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

Large language models (LLMs) are increasingly evaluated on knowledge, reasoning, and robustness across professional domains. However, their ability to accurately understand and follow human instructions remains underexplored, despite text-based interaction being the primary interface between humans and LLMs. In this work, we introduce a realistic and robust evaluation framework for instruction understanding that integrates both LLM-as-Judge and Human-Judge assessments across multiple levels of instruction prompts. Our framework begins with a set of base instruction prompts under which LLMs generate explanations for professional-level concepts spanning diverse domains. We then perform pairwise comparisons of model outputs using LLM-as-Judge with an Elo-style ranking scheme to obtain relative performance scores. Based on the comparison outcomes, the LLM analyzes systematic weaknesses in instruction following and iteratively refines the original prompts, producing a second set of improved prompts. This design enables controlled evaluation of both instruction sensitivity and prompt evolution. We evaluate model performance under both the original and refined prompts and present leaderboards. Our results reveal consistent gaps between LLM-based judgments and human preferences, as well as measurable improvements from iterative prompt refinement. Together, these findings provide new insights into instruction-following behavior, prompt design, and the limitations of automated evaluation for real-world human interaction with LLMs.

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

1 Introduction

1.1 Background

In the rapidly evolving field of artificial intelligence, evaluating the behavior and performance of large language models (LLMs) has become increasingly important. Existing benchmarks primarily focus on knowledge accuracy, reasoning ability, and robustness across professional domains. While these metrics assess factual correctness, they do not fully capture how effectively the model communicates that knowledge to users.

Since text-based interaction remains the primary interface between humans and LLMs, the ability to correctly interpret and respond to human instructions is fundamental for both real-world applications and scientific evaluation. In an educational context, models need to clearly structure explanations for particular audiences and abstraction levels in addition to offering accurate answers.

Two evaluation paradigms are commonly used to assess LLM behavior: LLM-as-Judge and Human-Judge. LLM-as-Judge offers scalable, reproducible, and knowledge-driven assessments by leveraging pretrained language models to evaluate generated content. In contrast, Human-Judge evaluation relies on human perception, contextual reasoning, and subjective preference, capturing nuanced judgments that automated evaluators may overlook. While both paradigms are widely adopted, it is unclear whether LLM judgments reliably reflect human preferences, particularly under varying instruction constraints.

1.2 Motivation

Despite the progress in LLM evaluation, it still does not fully address several important limitations. First, while prior work studies correctness and reasoning accuracy, fewer efforts systematically analyze how instruction structure and prompt framing influence explanation quality across domains. Second, although both Human-Judge and LLM-as-Judge paradigms are widely used, the alignment between LLM and human evaluation remains unclear.

To address this gap, our project investigates how LLMs generalize across multiple levels of instruction prompts by jointly analyzing LLM-as-Judge and Human-Judge evaluations. We design a structured prompt framework that systematically varies instruction specificity and constraint, enabling controlled analysis of instruction sensitivity in model-generated explanations across diverse professional domains.

Beyond static evaluation, we introduce an iterative prompt refinement mechanism that enables LLMs to learn from evaluation outcomes. Specifically, after conducting pairwise comparisons of model outputs under different prompts, the LLM analyzes systematic weaknesses in instruction following and generates refined instruction prompts. This process allows direct comparison between original and refined prompts under both evaluation paradigms. Through this design, we study not only instruction-following performance, but also the effects of prompt evolution on model behavior.

Using both LLM-as-Judge and Human-Judge assessments, we construct comprehensive leaderboards. By comparing rankings and preferences across judging paradigms, we analyze the consistency and divergence between automated evaluations and human choices. This evaluation framework enables us to assess framework enables us to assess alignment gaps between model-inferred interpretations of human intent and actual human preferences.

In short, these components enable a systematic examination of how instruction variation affects model performance, how consistent different evaluation paradigms are, and how these effects generalize across domains.

1.3 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, these evaluation rubrics primarily focus on factual alignment and overall coherence, rather than explicitly assessing how well a model structures explanations for learners.

Recent work has further highlighted the sensitivity of LLM evaluation to prompt framing. Mizrahi et al. (2024) argue 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. Similarly, He et al. (2024) investigate how format and structure influence model behavior, showing that presentation differences can affect 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.

