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

Jessie Zhang jiz199@ucsd.edu
Haoyang Yu hay034@ucsd.edu
Jessica Zhang jez004@ucsd.edu
Anduo Wang anw043@ucsd.edu
Rajeev Chhajer rajeev_chhajer@honda-ri.com
Ryan Lingo ryan_lingo@honda-ri.com

Abstract

With the rapid development of large language models (LLMs), existing benchmarks evaluate their knowledge and robustness in various professional fields. However, the ability of such systems to understand and correctly respond to human instructions is equally crucial, as text-based interaction remains the most direct way for humans to communicate with LLMs. Ensuring that these systems interpret human commands and return output as expected is a challenge in both daily-use and research conditions. In this paper, we introduce a realistic and robust approach to evaluate systems by using both LLM-as-Judge and Human-Judge under different levels of instruction prompts to analyze LLMs' performance in understanding human needs based on both LLMs' judgment and realistic human decisions. We involve multiple top professional-level vocabulary datasets in various domains that were chosen by LLM, and well-designed levels of instruction commands, followed by different conditions and settings. As an outcome, we present a comprehensive leaderboard that integrates both LLM-based and human evaluations, offering insights into performance rankings as well as the underlying reasons behind system behavior.

Code: https://github.com/Vica1106/Judging-the-Judges

1 Introduction

In the fast-developing area of artificial intelligence, two primary evaluation methods are commonly used to assess system behavior and performance: LLM-as-Judge and Human-Judge. These approaches reflect different perspectives, one automated and scalable, the other grounded in human intuition. Our research focuses on analyzing the model's generalization differences under multiple levels of prompts using both judging paradigms to better understand how models interpret and respond to varied human instructions.

LLM-as-Judge leverages one or more large language models to evaluate specified content, offering consistent, knowledge-based, and scalable judgments derived from extensive training data. In contrast, Human-Judge evaluation relies on human perception and contextual understanding, capturing subtle nuances and reasoning patterns that models may overlook. Together, these two methods provide a balanced framework for analyzing model performance from both computational and human-centric perspectives.

Our research aims to discuss a prompt-writing strategy that improves human interaction with LLMs on a simple and time-cost-effective function and analysis on the difference between LLMs' judgment and Human choice in multiple prompt layers. Specifically, we will use LLM-as-judge to choose professional-level vocabulary datasets from multiple domains to ensure variance. We will design different levels of instruction prompts, from simple and direct to restricted and complicated, to verify the difference in the model responses generated under different prompts. This approach will provide insights into the model's generation and explanation ability for extremely difficult content and the analysis model's generation under different prompts. Human evaluation leaderboard will be conducted after collecting the responses and the leaderboard of the LLM-as-judge system. We will analyze the consistency between Human evaluations and LLM judgments.

The project's objective is to enhance LLMs' robustness and generalizability by comparing the response consistency in the choice of the best prompt between LLMs and humans. This approach will provide a view of the gaps between LLMs' understanding of human needs and human preferences for needs.

2 Related Work

Traditional benchmarks such as MMLU Hendrycks et al. (2020) and HELM Liang et al. (2022) emphasize correctness, factual coverage, and reasoning ability. These static benchmarks answer whether a model knows the right information but not whether it can communicate that knowledge effectively at different levels to college students. Interactive evaluation platforms such as Chatbot Arena Chiang et al. (2024) have incorporated human preferences or model judges to evaluate answer quality in open-ended questions. However, those evaluation rubrics primarily focus on factual alignment and overall coherence, rather than explicitly assessing how well a model structures explanations for learners.

While Large Language Models (LLMs) have shown impressive capacity to generate expla-

nations of complex or abstract concepts across statistics, computer science, and artificial intelligence, assessing the quality of such explanations remains a significant challenge. Recent work suggests that current LLM evaluations show instability because they rely on fixed prompts; Mizrahi et al. (2024) argues that single-prompt evaluations are unreliable, as small changes in wording and phrasing can significantly alter model performance. They advocate for multi-prompt evaluation, where performance is aggregated across diverse prompt variations. He et al. (2024) similarly investigates how format and structure influence model behavior, showing that presentation framing can alter reasoning performance even when semantic content is identical.

