rationales

Julie Jung¹, Max Lu¹, Dogus Darici², Emilia Brüegge³

¹Harvard Graduate School of Education, ²Medical University of Lausitz-Carl Thiem,

³University Hospital of Münster

Correspondence: juliejung@g.harvard.edu

Abstract

Comparing LLM-human rater rationales using semantic similarity risks conflating textual proximity with evaluative agreement. We test whether embedding-based similarity reflects qualitative coding distinctions across rationale pairs in medical education. Similarity declined as rationale length differences grew and was less effective at distinguishing whether LLMs preserved the human's central claim.

1 Introduction

As Large Language Models (LLMs) are increasingly used as raters of text and discourse data, researchers have begun comparing not only the scores between LLM and human raters but also the rationales they generate to justify the scores (Hua et al., 2025; Hou et al., 2025). As noted by scholars, score agreement alone does not establish that LLMs and humans reason in alignment, particularly in formative educational contexts where the rationale behind a score becomes the feedback for the learner.

Traditionally, semantic similarity was used as the main method for comparing rationale similarities across LLMs and humans. However, we argue that this dominant approach has important methodological limitations. Most importantly, semantic similarity metrics are built to detect shared vocabulary and textual proximity, not necessarily to detect whether two texts agree evaluatively. Two rationales may have high similarity by addressing the same aspect of a student's performance and still differ in ways that matter for formative use: one may preserve a central claim and provide additional evidence, while another may shift its valence, qualify it in the opposite direction, or omit it altogether. Treating semantic similarity as evidence of rationale alignment therefore risks construct misspecification.

In this study, we investigate whether similarity scores can reflect the evaluative distinctions

between LLM and human rationales in a clinical reasoning assessment context. Using LLM-human rationale pairs from one human rater and 27 differently configured ChatGPT-5.4 raters evaluating student-patient dialogues, we compare similarity scores against a two-stage coding qualitative scheme that classifies each pair by whether the LLM rationale covers the human rater's central evaluative claim and, if partially, in what way.

By testing whether a widely used comparison method can recover distinctions that matter for evaluating LLM-generated rationales, this study contributes evidence on the limits of semantic similarity as a proxy for rationale alignment in this setting. This question has direct implications for how validity evidence is constructed for LLM raters in formative assessment contexts.

2 Literature

Existing studies have examined the consistency between LLM and human scoring (e.g., Bui and Barrot, 2024; Pack et al., 2024), but few have explored the alignment between the reasoning processes underlying these scores. Hua et al. (2025) uses cosine similarity on rationale embeddings and principal component analysis to compare LLM and human rater reasoning in automated essay scoring. Hou et al. (2025) also goes beyond score comparison to examine evidence selection overlap between LLM and human raters in teacher questioning assessment, using character-level Intersection Over Union. These studies reflect a growing recognition that similarity-score-only comparisons do not tell the complete story, particularly in formative assessment contexts where the reasoning behind a rating matters for uses such as feedback generation. Specifically, the textual similarity metrics researchers have employed to examine rationale alignment may not capture the evaluative structure of the construct being measured.

Measures of text similarity have been applied across fields to compare and classify text, built on lexical and semantic matching between words and sentences (Corley and Mihalcea, 2005; Reimers and Gurevych, 2019). While these measures capture shared vocabulary and semantic proximity, they are insensitive to properties that matter when the comparison target is not topical overlap but reasoning alignment. Wu et al. (2024) provides evidence of this gap in the scoring context. They used a rubric-generation approach and found that LLMs reach the same scores as humans via different reasoning paths, suggesting that score agreement does not imply reasoning agreement. Organisciak et al. (2023) established the limits of semantic similarity in divergent thinking assessment, showing that similarity metrics capture surface-level textual proximity instead of the construct of interest, originality, and that stronger embedding models do not close this gap. Xu et al. (2024) also recognizes this limitation in medical text comparison and proposes a partial solution of using GPT-4-generated labels as an intermediate representation before computing similarity. This approach improves on raw similarity but still computes a symmetric proximity score on compressed representations.