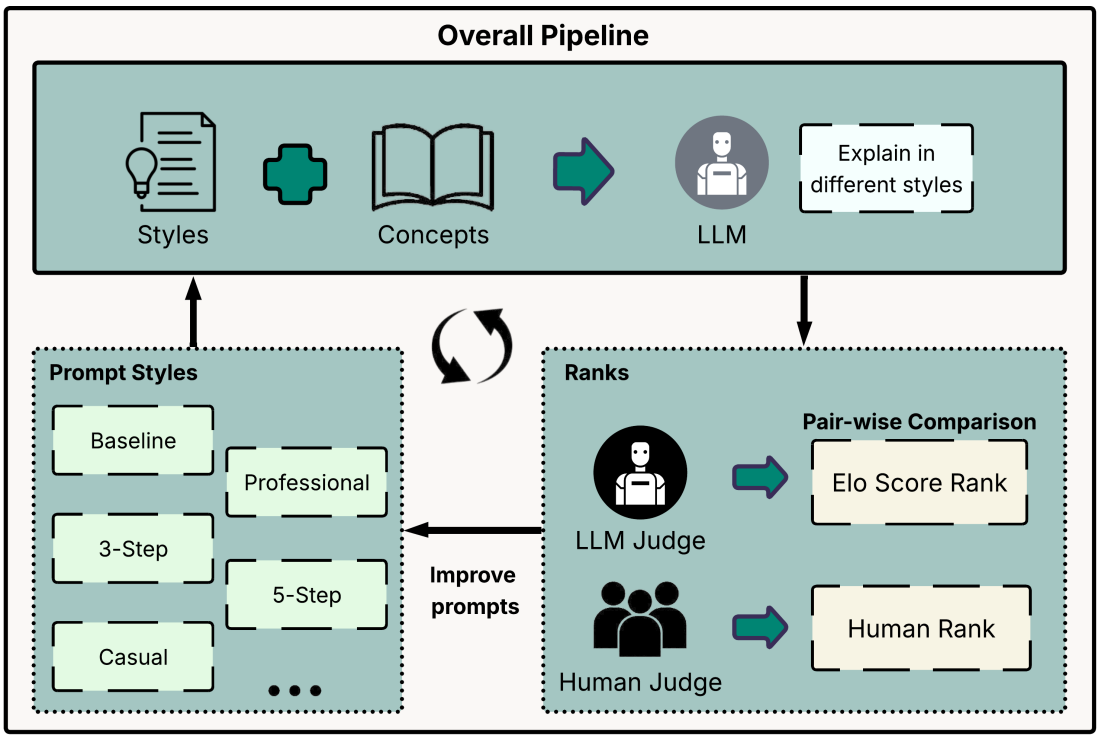

Figure 1: Full pipeline of our evaluation framework, including concept selection, structured prompt generation, LLM explanation generation, pairwise LLM judgment, Elo score aggregation, and final leaderboard construction.

2 Methods

2.1 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, from concept selection based on domain glossaries, through multi-level prompt construction and LLM-based explanation generation, to pairwise LLM-as-Judge comparisons, reverse-order and retry strategies to counter positional and stochastic bias, and finally aggregation of results using Elo scoring to rank the quality of the prompts. Based on these evaluation results, the model identifies systematic weaknesses and regenerates refined instruction prompts, which are subsequently re-evaluated within the same pipeline. This modular design allows individual components, such as prompt templates, judging models, regeneration strategies, or scoring mechanisms, to be swapped or extended without modifying the overall architecture.

2.2 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

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.

Prompt Design. To evaluate how different prompts styles affect the quality of generated explanations, we designed 5 prompts that vary in structure and personas. The baseline is a single instruction with normalize word limit: "Explain the following concept in plain language in 200 words or fewer". For the second prompt ask the model to include realistic examples in its explanation. The third prompt structures the response into 5-step content, which is supposed to generate a more detail explanation. The remaining two prompts mainly focus on different personas: one is asked the model to explain the term as speeking to a friend and another one is teach more prefeesional as a course professor. In this variation, we aim to explore a better way to design prompts that generate most effective and comprehensive explanations.

