# Comparative Evaluation of AI-Generated vs. Expert-written Answer Explanations for a Medical Education Self-Assessment

**Yiyun Zhou, Francis O'Donnell, and Victoria Yaneva**
National Board of Medical Examiners, Philadelphia, USA
{YYZhou, FODonnell, VYaneva}@nbme.org

## Abstract

Answer explanations for medical multiple-choice questions (MCQs) are a valuable learning tool, but producing them is resource-intensive. Writing high-quality explanations requires specialized medical expertise and careful alignment with the keyed answer, distractors, and the clinical vignette. This paper evaluates whether a template-aware, retrieval-guided large language model (LLM) workflow can support this production task in a real formative assessment setting. Using a 50-item medical education self-assessment, we compared AI-generated and expert-written MCQ explanations in a blinded study involving eight medical faculty and sixteen medical students. Each participant rated 25 of 50 paired explanations on clarity, amount of information, and structure. The clearest empirical difference was in amount of information: AI-generated explanations were rated significantly higher than expert-written explanations in a cumulative link mixed model analysis (OR $= 1.99$, 95% CI $[1.33, 2.99]$, $p = 0.001$). Ratings of clarity and structure did not differ significantly between conditions. Based on faculty ratings, a smaller proportion of AI-generated explanations were judged to require correction (20%) compared with expert-written explanations (38%). These findings suggest that AI can reduce first-draft authoring effort in explanation writing while still requiring expert review to ensure content accuracy.

## 1 Introduction

Multiple-choice questions (MCQs) are a central component of clinical education and assessment, yet their instructional potential often extends beyond the selection of a correct answer. In many clinical learning tools, MCQs are accompanied by narrative explanations that justify why the correct option is appropriate and why the alternative options are incorrect (see Appendix A for an example MCQ and explanation). These explanations typically outline the relevant clinical concepts, diagnostic reasoning, or guideline-based considerations that underlie the item, making the expert reasoning process more transparent to learners. By explicitly articulating the logic behind answer choices, answer explanations can help students connect factual knowledge with clinical decision-making, reinforce key principles, and clarify common misconceptions. As a result, they have the potential to transform MCQs from purely evaluative instruments into powerful learning tools that support the development of clinical reasoning.

While valuable as learning support instruments, writing answer explanations for medical self-assessment questions is resource-intensive, which limits the availability of such materials. This is largely due to the strict criteria these explanations must satisfy to be pedagogically useful. Each explanation must justify the correct answer, explain why distractors are incorrect, remain faithful to the clinical vignette, and conform to an editorial style that supports instruction (Ch'en et al., 2025). In practice, therefore, producing high-quality explanations involves more than clinical expertise alone; it also requires careful editing and review to ensure clarity, accuracy, and consistency.

Large language models (LLMs) offer a possible way to reduce this drafting burden. Prior work has shown strong LLM performance on medical question answering and medical explanation tasks (Singhal et al., 2023, 2025). However, most of this literature has focused on answer accuracy and explanation correctness on benchmark datasets rather than the production of pedagogically valuable explanation drafts in real assessment workflows (Kim et al., 2024; Alonso et al., 2024; Chen et al., 2025). That gap matters because educational explanation writing requires not only domain reasoning, but also adherence to instructional goals, editorial conventions, formatting requirements, and expert qual-



*Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications*, pages 455–462

ity control; it is also a workflow task shaped by editorial and quality control procedures.

This paper makes two main contributions. First, we develop a pipeline for generating answer explanations for clinical MCQs using AI. The pipeline is designed to produce explanations that address both the correct answer and the distractors while adhering to instructional conventions used in medical education materials. Second, we evaluate the quality of the generated explanations in a comparative study with expert-written explanations. Using blinded evaluations from 8 clinical faculty members and 16 medical students, we find that AI-generated explanations perform comparably to or better than expert-written explanations on clarity, amount of information, and structure.