We contribute to this literature by demonstrating that semantic similarity does not effectively capture alignment between LLM and human rationales in formative clinical reasoning assessment. In examining LLM-generated rationales in this context, we observed three patterns that similarity metrics cannot recover effectively: LLM rationales were considerably longer than human rationales, frequently elaborated on the same evaluative content as human raters but with qualifications in the opposite direction, and did not always cover the human rater’s central evaluative claim even when addressing the same topic. These patterns motivated an alternative solution using an LLM-as-rater, human-in-the-loop qualitative coding approach that extracts and compares the evaluative claims in LLM and human rationales directly, rather than just measuring their textual proximity.

Specifically, we address the following research questions:

RQ1: Do systematic differences in rationale length between LLM-generated and human-written explanations inflate or suppress measured similarity?

RQ2: Can semantic similarity scores distinguish different patterns of rationale alignment, including

evaluative direction and inclusion of the human rater’s central claim?

3 Data

The data come from a clinical reasoning assessment context in which 21 third-year medical students at a German medical school completed four AI-simulated patient encounters, scored on eight items of the Clinical Reasoning Indicators-History Taking Scale on a 1-5 Likert scale. Rationale comparison draws on two rater sources: one human rater at the postdoctoral level who produced written rationales for 216 instances of student-encounter-item for the first two patient encounters and 27 ChatGPT-5.4 LLM raters varying in model size (nano, mini, full), reasoning effort (none, medium, xhigh), and prompt condition (zero-shot, one-shot with example scores, and one-shot with example scores and rationales). Pairing the human rater’s rationales with the ChatGPT-5.4 rationales led to 5,790 valid comparison pairs.

A qualitative coding scheme was developed in a parent study (Lu et al., 2026) to categorize the rationale comparisons by the alignment of the main evaluative claims within the text. A two-stage human-calibrated LLM coding pipeline labels each rationale pair. The first stage assigns one of four labels based on whether the LLM rationale covers the human rater’s central evaluative claim: FULLY_ALIGNED, PARTIALLY_ALIGNED, NOT_ALIGNED, and UNCLEAR. The second stage refines PARTIALLY_ALIGNED pairs into three subcodes: not contained (LLM misses the human’s main point), EXTEND (LLM covers the human’s main point and adds content of the same valence), and OTHER_VALENCE (LLM covers the human’s main point and adds content of the opposite valence). The pipeline runs on ChatGPT-5.4 with no reasoning effort and was calibrated against two human-coded samples ($n = 100$ for the first stage, $n = 60$ for the second stage) coded independently by two human raters and reconciled to consensus. Inter-rater reliability between the two human raters was 79.0 percent agreement ($\kappa = 0.639$) for the Stage 1 labels and 66.7 percent agreement ($\kappa = 0.499$) for the Stage 2 labels.

4 Methods

To address the first research question, we computed word or token length distributions for the

human rationales versus ChatGPT-5.4 rationales across the corpus and computed similarity scores for all pairs using an embedding model, which transforms text into numerical values to measure the relatedness between two pieces of text. We used text-embedding-3-large from OpenAI. Prior work suggests that embedding-based similarity measures may be sensitive to differences in text length (Lee et al., 2024), which prompted us to examine the association between the absolute difference of the rationale lengths in each pair and the similarity scores, using a linear regression model with the type of LLM as a fixed effect.

To answer the second research question, we compared similarity score distributions across the four Stage 1 qualitative codes and within the PARTIALLY_ALIGNED rationale pairs, three Stage 2 qualitative subcodes. We fit linear regression models predicting similarity from qualitative category, with language, rationale length difference, and LLM rater fixed effects included as covariates, and standard errors clustered by observation unit.

