Judging the Judges: Using AI and Humans to Evaluate LLM Explanations

Anduo Wang amw043@ucsd.edu
Jessica Zhang jez004@ucsd.edu
Haoyang Yu hay034@ucsd.edu
Jessie Zhang jiz199@ucsd.edu
Ryan Lingo ryan_lingo@honda-ri.com
Rajeev Chhajer rajeev_chhajer@honda-ri.com

1 Context

As AI systems become more capable, people increasingly rely on Large Language Models (LLMs) not just for answers but for explanations. Students use AI to understand difficult topics, professionals use AI-generated reasoning to support decisions, and everyday users depend on explanations to judge whether an AI's answer is trustworthy. However, different users have different needs: a middle-school student, a college learner, a working professional, and an expert researcher all require explanations at different levels of depth and clarity. This makes prompt design critically important. The way we ask an LLM to explain something can dramatically change how understandable, accurate, or helpful the explanation is. Our project aims to evaluate how different prompting strategies and LLM-generated explanations align with real human preferences, particularly for college-level learners studying technical subjects such as statistics, machine learning, and data science. By building and testing this evaluation pipeline over 10 weeks, we hope to understand what makes an explanation not only accurate but genuinely helpful, trustworthy, and aligned with how students actually learn abstract concepts.

2 Problem Statement for Domain Expert

2.1 Overview

Given a set of difficult concepts $C = \{c_1, \ldots, c_n\}$, our goal is to evaluate the quality of explanations produced by multiple Large Language Models (LLMs) under different prompting strategies. Each concept is explained at three levels of depth (baseline, level 2, level 3). Specifically, we operationalize each level by prompt attributes such as constraint density (number of instructions), audience specification, and explanation style control (e.g., examples required, analogy allowed, persona desired). Then, the explanations are evaluated by both humans and LLM-based judges. Our aim is to quantify how well automated evaluation methods approximate human preferences, identify disagreement patterns, and determine

whether improved prompting leads to measurably better explanations.

In short, our primary research goal is to measure how strongly prompt design influences explanation quality for college-level learners, using human judgments as the ground truth of understandability.

2.2 Problem Definition

For each concept c_i, model m_j, and explanation level $\ell \in \{\text{baseline}, 2, 3\}$, an LLM generates an explanation

$$e_{i,j,\ell} = f_{m_j}(c_i, \ell).$$

Each explanation is scored along three main dimensions: explainability, complexity, and familiarity, using two sources.

(1) Human Evaluation. Human annotators assign rubric-based scores

$$h(e_{i,j,\ell}) \in \mathbb{R}^3,$$

corresponding to the three evaluation dimensions.

(2) LLM-as-Judge Evaluation. An evaluator model performs pairwise comparison between two explanations e_a and e_b:

$$g_{\text{judge}}(e_a, e_b) \to \{a, b\},$$

which are then aggregated into Elo ratings

$$\text{Elo}(e_{i,j,\ell}).$$

Additionally, a critique model generates feedback on each explanation:

$$r_{i,j,\ell} = g_{\text{critique}}(e_{i,j,\ell}),$$

which is used to guide prompt refinement and evaluate whether LLM-generated critiques lead to measurable improvement in explanation quality.

2.3 Research Questions

The project investigates several key questions:

1. **Prompt-Level Performance:** Do more structured prompting strategies consistently produce better explanations? That is, does the ordering

   $$\text{Elo}(e_{i,j,3}) > \text{Elo}(e_{i,j,2}) > \text{Elo}(e_{i,j,\text{baseline}})$$

   hold across concepts and models?

2. **Human–LLM Alignment:** To what extent do LLM-based evaluation scores correlate with human judgments of explanation quality?
3. **Prompt Optimization Impact:** Does LLM-generated critique and prompt rewriting measurably improve explanation quality across concepts?

2.4 Relation to Prior Work

Existing benchmarks for large language models tend to fall into two broad categories, each with important limitations for evaluating explanation quality. Knowledge-oriented benchmarks such as MMLU(Hendrycks et al. 2020) and MT-Bench(Zheng et al. 2023) primarily measure a model's factual recall or problem-solving ability using fixed question sets. While valuable, these benchmarks provide little insight into how readable or understandable a model's explanations are for users with different ages, backgrounds, or expertise levels. They evaluate correctness, but not the usability or pedagogical quality of the generated output.