## 2 Related Work

Prior work shows that LLMs can perform strongly on medical question answering (QA) benchmarks and that domain adaptation and prompting can further improve performance on medical QA and long-form medical responses, but this literature has largely emphasized answer accuracy rather than the quality, reliability, or educational usefulness of the accompanying explanations (Jin et al., 2021; Singhal et al., 2023, 2025; Chen et al., 2025). This motivates our focus on evaluating explanation quality directly rather than treating the explanations as a by-product of answer prediction.

Our method is also related to work on retrieval-augmented generation, in-context example retrieval, and post-generation refinement and verification. Retrieval can improve knowledge-intensive generation, and the choice of in-context examples can substantially affect downstream LLM performance (Lewis et al., 2020; Wang et al., 2024). Our expert-guided prompt iteration and post-generation checks are also informed by work on automatic prompt optimization, self-refinement, and verification-based decoding (Pryzant et al., 2023; Madaan et al., 2023; Dhuliawala et al., 2024). However, prior work does not directly address the question studied here: whether a retrieval- and prompt-based LLM workflow can generate answer explanations for real medical self-assessment items that are acceptable under expert review and comparable to expert-written explanations.

## 3 Dataset

The study used two data sets, each consisting of 50 MCQs, from a self-assessment program for medical students in the clinical stage of their training.

Set 1 served as the reference set and included MCQs with corresponding explanations that were used in the system prompts. Set 2 was the experimental set. Although it also contained expert-written explanations, these were not used during AI generation; instead, they were reserved for the later human evaluation, where AI-generated explanations were compared with the expert-written explanations for Set 2.

In addition to the two sets of items and explanations, two internally developed explanation templates reflecting existing editorial standards were provided – a *General* template for areas like basic science and biostatistics, and a *Clinical* template for areas that focus on patient management workflows. These templates guided the expected organization and level of detail in the generated explanations.

Each item within the sets included three distinct metadata tags: *Content Area* - the specific medical topic, coded hierarchically, e.g., `Multi: acid-base disorders`; *Competency Group* - the targeted physician competency, e.g., `B2 Laboratory/diagnostic studies`; and *Template Type* - the explanation template structure as either General or Clinical, determined by an LLM. There were 49 content areas and 13 competency groups represented in the dataset.

## 4 Methodology

When creating answer explanations for MCQs, we face three main challenges.

First, different explanations need to follow different templates depending on their content, as explained in the previous section. Therefore, one of our first tasks given a new item is to perform a binary classification step to correctly classify it into the template most appropriate for its content. Our approach to this task is described in section 4.1. Next, we need to determine the most appropriate items from our reference set that can be used for few-shot prompting. This task is challenging, because the items in the reference set cover multiple topics, content areas and tasks, and relevance can be determined across multiple dimensions. Our approach to this task is described in section 4.2. Finally, the generation is performed using an agentic

structure with *Expert-in-the-Loop* (EITL) calibration, as described in Section 4.3.

Figure 1 provides an overview of the approach and the following sections describe each step.

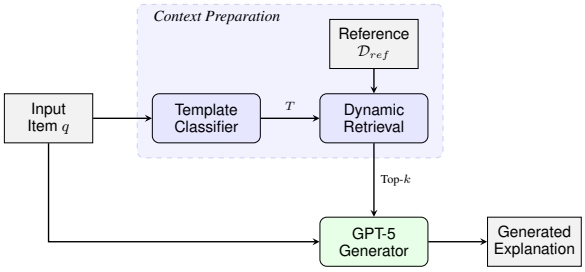


Figure 1: The workflow Architecture. Input items are routed via a classifier to retrieve context, which is then fed alongside the original input into the generator.

### 4.1 Dataset Stratification and Template Classification

To optimize the generation process for domain specificity, we employ a schema-aware preprocessing strategy that segments the reference corpus (Set 1), $\mathcal{D}_{ref}$, into distinct reasoning templates - *Clinical* template and *General* template. Correctly routing input items to these templates is critical for correctly organizing the information in the explanations. We implement a hybrid classification pipeline that processes each target item $q$ to assign a template category $T \in \{T_{clinical}, T_{general}\}$. The primary classification is performed by a large language model (GPT-5) which analyzes the semantic content and the metadata tags of the item. To reduce misclassification, we also used a deterministic keyword-based fallback for items with uncertain template assignments, like in domains such as biostatistics and epidemiology. This hybrid approach significantly reduces noise in specialized domains where purely semantic vector matching may struggle to distinguish between subtle subcompetencies.