Scale of Evaluation. We select 10 concepts from each domain—Artificial Intelligence, Computer Science, and Statistics—resulting in a total of 30 concepts. In the first round, each concept is paired with 5 base prompt templates, generating 5 explanations per concept and 150 explanations in total. In the second round, the LLM generates an additional 5 refined prompts based on weaknesses identified in the first-round evaluations, producing 10 explanations per concept and 300 explanations overall across both rounds. For each evaluation round, prompt quality is assessed through exhaustive pairwise comparisons. In the first round, five prompts yield 10 unique prompt pairs per concept, and with reverse-order evaluation this results in 20 comparisons per concept. The second round producing 45 prompt pairs and 90 comparisons per concept with reverse-order evaluation. Across 30 concepts, the system performs a total of 600 pairwise judgments in the first round and 2,700 in the second round. All pairwise outcomes are aggregated using the Elo scoring system to produce the final prompt leaderboard.

2.3 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 assigns numerical scores across these three dimensions. We then combine these scores to select concepts that are both abstract, unfamiliar, and difficult to simplify, characteristics which make them perfect for stress-testing the explanation quality of various prompt designs. Concepts with high aggregate difficulty scores are chosen for evaluation, thereby ensuring that our benchmark is centered on terms that are truly challenging for both LLM reasoning and human interpretability.

2.4 Evaluation System

After generating word explanations using three prompt variants, we conducted pairwise LLM-based evaluations to determine which explanation is most understandable for non-expert college students. The evaluator is instructed to act as an experienced educator and judge explanations based on readability, clarity, approachability, appropriate length, and avoidance of unnecessary jargon. The goal is not factual accuracy scoring, but selecting the explanation that best supports real human understanding.

For each term, we perform reverse judgments (A→B and B→A) to reduce order bias. A prompt wins only if both judgments agree; otherwise, the result is considered a tie. To ensure output validity, the system retries up to three times if the evaluator returns a response other than "A," "B," or "tie."

2.5 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.

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.

3 Results

3.1 Prompt sensitivity analysis

The impact of prompt refinements was analyzed by comparing the rankings of original and improved prompts using slopegraphs for both human and LLM evaluations (Figure 2). Each line represents a prompt style, connecting its rank before and after refinement, where downward movement indicates an improvement in rank. Across both evaluation settings, most prompt styles exhibit downward slopes, indicating that refinement generally leads to improved relative performance.

Under human evaluation, the baseline prompt shows the most pronounced improvement, moving from a low initial rank to the top position after refinement. The level2 and casual

prompts also improve, though to a lesser extent, while the 5-step prompt shows a smaller but still observable rank gain. In contrast, the academic prompt remains consistently low-ranked before and after refinement, suggesting limited sensitivity to the refinement process in human judgments.

A similar pattern is observed in the LLM-based evaluation. The level2 prompt demonstrates a clear improvement, rising to the top rank after refinement, while baseline and 5-step prompts show moderate upward movement. As in the human evaluation, the academic prompt remains near the bottom of the ranking with minimal change. Overall, the slope-graphs reveal that prompt refinement tends to benefit prompts that emphasize accessibility and guided structure, while highly formal academic prompts show comparatively limited rank changes.

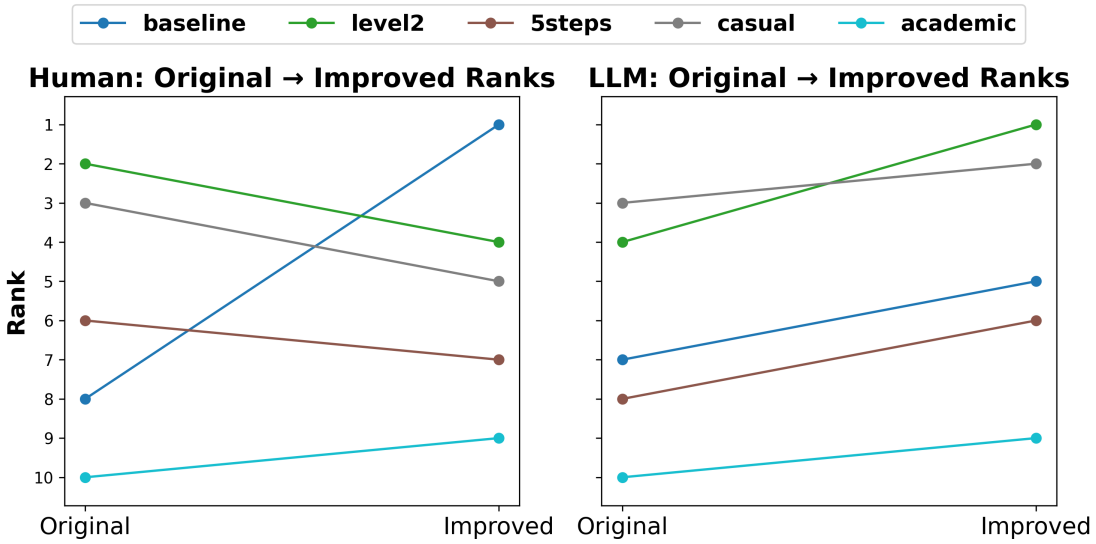

Figure 2: Slopegraphs for human and LLM ranks before and after prompt improvement.