Moreover, human evaluation remains subjective and difficult to scale. Shankar et al. (2024) identify a phenomenon called "criteria drift," in which human evaluators' standards shift over time as they interact with LLM outputs. Their findings emphasize the need for iterative calibration rather than assuming a static notion of correctness or quality.

Prior work has also explored LLMs as explanatory systems. The ELI5 dataset Fan et al. (2019) was an early attempt to test whether models could simplify complex topics for non-experts, helping establish "explanation clarity" as a measurable capability distinct from reasoning accuracy.

Together, these findings reveal two critical gaps. First, prompt framing and structure have not been systematically studied in the context of conceptual explanation for non-experts. Second, existing evaluation frameworks lack consistent alignment between human and LLM judgments. Building on these insights, we aim to develop a contextual explanation evaluation benchmark focused on college learners, examining how explanation quality varies across prompts targeting different levels of abstraction.

3 Methods

3.1 Pipeline Overview

To contextualize our evaluation workflow, we present the full system pipeline in Figure 1. The pipeline outlines each major stage of the process, beginning with concept selection from domain glossaries, followed by multi-level prompt construction and LLM-based explanation generation. Next, the system performs pairwise LLM-as-Judge comparisons, applies reverse-order and retry mechanisms to reduce bias, and aggregates outcomes using Elo scoring to produce a ranking of prompt quality. This modular design allows components—such as prompt templates, judging models, or scoring mechanisms—to be swapped or extended without changing the overall architecture.

3.2 Experimental Setup

To ensure consistency across all stages of the pipeline, we use a single model family "`GPT-5-Nano`" for both explanation generation and automated evaluation. Using the same lightweight

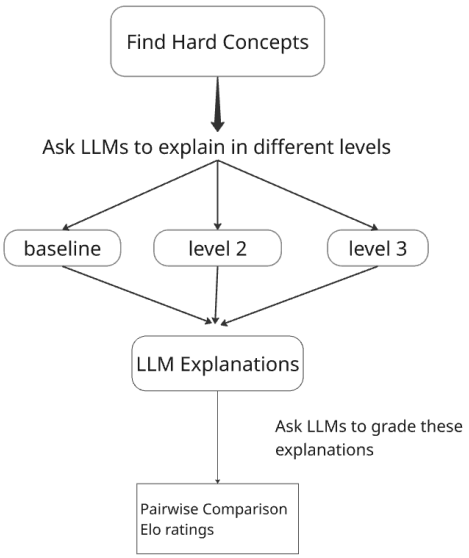

Figure 1: Full pipeline of our evaluation framework, including concept selection, structured prompt generation, LLM explanation generation, pairwise LLM judgment, Elo score aggregation, and final leaderboard construction.

model architecture allows us to isolate the effect of prompt structure without introducing additional variability from model scale or training differences.

Generation Model. All explanations are generated using `GPT-5-Nano` with a fixed decoding configuration (temperature = 1, top-p = 0.95). This setup encourages moderately diverse outputs while ensuring that explanations remain concise and coherent, enabling fair comparison across prompt templates.

LLM-as-Judge Model. We also use `GPT-5-Nano` as the evaluator in the pairwise comparison stage. Although larger models may yield more stable or human-aligned judgments, using a lightweight model helps surface differences caused specifically by prompt structure rather than model capacity. Each pair of explanations is evaluated twice (A→B and B→A) to reduce order bias, with up to three retries allowed if the model outputs an invalid comparison label.

Scale of Evaluation. We select 10 concepts from each domain (Artificial Intelligence, Computer Science, and Statistics), resulting in a total of 30 concepts. For each concept, we generate three explanations—one from each prompt template—yielding 90 generated explanations. Each concept produces three pairwise comparisons (Prompt 1 vs Prompt 2, Prompt 1 vs Prompt 3, Prompt 2 vs Prompt 3), and with reverse-order evaluation this results in 6 comparisons per concept. In total, the evaluation consists of 540 pairwise judgments, which are aggregated using the Elo scoring system to produce the final prompt leaderboard.