5 Results

Table 1: Linear Regression Predicting Similarity Scores

	Similarity Score
Rationale Length Difference	-0.0029*** (0.000)
LLM Fixed Effects	Yes
Intercept	0.525*** (0.010)
Observations	5,790
R^2	0.113
Adjusted R^2	0.109
F Statistic	12.32

Note. Standard errors in parentheses. LLM fixed effects included but omitted from the table for brevity. *** $p < .001$.

First, we examined how the difference of the lengths of each LLM-human rationale pair was associated with semantic similarity. In approximately 89 percent of the rationale pairs, the LLM rationale had a greater number of words than the human rationale. The mean absolute difference of the number of words between the rationales in each pair is about 19 words ($SD = 10$). The results of the linear regression of the absolute difference of rationale length on similarity score with LLM rater fixed effects and cluster-robust standard errors show that

greater differences in rationale length were associated with lower similarity scores, controlling for LLM fixed effects ($\beta = -0.0029$, $p < .001$, see Table 1).

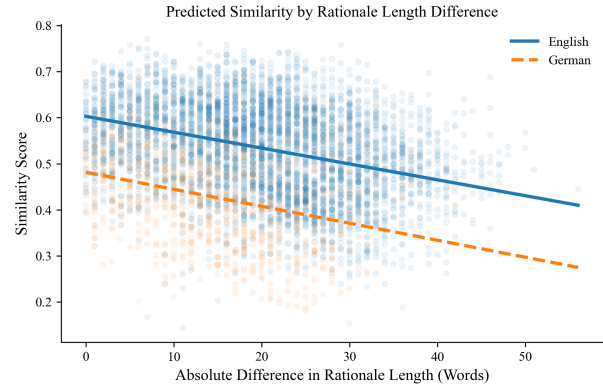


Figure 1: Predicted similarity as a function of rationale length difference, with language as an interaction. Lines show predicted values from a linear regression of similarity on rationale length difference (German vs. English), their interaction, and LLM rater fixed effects. Standard errors are clustered by observation unit.

Because 902 of the rationales that the LLMs generated were in German (15.6 percent of the total), we examined whether the language influenced the association between the rationale length differences and similarity scores by creating an interaction between the rationale length difference and whether the language was in German (see Figure 1). We found that rationale pairs with the LLM rationale in German showed significantly lower similarity scores than English rationale pairs, controlling for rationale length differences and LLM fixed effects, with standard errors clustered by observation unit ($b = -0.121$, $SE = 0.009$, $p < .001$). Specifically, these rationale pairs with German were associated with similarity scores that were approximately 0.12 points lower on average than English rationale pairs. However, the interaction between rationale length difference and language was not statistically significant ($b = -0.0002$, $SE < 0.001$, $p = .698$), indicating that the association between rationale length differences and similarity was comparable for German and English rationale pairs.

For our second research question, we examined how the similarity scores compare with our qualitative coding. There were 1,642 LLM-human rationale pairs labeled FULLY_ALIGNED, 3,319 labeled as PARTIALLY_ALIGNED, 828 labeled as NOT_ALIGNED, and one labeled as UN-

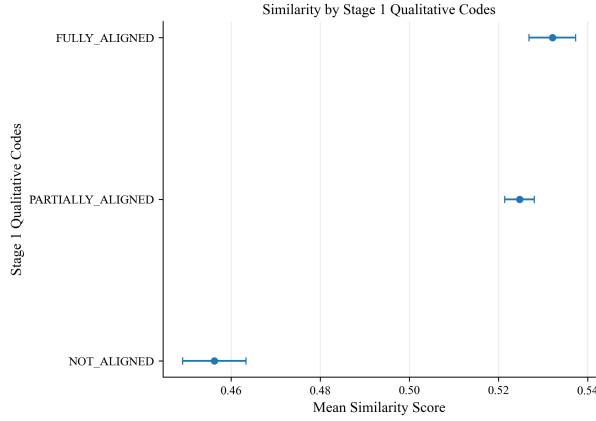


Figure 2: Mean similarity by first-stage qualitative code. Error bars represent 95% confidence intervals.