3.2 Human VS. LLM ranking comparison

We compare prompt rankings produced by Human-Judge and LLM-as-Judge evaluations to analyze consistency and divergence between human preferences and automated judgments.

The rankings assigned by humans and the LLM were compared using Spearman's rank correlation coefficient and Kendall's tau. The observed Spearman coefficient of approximately 0.73 indicates a moderate positive correlation, suggesting that the LLM's rankings align reasonably well with human evaluations. Kendall's tau further supports this conclusion, highlighting the consistency of relative rankings between the two evaluators.

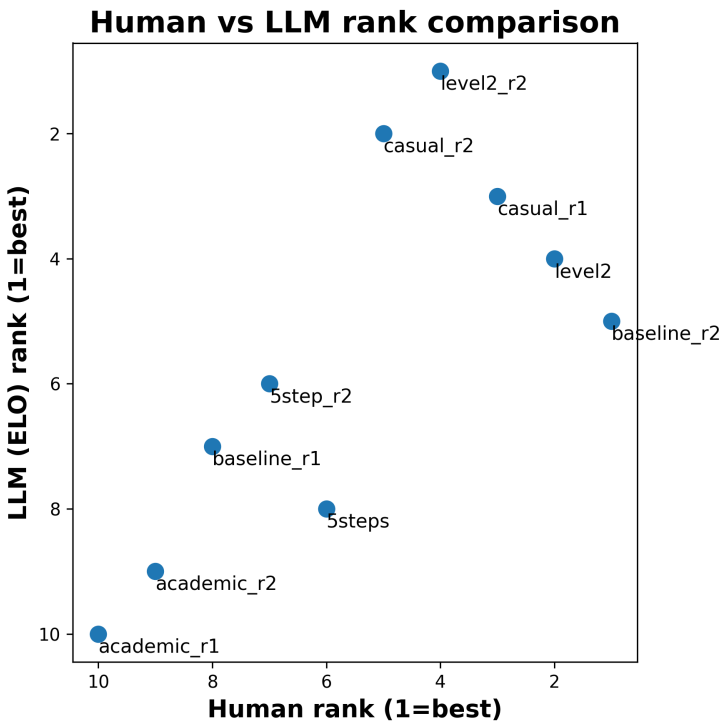

Figure 3: Human ranking VS. LLM ranking scatter plot.

The bar chart quantifies the rank changes for each prompt pair, comparing human and LLM evaluations. Negative values indicate that the improved prompt was ranked better than the original. The chart provides a clear comparison of the magnitude of changes, highlighting prompts where improvements had the most significant impact.