3.3 Data Collection

We construct our concept dataset using several Wikipedia domain glossaries, including the Glossary of Artificial Intelligence, Glossary of Computer Science, and Glossary of Statistics. These sources provide a broad set of domain-specific terms along with concise definitions, enabling consistent coverage across AI, data science, and related technical fields.

To identify which concepts are most suitable for evaluating explanation quality, we employ a large language model (LLM) as an automated difficulty assessor. Rather than selecting terms randomly, we prompt the LLM to rate each concept along three dimensions:

- **Complexity** — How difficult the concept is for a non-expert to understand (1 = easily understood; 10 = requires advanced theoretical background or integration of multiple sub-concepts).
- **Familiarity** — How likely an average college student is to have encountered the term (1 = widely familiar; 10 = rarely known outside specialized domains).
- **Explainability** — How easily the concept can be summarized in a short, non-technical sentence (1 = very easy to simplify; 10 = difficult to simplify without significant loss of meaning).

For each term, the LLM provides numeric scores on these three scales. We then aggregate these ratings to identify concepts that are simultaneously abstract, unfamiliar, and hard to simplify—properties that make them ideal for stress-testing explanation quality across different prompt designs. Concepts with high combined difficulty scores are selected for downstream evaluation, ensuring that our benchmark focuses on terms that genuinely challenge both LLM reasoning and human interpretability.

3.4 Prompt Design and Generative Process

To examine how instruction design influences explanation quality, we constructed three structured prompt templates that vary in complexity and level of guidance. Each prompt targets a different style of explanation, ranging from minimal instruction to multi-dimensional framing, allowing us to capture how LLMs respond to increasing instructional constraints. All explanations for each concept were generated using the same model to isolate the effect of prompt structure alone.

Prompt 1: Baseline. A minimal, direct instruction designed to elicit short, simplified explanations:

> *"Explain the following concept in plain language as short as possible: {concept}."*

This prompt serves as a control condition, testing how the model explains a term when given almost no structural guidance.

Prompt 2: Level 2 (Multi-Aspect). A moderately structured prompt that asks the model to explain the concept using three explicit components:

> *"Explain the concept {concept} for a non-expert. Please cover: (1) its basic meaning, (2) a simple real-world example, and (3) why it is important. Please keep the entire explanation under 200 words."*

This version introduces light scaffolding to encourage clearer, more audience-aware explanations.

Prompt 3: Level 3 (Multi-Perspective). A highly structured prompt that requests multiple explanatory perspectives, encouraging depth and detail:

> *"Provide a comprehensive explanation of the concept {concept}. Your explanation should integrate multiple perspectives, including: (1) an intuitive perspective that conveys the core idea, (2) a formal perspective with a precise definition or theoretical framing, (3) a practical perspective describing where or how it appears in real applications, and (4) any relevant background knowledge or related concepts that help deepen understanding. Conclude with a short analogy that ties the perspectives together. Please keep the entire explanation under 200 words."*

This prompt aims to test whether heavy structure improves clarity or whether excessive constraints reduce comprehensibility for non-experts.

Together, these three prompts enable a controlled comparison of how prompt complexity affects the structure, clarity, and human interpretability of LLM-generated explanations.

3.5 Evaluation System

After we got all the word explanations that generated instructions by our three unique prompts, we conducted a pairwise large language model (LLM) judgment parallel in words on the response of each prompt in a way that the LLM thinks humans will understand the best. Each judgment uses an LLM evaluator with a system prompt that instructs the model to act as an experienced educator evaluating explanations from the perspective of a non-expert college student with limited patience and no background in the major. The system prompt requires focusing on:

- How easy the explanation feels to read on the first pass.
- Whether it gives a clear, intuitive "now I get it" feeling.
- How approachable and non-intimidating the wording is.
- Whether it avoids unnecessary jargon or complexity.
- Whether the explanation is the right length, not too long or overwhelming. (Humans lose patience quickly; long, dense explanations reduce understanding.)

- Overall, which explanation a real student would *actually prefer* because it is easier to follow and more helpful.
- You are not grading research papers. You are judging which explanation best supports real human understanding.

The user prompt provides the major, term, and both explanations labeled by their prompt variant names, asking which explanation a typical non-expert college student would find easier to understand, more readable, and more helpful.