### 4.2 Hierarchical Dynamic Few-Shot Retrieval

Because LLMs outputs can be sensitive to the choice of in-context examples, fixed few-shot prompts may be poorly aligned when a query item differs in subdomain, competency, or explanation format. Prior work has shown that retrieval and example selection can substantially affect downstream LLM performance.To address this, we dynamically retrieve a small set of exemplars for each query item $q$ from a reference pool $\mathcal{D}_{ref}$. Rather than relying only on semantic similarity, we rank candidates using a hierarchical relevance scheme that first prioritizes metadata alignment and then uses textual overlap as a secondary signal.

Specifically, the ranking favors exact matches on the full Content Area code, then broader partial matches on the Content Area prefix, followed by matches on Competency Group and Template type. Lexical overlap between the query and candidate item is used to refine the ordering among otherwise suitable examples. This strategy encourages the retrieved exemplars to match not only the topic of the item but also its instructional framing and expected explanation style. We then insert the top-$k$ examples (with $k = 3$) into the prompt, yielding a query-specific few-shot context that is more stable and better aligned than a static prompt.

### 4.3 Iterative Optimization and Generation

The generation phase employs GPT-5 as the backbone reasoning engine. Prior to large-scale inference, we perform an Expert-in-the-Loop calibration phase to align the model's output with the specific pedagogical voice of the assessment. This process treats prompt engineering as a discrete optimization problem (Pryzant et al., 2023), where the objective function is the satisfaction score of a human Subject Matter Expert (SME). In our case, the SME was an editor expert with no clinical background but with experience in editing MCQ explanations for clarity and consistency. In each iteration $t$, the system generates a batch of explanations $R_t$ using prompt $P_t$. The SME evaluates these outputs for consistency, format, style guidelines, and presentation logic, providing unstructured natural language feedback $F_t$. We utilize a meta-prompting strategy to inject this feedback into the prompt instructions, deriving a refined prompt $P_{t+1}$. This iterative cycle continues until the SME confirms that the outputs consistently meets the required standard without further intervention, ensuring the instructions generalize well to unseen data (Shah, 2025). Following prompt calibration, the system executes the inference pipeline. To mitigate the risk of hallucinations—a critical failure mode in medical QA—we implement a deterministic self-correction loop. The system parses the generated output to verify that every distractor option is explicitly addressed and logically refuted. If the parser detects missing components or structural inconsistencies, an automated "nudge" prompt triggers a regeneration pass. This iterative refinement

helps make the final output more structurally complete and better aligned with the target explanation style (Madaan et al., 2023) . Medical consistency was assessed separately through expert review and the blinded comparative study.

## 5 Study Design and Findings

We evaluated the 50 AI-generated explanations through an external rating study by comparing them to the 50 existing expert-written explanations for corresponding MCQs in Set 2.

The study used a balanced incomplete block design implemented in Qualtrics. For each Set 2 MCQ, participants evaluated two explanations: one AI-generated explanation and one expert-written explanation. The source of each explanation was blinded. Each participant rated explanations for 25 of the 50 MCQs. Participants included 8 medical faculty members and 16 third-year medical students. In the faculty group, each MCQ was evaluated by four faculty members; in the student group, each MCQ was evaluated by seven to nine students.

After reading an explanation, participants rated its *clarity*, *amount of information*, and *structure* on a three-point scale (Poor, Fair, Good). Faculty also indicated whether an explanation needed correction, and both groups could provide optional open-ended comments. For inferential analysis, we fit separate cumulative link mixed models for the three rated dimensions with fixed effects for explanation type, rater group, and their interaction, plus random intercepts for rater and question.

Table 1 shows that average ratings were high for both sources, especially for clarity and structure, where scores clustered near the upper end of the three-point scale. This ceiling effect helps explain why source differences on those dimensions remained small. Even so, AI explanations received slightly higher mean ratings than expert-written explanations in all six source-by-group comparisons. Median-based summaries pointed in the same direction (see Table 2).