CLEAR, which was omitted. The mean similarity score was 0.532 for FULLY_ALIGNED, 0.525 for PARTIALLY_ALIGNED, and 0.456 for NOT_ALIGNED (see Figure 2).

Table 2: Pairwise Comparisons of Similarity Across Alignment Categories

Comparison	Mean Diff.	95% CI	<i>p</i>
FA – NA	0.064	[0.034, 0.094]	< .001
FA – PA	0.006	[-0.015, 0.027]	.576
PA – NA	0.058	[0.035, 0.081]	< .001

Note. FA = Fully Aligned; PA = Partially Aligned; NA = Not Aligned. Estimates from a linear model regressing semantic similarity on alignment category, language, rationale length difference, and LLM rater fixed effects, with standard errors clustered by observation unit.

After adjusting for language, rationale length difference, and LLM rater fixed effects, with clustered standard errors, similarity scores were lower for pairs classified as NOT_ALIGNED than for those classified as either FULLY_ALIGNED ($b = 0.064$, $p < .001$) or PARTIALLY_ALIGNED ($b = 0.058$, $p < .001$). FULLY_ALIGNED and PARTIALLY_ALIGNED pairs did not differ significantly from each other ($b = 0.006$, $p = .576$). (see Table 2).

For the Stage 2 qualitative codes (see Table 3 for distributions), after adjusting for language, rationale length difference, and LLM rater fixed effects, with clustered standard errors, none of the pairwise differences in similarity across the subcategories were significant (EXTEND vs. OTHER_VALENCE: $b = -0.017$, $p = .172$; EXTEND vs. not contained: $b = -0.006$, $p = .598$; OTHER_VALENCE vs. not contained: $b = 0.011$, $p = .286$).

Table 3: Similarity by Partially Aligned Categories

Category	<i>N</i>	Mean	95% CI
Not Contained	2,017	0.534	[0.530, 0.538]
Other Valence	950	0.512	[0.506, 0.519]
Extend	352	0.505	[0.495, 0.516]

Note. Means represent similarity scores within each second-stage category among responses classified as partially aligned. Confidence intervals are 95% confidence intervals around the mean.

6 Discussion

This study examined whether semantic similarity provides a meaningful basis for comparing LLM-generated and human-written rationales in formative clinical reasoning assessment. Our findings suggest that it does not, at least not on its own. One reason is that similarity scores were systematically related to differences in rationale length. In our data, LLM rationales were longer than human rationales in 89 percent of pairs, with a mean absolute length difference of 19 words, and larger length differences were associated with lower similarity scores ($\beta = -0.0029$, $p < .001$). This matters because lower similarity may reflect differences in verbosity or explanatory style rather than substantive disagreement in evaluative judgment.

A second issue is that semantic similarity was not sensitive to some of the distinctions that matter most in formative assessment. Similarity scores clearly separated rationale pairs coded as NOT_ALIGNED ($M = 0.456$) from those coded as FULLY_ALIGNED ($M = 0.532$) or PARTIALLY_ALIGNED ($M = 0.525$), but they were much less useful for distinguishing finer-grained differences in evaluative meaning. In particular, within the partially aligned cases, pairs coded as not contained ($M = 0.534$) could still receive similarity scores as high as or higher than pairs in which the LLM did preserve the human rater’s central claim but extended it ($M = 0.505$) or qualified it in a different evaluative direction ($M = 0.512$). In other words, semantic proximity did not reliably indicate whether the LLM rationale captured the human rater’s main point.

These findings matter for research on LLM use in education and assessment, especially in formative settings where the rationale attached to a score may influence how feedback is interpreted and used. A semantically similar rationale should not be assumed to be interchangeable with a human ratio-

nale, and a semantically dissimilar rationale should not automatically be treated as inferior. Our point is not that human rationales are inherently better than LLM rationales. Rather, the point is that semantic similarity alone is too limited a basis for claims about substitutability.