We do not judge based on the best explanation that LLM determines by scoring on an accuracy or understanding scale, but ask them to be in a human position to decide. We also design a reverse judgment in which each comparison is judged twice (in both orders: A→B and B→A) to reduce order bias that the LLM might introduce. In this way, after both judgments are obtained, the function determines the final winner: if both judgments agree that prompt A wins (judgment_ab says "A" and judgment_ba says "B", which means prompt A wins in the swapped order), then A wins; if both agree that prompt B wins, then B wins; otherwise, the result is a tie. To reduce the model's hallucinations, the system will retry up to 3 times if a judgment returns a response other than 'A', 'B', or 'tie'.

3.6 Elo Scoring and Leaderborad

After judging, we analyze all comparison results by calculating Elo ratings for each prompt based on win/loss/tie records. The Elo score works by updating the rating of each prompt after each comparison, according to the expected probability of winning; beating a strong opponent therefore, increases the score more than beating a weak one does. Elo scores are computed by treating each pairwise comparison between prompts as a "match" in an Elo rating system. Every prompt starts with an initial rating (defaulting to 1500), and for each comparison, the LLM judge will determine which prompt won (A, B, or tie) and update both prompts' ratings accordingly.

For the comparison process, it reads the current ratings of prompt A (R_A) and prompt B (R_B), and converts them into expected scores using the standard Elo logistic formula:

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}}, \qquad E_B = \frac{1}{1 + 10^{(R_A - R_B)/400}}.$$

The actual scores are calculated based on the comparison outcome:

$$S_A = \begin{cases} 1, & \text{if A wins,} \\ 0.5, & \text{if tie,} \\ 0, & \text{if A loses,} \end{cases} \qquad S_B = 1 - S_A.$$

Each rating is updated using the Elo update rule:

$$R'_A = R_A + K \cdot (S_A - E_A), \qquad R'_B = R_B + K \cdot (S_B - E_B),$$

where K (default 32) controls the sensitivity of the rating to new results.