The clearest advantage for AI appeared in the *amount-of-information* dimension. A cumulative link mixed model estimated that AI explanations were nearly twice as likely as expert-written explanations to receive a higher rating ($\mathrm{OR} = 1.99$, 95% CI $[1.33, 2.99]$, $p = 0.001$). Neither rater group nor the interaction between rater group and explanation type was statistically significant, suggesting that faculty and students did not differ systematically in their rating patterns and that the relative advantage of AI explanations was similar across both groups.

Lastly, faculty flagged fewer AI explanations than expert-written explanations for needing corrections: 10 of 50 AI explanations (20%) versus 19 of 50 expert-written explanations (38%). Most flagged cases were identified by only one reviewer, suggesting subjectivity.

## 6 Conclusion

This study demonstrates that a retrieval-guided LLM pipeline can generate explanations for MCQs that are comparable to expert-authored explanations under blinded human evaluation. Across both faculty and student raters, AI explanations matched expert explanations on clarity and structure and were rated significantly higher on the amount-of-information dimension. These results suggest that LLM-assisted workflows can meaningfully support the production of MCQ explanations in medicine. A strength of this study is its inclusion of both clinical faculty and medical students, allowing explanation quality to be assessed from the perspectives of expert reviewers and intended learners.

However, even highly rated AI explanations occasionally included verbosity, stylistic inconsistencies, or unsupported clinical details, necessitating expert review. Rather than investing effort in drafting explanations from scratch, experts can focus on review, calibration, and targeted revision to ensure clinical accuracy and pedagogical quality. More broadly, these results point to a practical role for AI in assessment development pipelines: accelerating the generation of high-quality first drafts while preserving expert oversight as the central mechanism for maintaining reliability and trust.

## 7 Limitations and Ethical Considerations

This study has several limitations. First, the evaluation was conducted on a single set of medical self-assessment explanations from one assessment context and therefore reflects one content domain and one editorial style rather than a fuller range of medical assessment materials. Second, the external review sample was modest (8 faculty and 16 third-year medical students, all U.S.-based), and each participant rated only a subset of explanation pairs under a balanced incomplete block design. Third, the three-point Poor/Fair/Good scale reduced re-

| | Students | | Faculty | |
|---|---|---|---|---|
| **Dimension** | **Expert** | **AI** | **Expert** | **AI** |
| Clarity | 2.82 (0.44) | 2.85 (0.41) | 2.79 (0.46) | 2.86 (0.38) |
| Amount of information | 2.32 (0.70) | 2.50 (0.66) | 2.46 (0.57) | 2.62 (0.63) |
| Structure | 2.78 (0.46) | 2.81 (0.44) | 2.75 (0.51) | 2.82 (0.41) |

Table 1: Mean (SD) ratings by rater group and explanation source.

| | Students | | Faculty | |
|---|---|---|---|---|
| **Dimension** | **Expert** | **AI** | **Expert** | **AI** |
| Clarity | 3 (0) | 3 (0) | 3 (0) | 3 (0) |
| Amount of information | 2 (1) | 3 (1) | 2 (1) | 3 (1) |
| Structure | 3 (0) | 3 (0) | 3 (0) | 3 (0) |

Table 2: Median and interquartile range of medians across explanations.

spondent burden for a lengthy rating task, but it also compressed variation and likely contributed to ceiling effects on clarity and structure.

From an ethical and operational perspective, we frame this work as AI-assisted content generation within a human-reviewed editorial process, not as replacement of human expertise. In this study, generated explanations remained subject to editor screening and expert review for clarity, alignment with the vignette, medical consistency, and bias. The need for this oversight is underscored by observed issues such as redundancy, imprecise wording, and occasional content misrepresentation in generated outputs. Because the study involves proprietary assessment materials and externally authored comparison explanations, we intentionally abstract some operational details and refer to external contributors generically. More broadly, the goal of this work is not to suggest that explanation authoring can be fully automated, but to evaluate whether retrieval-guided generation can support a more efficient and responsible human-centered authoring workflow for educational content.