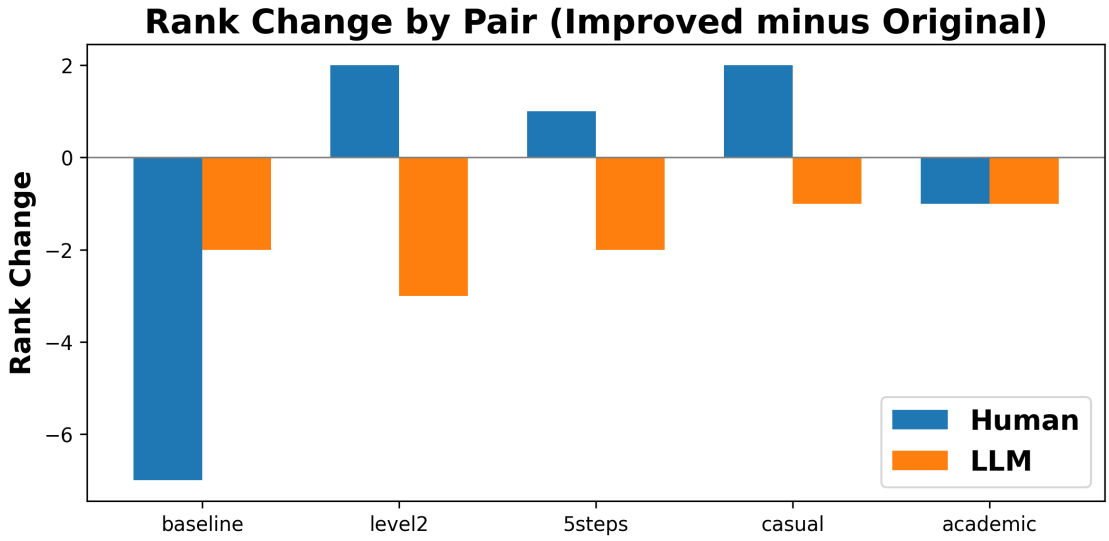

Figure 4: Bar Chart of Rank Changes (Improved - Original). Negative values indicate better rankings for improved prompts.

3.3 High-score preference comparison

We would like to examine the similarity in top LLM-as-judge responses to see if we could find some interested conclusion on LLM-as-judge responses and the same for top human evaluation responses. However, the similarity matrix above shows that there is little lexical overlap in the top responses chosen for different terms. The off-diagonal elements of the cosine similarity matrix are mostly close to zero, which means that the top responses with high ratings are more term-specific rather than formulaic. Even when comparing the top responses preferred by LLM and human judges, there is no strong evidence of overlapping wording patterns for different concepts. This led to the further structural and stylistic analysis in the next section.

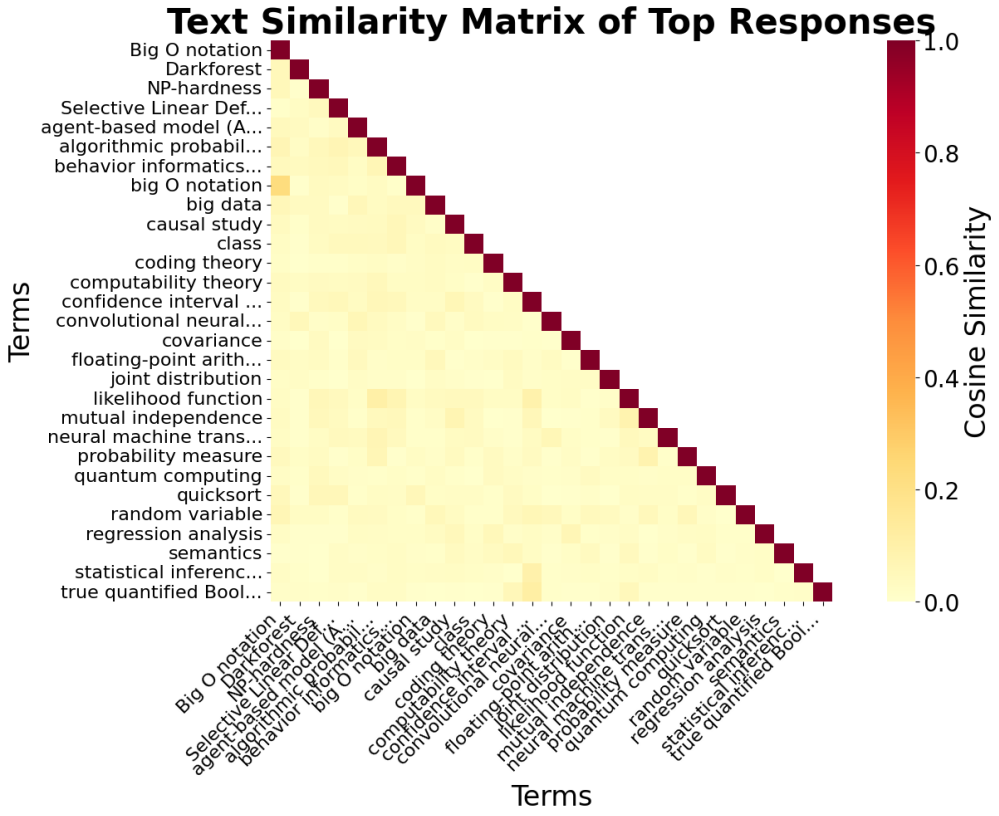

Figure 5: Similarity matrix.

Table 1: Lexical and Structural Feature Comparison Between Human and LLM Responses

Feature	Human Mean	LLM Mean	Difference	Relative Diff (%)
Word count	187.83	116.72	71.10	60.92
Sentence count	11.48	6.03	5.45	90.29
Bullet count	6.07	2.90	3.17	109.52
Numbered list count	4.14	0.00	4.14	0.00
Potential headers	1.72	0.45	1.28	284.62

The LLM-preferred top responses are more compact and shorter, with fewer structural features. They do, however, have slightly higher vocabulary diversity and longer average sentence length.