On the other hand, human-preference frameworks such as Chatbot Arena focus heavily on which responses humans prefer in pairwise comparisons. Although these systems capture real user preferences, they do not explicitly evaluate pedagogical clarity or whether an explanation supports learning outcomes. As a result, models that produce persuasive but potentially incorrect explanations may still rank highly.

Our project aims to address this gap by constructing a benchmark centered specifically on explanation quality for professional or abstract concepts. We generate explanations using prompt templates with varying levels of constraint (baseline, intermediate, and highly structured prompts) and evaluate these outputs using both human judgments and LLM-as-judge pairwise comparisons. The pairwise preferences are aggregated through Elo scoring to produce a ranking of which prompting strategies yield explanations that humans are most likely to find helpful and comprehensible. This combined knowledge–preference approach allows us to evaluate not only whether an explanation is correct, but also whether it is accessible to diverse users and tailored to their needs.

2.5 Data Collection

We construct our concept dataset from publicly available domain glossaries, including Wikipedia glossaries in Artificial Intelligence, Computer Science, and Statistics Wikipedia contributors (2025a,b,c), which provide broad coverage of terminology across AI, statistics, and data science.

From these sources, we collect several hundred candidate concepts and use an LLM as a difficulty classifier to identify terms that require abstract reasoning, statistical intuition, or cross-domain understanding. Each concept is assigned a difficulty score, and only the highest-ranked subset is retained for evaluation. This filtering step ensures that our benchmark focuses on concepts where explanation quality is most informative and where differences between prompt strategies and models are more likely to emerge.

Human evaluation will primarily be conducted by the four project team members, all of whom have completed college-level coursework in statistics or machine learning. Each explanation will be rated using a fixed rubric covering explainability, complexity, and familiarity. While the number of raters is limited, this design allows for a controlled pilot study focused on rubric refinement and human–LLM alignment analysis. If feasible, we will recruit additional student evaluators to validate the reliability of our results.

This project is feasible within the timeline because both the data pipeline and the core evaluation components have already been prototyped in Quarter 1. Specifically, we have implemented those for a smaller setting: selecting concepts from the three glossaries, generating explanations under multiple prompt templates, conducting pairwise LLM-as-judge comparisons with reverse-order controls, and aggregating results using Elo scoring to construct prompt rankings. These experiments were successfully conducted on 30 concepts (10 per domain), producing 90 explanations and 540 pairwise judgments, with stable and interpretable Elo rankings across prompt types.

Scaling this study is practical. The glossaries contain hundreds of additional terms, eliminating the need for new data collection. Expanding from 30 to a few hundred concepts increases API usage linearly but remains computationally manageable, as the system relies solely on hosted LLM APIs rather than model training. Our pilot experiments also confirm that runtime and costs are well within feasible bounds.

3 Primary Output

The primary output of this project will be a comprehensive research report accompanied by an interactive website that presents our evaluation results. The report will document the full methodology, including explanation generation, human evaluation procedures, LLM-as-judge pairwise comparison, Elo rating computation, and statistical analyses of agreement and disagreement between humans and models. It will also include qualitative analyses of failure modes, examples of misleading or unclear explanations, and case studies of prompt refinement based on LLM critique.

In addition to the written report, we will develop an interactive leaderboard and visualization dashboard that communicates model performance across explanation levels, prompting strategies, and concept categories. This website will display representative explanations, human and LLM evaluation metrics, Elo-based rankings, and disagreement heatmaps. These visualizations will enable users to examine how different models respond to various prompt types and how their explanations differ in clarity, accuracy, and usefulness. More importantly, they directly support our research questions by revealing (1) how prompt depth influences explanation quality, (2) where human and LLM-judge evaluations diverge, and (3) whether prompt optimization meaningfully improves understandability.

References

Hendrycks, Dan, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. "Measuring massive multitask language understanding." *arXiv preprint arXiv:2009.03300*

Wikipedia contributors. 2025a. "Glossary of Artificial Intelligence." https://en.wikipedia.org/wiki/Glossary_of_artificial_intelligence

Wikipedia contributors. 2025b. "Glossary of Computer Science." https://en.wikipedia.org/wiki/Glossary_of_computer_science

Wikipedia contributors. 2025c. "Glossary of Statistics." https://en.wikipedia.org/wiki/Glossary_of_statistics

Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing et al. 2023. "Judging llm-as-a-judge with mt-bench and chatbot arena." *Advances in neural information processing systems* 36 : 46595–46623

Appendices