Bias is another important reason for human oversight. Language models trained on large corpora may reflect biases related to patient demographics, disease prevalence, or clinical assumptions. If such biases appear in explanations, they could reinforce stereotypes or present an unbalanced view of clinical practice. Systematic evaluation for bias and careful editorial review are therefore necessary components of responsible deployment.

Transparency and accountability are also important in assessment contexts. Organizations using AI-assisted workflows must determine how to communicate the role of AI in content development and ensure that responsibility for the final content remains clearly assigned to human experts. More broadly, AI systems should be positioned as tools that support expert authorship rather than replace it, helping to reduce drafting effort while preserving expert oversight and responsibility for educational quality.

## References


Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Artificial intelligence in medicine*, 155:102938.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599.

Peter Y Ch'en, Wesley Day, Ryan C Pekson, Juan Barrientos, William B Burton, Allison B Ludwig, Sunit P Jariwala, and Todd Cassese. 2025. Gpt-4 generated answer rationales to multiple choice assessment questions in undergraduate medical education. *BMC Medical Education*, 25(1):333.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*, pages 3563–3578.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Yunsoo Kim, Jinge Wu, Yusuf Abdulle, and Honghan Wu. 2024. Medexqa: Medical question answering benchmark with multiple explanations. In *Proceedings of the 23rd Workshop on biomedical natural language processing*, pages 167–181.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594.

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with" gradient descent" and beam search. *arXiv preprint arXiv:2305.03495*.

Chirag Shah. 2025. From prompt engineering to prompt science with humans in the loop. *Communications of the ACM*, 68(6):54–61.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature medicine*, 31(3):943–950.

Liang Wang, Nan Yang, and Furu Wei. 2024. Learning to retrieve in-context examples for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1752–1767.

# A Sample Item and Explanation

## A.1 Item

A 58-year-old woman is brought to the emergency department 4 hours after the sudden onset of left-sided facial, arm, and leg weakness. She has hypertension treated with a diuretic. She reports the sensation of a fluttering heart for 1 week. Her pulse is 125/min and irregular, and blood pressure is 135/80 mm Hg. Physical examination shows a left-sided facial droop. There is flaccid weakness of the left upper and lower extremities. An ECG shows a rapid, irregular ventricular rate with no discernible P waves. Which of the following therapies would best improve function of the weakened muscles over the next 3 hours?

**A.** Acetylcholine agonist
**B.** Acetylcholine antagonist
**C.** Antifibrinolytic
**D.** Fibrinolytic
**E.** Procoagulant

**Correct Answer.** D.

## A.2 Explanation

This patient with a likely thromboembolic cerebral infarction (stroke) from atrial fibrillation would benefit from fibrinolytic therapy. Strokes occur due to ischemic or hemorrhagic loss of blood supply to the brain and manifest as focal neurological deficits related to the dysfunction of the affected brain region. Approximately 80–85% of strokes are ischemic, commonly arising from thromboembolic disease, for which atrial fibrillation is a major risk factor. Atrial fibrillation typically presents with palpitations, fatigue, lightheadedness, and mild dyspnea if symptomatic. On ECG, atrial fibrillation demonstrates irregularly irregular RR intervals without discernible P waves. Prolonged atrial fibrillation leads to left atrial hemostasis and increases the risk of thrombosis. A thrombus from this patient's left atrium likely embolized to the right internal carotid artery supplying the right precentral gyrus, resulting in left-sided hemiparesis. Alteplase (tissue plasminogen activator) is a fibrinolytic medication utilized in ischemic strokeso that binds to fibrin clots and converts plasminogen to plasmin, which lyses clots. If given early (generally within 3–4.5 hours after onset of the event), alteplase may promote neurological recovery.

**Incorrect Answers: A, B, C, and E.**

Acetylcholine agonists (Choice A) and antagonists (Choice B) are unlikely to be helpful in lysing this patient's thromboembolic clot, as the interaction of acetylcholine with the coagulation cascade is poorly defined. Acetylcholine agonists (eg, bethanechol) may improve cognition in patients with dementia, decrease heart rate, improve gut peristalsis, increase bladder contraction, and increase exocrine gland secretions. Acetylcholine antagonists (eg, benztropine) typically increase heart rate, decrease gut and bladder activity, and worsen cognitive function. They also act on muscle at the motor end-plate; however, this patient's weakness results from central nervous system dysfunction without pathology at the muscle fiber itself.