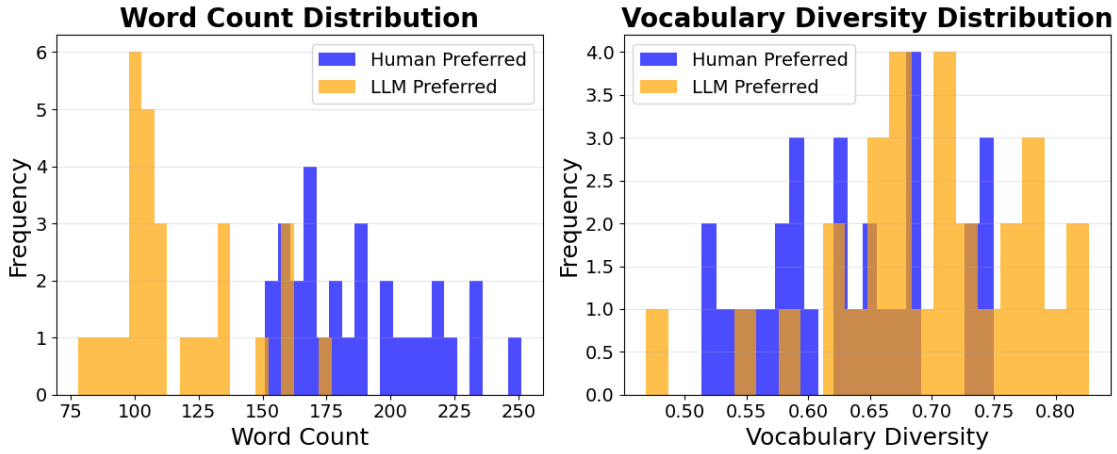

Figure 6: Word count and vocabulary diversity distribution.

The distribution plots further support these results. Human-preferred responses are grouped around higher word counts and paragraph counts, while LLM-preferred responses are grouped around lower word counts and higher sentence lengths.

3.4 Word-level winning rate across prompts and domains

Across all domains combined, the winning-rate analysis shows a clear stratification among prompt styles. Prompts refined in Round 2 consistently outperform their Round 1 counterparts, indicating that iterative prompt engineering substantially improves response quality as judged by both human evaluators and LLM-based ranking. In particular, Casual round2 and Level2 round2 achieve the highest overall win rates, suggesting that prompts which balance accessibility with structured guidance are most effective. In contrast, academic-style prompts, especially the highly formal academic variant, exhibit the lowest win rates, consistently underperforming relative to other styles. This pattern implies that excessive formality or rigid academic framing may hinder clarity, usefulness, or perceived helpfulness in generated responses.

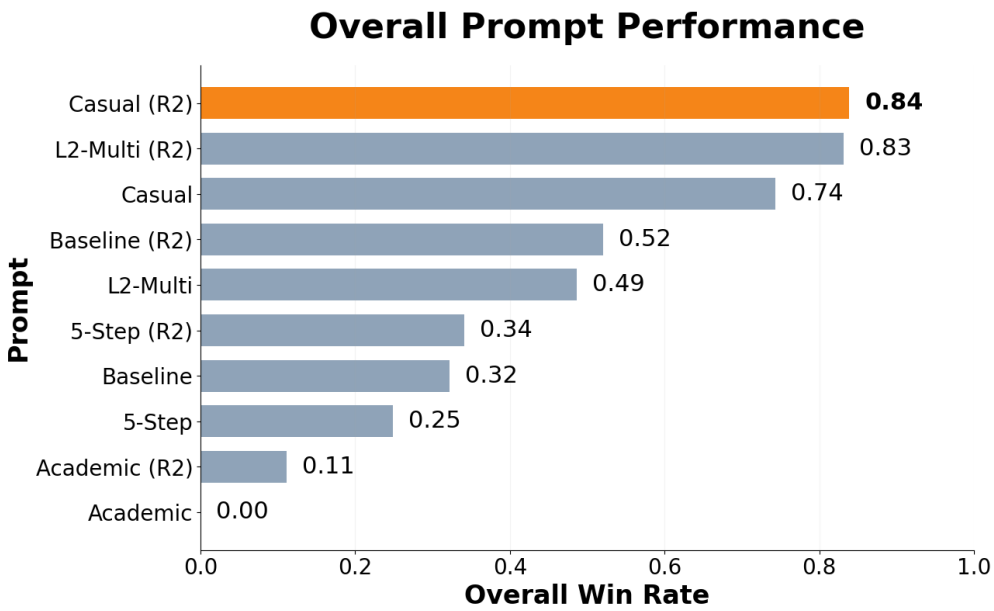

Figure 7: Overall prompt performance in terms of their winning rate.