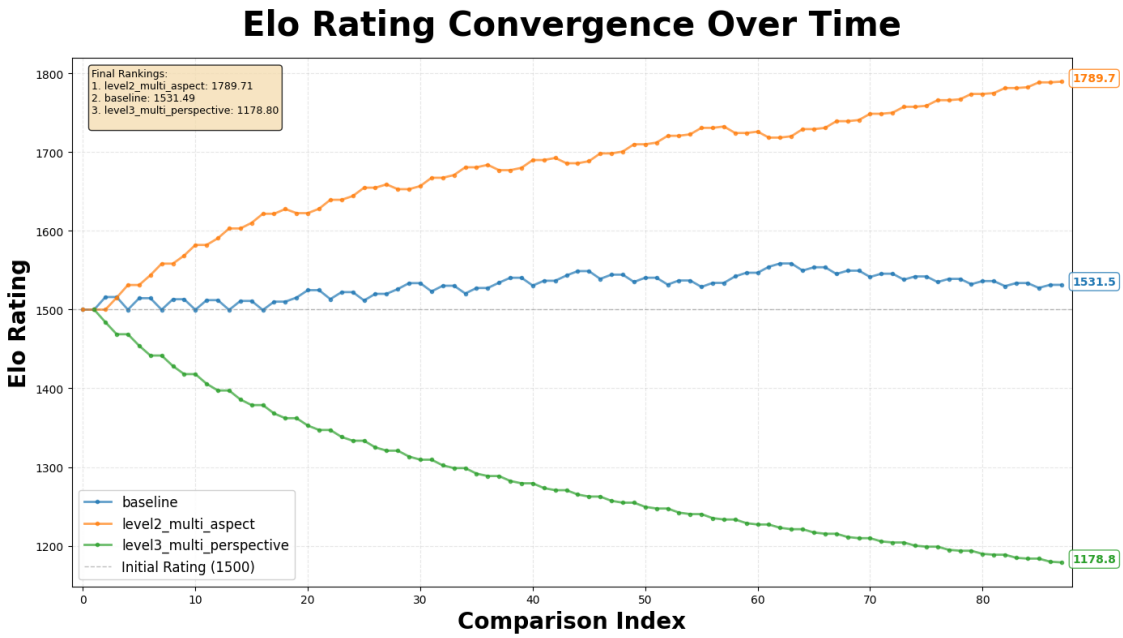

Figure 2: Elo Rating Convergence Curve

This update process is repeated for all comparisons in the evaluation file. Prompts that consistently win gain rating, while those that lose drop, and ties pull their ratings closer together. The Elo Score is the perfect method for our project, as it naturally aggregates many noisy pairwise LLM judgments into a stable, comparable ranking without absolute ground-truth labels, thus rendering it robust to variability and uncertainty in LLM-based evaluation. We use the Elo score to calculate a ranking among three prompts as our final result.

4 Results

In our results, we got a leaderboard, shown in Table 1, where the prompt with aspect and examples reached the highest Elo score. The higher the score represent the higher the possibilities the Large Language model believes the response generated by that prompt will be more likely to be adapted by humans. In these cases, we found that sometimes some restrictions and rules might not help LLM to generate a good response based on our needs. For example, in the Level 3 prompt, we instruct the model to explain the terms from multiple perspectives, but it fails to get accepted by humans based on the evaluation process and scoring, which also emphasizes the cruciality of the prompt in the LLM generation process.

Figure 2 shows that the three prompts clearly separate as the comparison process converges. The level2_multi_aspect prompt steadily rises in rating, quickly becoming the dominant prompt and indicative of consistently preferred explanations across pairwise evaluations. The baseline prompt remains stable within the 1500–1550 range, reflecting moderate performance. The continuous decline throughout the course of evaluation for

Table 1: Elo rankings of the three prompt types based on pairwise LLM judgments.

Rank	Prompt	Elo Rating
1	level2_multi_aspect	1789.71
2	baseline	1531.49
3	level3_multi_perspective	1178.80

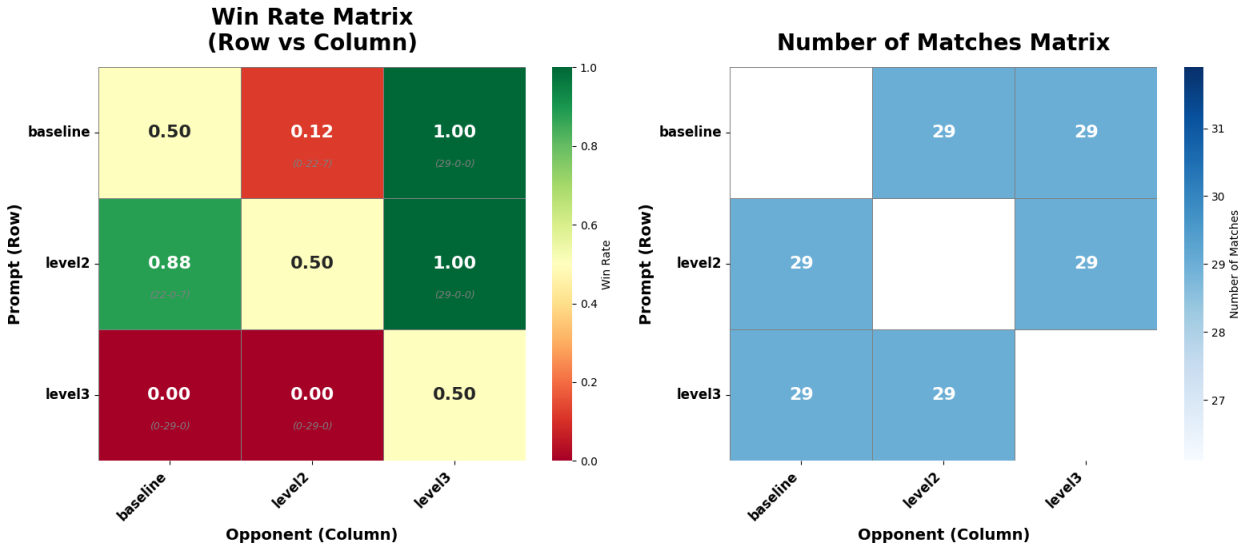

Figure 3: Win rate matrix across three prompts

the level3_multi_perspective prompt suggests its heavily structured explanations were less readable or less understandable for the non-expert audience. Overall, the pattern of convergence demonstrates the stability of the Elo scoring process and reliable differentiation of explanation quality across prompts.