Taken together, the findings suggest that comparisons between LLM and human rationales need to attend more directly to evaluative meaning rather than relying on semantic proximity alone. In formative assessment contexts, the relevant question is not simply whether two rationales sound similar, but whether they support similar interpretations and decisions for the people who rely on them.

7 Limitations and Future Directions

This study has several limitations. First, the qualitative labels used as the comparison target are not objective ground truth. The coding framework was developed by the research team, and the LLM-based coding pipeline was calibrated on relatively small human-coded samples reconciled by two researchers. Our claim, therefore, is not that these categories represent a definitive truth about rationale alignment, but that semantic similarity did not reliably distinguish the categories used in this study. Future work should test whether the same pattern holds with alternative coding schemes, larger human-coded samples, and stronger forms of human validation.

Second, the study is limited in scope. The data come from 21 medical students in a single clinical reasoning simulation context at one institution, and the human rationale set was produced by one rater. Although the total number of rationale pairs is substantial, the underlying assessment setting is relatively narrow, which limits the generalizability of the findings across domains, assessment types, and rater populations. Future studies should examine whether these issues persist in other formative assessment contexts, with multiple human raters and more diverse samples.

Finally, our analysis used a single embedding-based cosine similarity measure. Although this reflects a common approach in current research, the findings may not generalize to other similarity metrics or comparison frameworks. In addition, a subset of the LLM rationales was generated in German, raising a further question about multilingual robustness in similarity-based comparisons. Future work should replicate these analyses using

alternative similarity measures and should examine more directly how multilingual variation affects the interpretation of rationale similarity.

References

- Ngoc My Bui and Jessie S. Barrot. 2024. Chatgpt as an automated essay scoring tool in the writing classrooms: How it compares with human scoring. *Education and Information Technologies*, 30:2041–2058.
- Courtney Corley and Rada Mihalcea. 2005. [Measuring the semantic similarity of texts](#). In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 13–18, Ann Arbor, Michigan. Association for Computational Linguistics.
- Ruikun Hou, Tim Fütterer, Babette Bühler, Patrick Schreyer, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2025. Llm-human alignment in evaluating teacher questioning practices: Beyond ratings to explanation. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers*, pages 239–249, Wyndham Grand Pittsburgh, Downtown, Pittsburgh, Pennsylvania, United States. National Council on Measurement in Education (NCME).
- Haowei Hua, Hong Jiao, and Dan Song. 2025. [Comparison of scoring rationales between large language models and human raters](#). *arXiv preprint arXiv:2509.23412*.
- Reagan J. Lee, Samarth Goel, and Kannan Ramchandran. 2024. [Quantifying positional biases in text embedding models](#). *arXiv preprint arXiv:2412.15241*.
- M. Lu, J. Jung, E. Brügge, M. Holling, and D. Darici. 2026. When LLMs do not score like human raters: Separating score alignment from rationale evidence in clinical reasoning assessment. Manuscript under review.
- Peter Organisciak, Selcuk Acar, Denis Dumas, and Kelly Berthiaume. 2023. [Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models](#). *Thinking Skills and Creativity*, 49:101356.
- Austin Pack, Alex Barrett, and Juan Escalante. 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.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Xuansheng Wu, Padmaja Pravin Saraf, Gyeonggeon Lee, Ehsan Latif, Ninghao Liu, and Xiaoming Zhai. 2024. [Unveiling scoring processes: Dissecting the differences between llms and human graders in automatic scoring](#). *arXiv preprint arXiv:2407.18328*.

Shaochen Xu, Zihao Wu, Huaqin Zhao, Peng Shu, Zhengliang Liu, Wenxiong Liao, Sheng Li, Andrea Sikora, Tianming Liu, and Xiang Li. 2024. [Reasoning before comparison: Llm-enhanced semantic similarity metrics for domain specialized text analysis](#). *arXiv preprint arXiv:2402.11398*.