Antifibrinolytic (Choice C) therapy (eg, tranexamic acid) displaces plasminogen from fibrin clots to promote hemostasis during intraoperative bleeding, heavy menstrual bleeding, and traumatic hemorrhage. Procoagulant (Choice E) therapy (eg, protamine, coagulation factors) increases activation of the coagulation cascade. Both antifibrinolytic and procoagulant therapy would not lyse this patient's clot and may lead to further thrombosis.

**Educational Objective.** Patients with atrial fibrillation are at risk for thromboembolism due to left atrial hemostasis, which may result in ischemic strokes. Strokes manifest as focal neurological deficits related to the dysfunction of the affected brain region. Ischemic strokes are treated with fibrinolytic therapy, which promotes neurological recovery.

## B Additional Statistical Tables

| | | Students | | | Faculty | | |
|---|---|---|---|---|---|---|---|
| **Dimension** | **Source** | **Poor** | **Fair** | **Good** | **Poor** | **Fair** | **Good** |
| Clarity | Expert | 2.0 | 14.5 | 83.5 | 2.0 | 17.5 | 80.5 |
| | AI | 2.0 | 11.2 | 86.8 | 1.0 | 12.5 | 86.5 |
| Amount of information | Expert | 13.8 | 40.5 | 45.8 | 4.0 | 46.5 | 49.5 |
| | AI | 9.0 | 32.5 | 58.5 | 8.0 | 22.5 | 69.5 |
| Structure | Expert | 1.8 | 18.8 | 79.5 | 3.5 | 18.0 | 78.5 |
| | AI | 2.0 | 15.0 | 83.0 | 1.0 | 16.5 | 82.5 |

Table 3: Percentage of explanations by median rating, dimension, and source.

| **Predictor** | **Odds Ratio** | **95% CI** | $p$ |
|---|---|---|---|
| Explanation type [AI] | 1.64 | 0.94–2.86 | 0.084 |
| Rater group [Student] | 1.22 | 0.50–2.99 | 0.669 |
| Type × rater group | 0.80 | 0.40–1.59 | 0.525 |
| $\tau_{00}$ Questions | | 0.05 | |
| $\tau_{00}$ Raters | | 0.79 | |
| Marginal / Conditional $R^2$ | | 0.008 / 0.209 | |

Table 4: Cumulative link mixed model for clarity.

| **Predictor** | **Odds Ratio** | **95% CI** | $p$ |
|---|---|---|---|
| Explanation type [AI] | 1.99 | 1.33–2.99 | 0.001 |
| Rater group [Student] | 0.73 | 0.43–1.24 | 0.242 |
| Type × rater group | 0.84 | 0.51–1.37 | 0.478 |
| $\tau_{00}$ Questions | | 0.00 | |
| $\tau_{00}$ Raters | | 0.24 | |
| Marginal / Conditional $R^2$ | | 0.035 / NA | |

Table 5: Cumulative link mixed model for amount of information.

| **Predictor** | **Odds Ratio** | **95% CI** | $p$ |
|---|---|---|---|
| Explanation type [AI] | 1.48 | 0.85–2.57 | 0.168 |
| Rater group [Student] | 0.96 | 0.32–2.84 | 0.937 |
| Type × rater group | 0.86 | 0.44–1.69 | 0.670 |
| $\tau_{00}$ Questions | | 0.19 | |
| $\tau_{00}$ Raters | | 1.27 | |
| Marginal / Conditional $R^2$ | | 0.005 / 0.312 | |

Table 6: Cumulative link mixed model for Structure.

| **Error pattern** | **Questions** | **Percentage** |
|---|---|---|
| No errors in either explanation | 26 | 52% |
| Only expert explanation had errors | 14 | 28% |
| Only AI explanation had errors | 5 | 10% |
| Both explanations had errors | 5 | 10% |
| Total | 50 | 100% |

Table 7: Question-level error pattern based on faculty flags.