Figure 3 gives a detailed view of prompt performance through both a win-rate matrix and a match-count matrix. The win-rate heatmap on the left shows that the level2_multi_aspect prompt dominates all pairwise comparisons, winning 88% of matches against the baseline prompt and 100% of matches against the level3_multi_perspective prompt. In contrast, the baseline prompt performs moderately, splitting 50% of its comparisons with level 3 but losing heavily to level 2. The level 3 prompt performs the weakest, failing to win a single match against either of the other two prompts. The right-side matrix confirms that each prompt pair was evaluated an equal number of 29 comparisons, ensuring that the observed win-rate patterns are not artifacts of uneven sampling. As a result, these outcomes reinforce the consistency of the Elo ranking, on which level 2 clearly produces the most preferred explanations, baseline achieves moderate clarity, and level 3's highly structured format appears to hinder readability for non-expert audiences.

5 Discussion and Future Improvement

For our next step, we plan to involve prompt improvement and regeneration by giving the response with the weakness of the response returned by the judgment process to the generation model. In this way, we will regenerate the prompts as improved prompts. After we get the improved prompts, we will test them by doing the whole process of generative, evaluation, and scoring system to refresh the ranking to see if it has a better accomplishment on human tasks and needs. In our expectation, the improved prompt will have a higher Elo score than it had before. In our expectation, after multiple runs of improvement and regeneration, we will get a template of a prompt that could largely match human expectations on model generation.

To test whether LLM judgment is reliable on human preference, we will design a human evaluation leaderboard to compare with the real human decision, with the LLM leaderboard to see the differences between LLM thought and real human understanding. In this way, we could better evaluate the gap between humans and AI technology, which might help future research on human-AI interaction.

6 Conclusion

In this project, we examined how different prompt designs influence the quality of LLM-generated explanations, using both LLM-as-Judge and Human-Judge frameworks to understand how models interpret instructions and how well their responses align with human preferences. Our results show that prompt structure plays a substantial role in shaping explanation clarity: the multi-aspect prompt consistently outperformed the baseline and multi-perspective versions, achieving the highest Elo score across pairwise comparisons. These findings suggest that adding structured guidance—such as specifying aspects or examples—helps models generate explanations that are more accessible to human learners, while overly complex or rigid instructions can unintentionally reduce interpretability.

Beyond demonstrating the importance of prompt design, our study highlights the value of combining LLM based and human-centered evaluation. LLM judges provide scalable and repeatable assessments, but the divergence observed across prompt levels emphasizes the need to verify whether LLM preferences truly reflect human judgment. By integrating both perspectives, we create a more realistic evaluation pipeline that captures not only a model's knowledge but its ability to communicate that knowledge effectively.

Looking ahead, iterative prompt refinement guided by weaknesses identified during evaluation offers a promising direction for improving explanation quality. A future human evaluation leaderboard will enable a deeper analysis of alignment gaps between LLM scoring and real human understanding. Ultimately, our goal is to move toward explanation strategies and prompt templates that reliably satisfy human needs, contributing to more transparent, interpretable, and learner-aligned AI systems.

References

Chiang, Wei-Lin et al. 2024. "Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference." *arXiv preprint arXiv:2403.04132v1*

Fan, Angela et al. 2019. "ELI5: Long-Form Question Answering." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

He, Jia et al. 2024. "Does Prompt Formatting Have Any Impact on LLM Performance?" *arXiv preprint arXiv:2411.10541*

Hendrycks, Dan et al. 2020. "Measuring Massive Multitask Language Understanding." In *International Conference on Learning Representations (ICLR)*.

Liang, Percy et al. 2022. "Holistic Evaluation of Language Models." *arXiv preprint arXiv:2211.09110*

Mizrahi, Moran et al. 2024. "State of What Art? A Call for Multi-Prompt LLM Evaluation." *Transactions of the Association for Computational Linguistics* 12: 933–949

Shankar, Shreya et al. 2024. "Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences." In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*.