The consistency between human judgments and LLM-based rankings is notable. High Kendall's τ values (approximately 0.85–0.96) indicate strong rank-order agreement across evaluators, suggesting that LLM-based comparisons reliably approximate human preferences in this setting. This alignment strengthens the validity of using LLMs as scalable evaluators for prompt effectiveness, particularly when human annotation resources are limited. Importantly, both human and LLM rankings converge on the same top-performing prompt styles, reinforcing the conclusion that conversational yet structured prompts yield superior outcomes.

When disaggregating results by domain (Artificial Intelligence, Computer Science, and Statistics), prompt rankings remain highly stable. The relative ordering of prompt styles changes minimally across domains, indicating that domain-specific content has little influence on which prompt performs best. Casual and multi-aspect prompts consistently rank near the top, while academic-style prompts remain near the bottom regardless of subject matter. This robustness suggests that prompt style effects dominate over domain effects, and that well-designed prompts generalize effectively across technical disciplines.

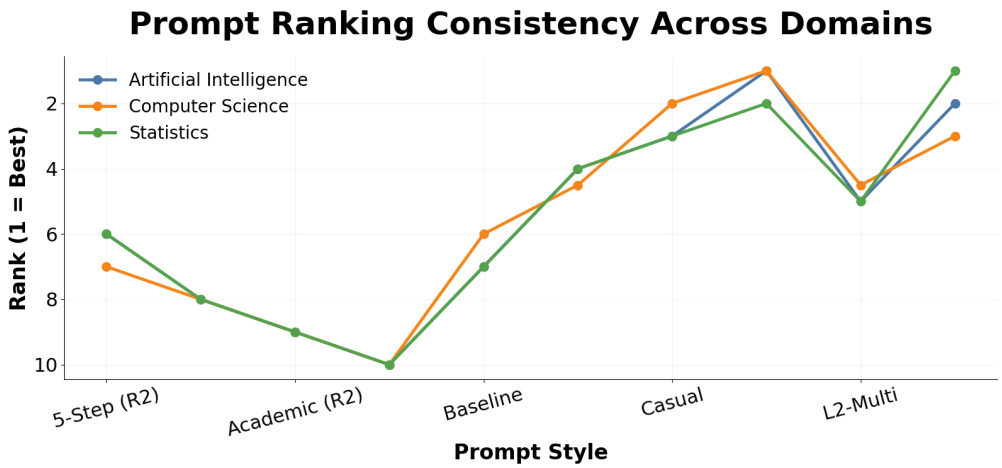

Figure 8: Prompt ranking consistency across three domains.

4 Discussion

4.1 Alignment Between Human and LLM Evaluation

Our results show a moderate to strong alignment between human judgments and LLM-as-Judge rankings, particularly for top-performing prompts. This finding suggests that LLM-based evaluators can approximate human preferences reasonably well when distinguishing higher-quality explanations. In practice, this means that LLM-as-Judge evaluations are effective at identifying prompts that consistently produce strong explanations, especially in comparative settings where relative quality is the primary concern.

However, alignment between human and LLM evaluations decreases for lower-ranked prompts. This divergence indicates that while LLM judges capture coarse-grained preferences, they may be less sensitive to subtle differences in explanation quality that humans perceive. As a result, LLM-based evaluation appears more reliable for ranking strong candidates than for finely discriminating among weaker or borderline responses. These findings highlight both the promise and the limitations of automated evaluation, suggesting that LLM-as-Judge should be viewed as a complementary tool rather than a full replacement for human judgment.

4.2 Differences in Human and LLM Quality Criteria

A deeper analysis of lexical and structural features helps explain the observed differences between human and LLM preferences. Human-preferred explanations tend to be longer, more segmented, and more explicitly structured. They frequently include multiple paragraphs, bullet points, numbered lists, and implicit section headers, which collectively make the content easier to scan and cognitively process. This pattern suggests that human eval-

uators value explanations that resemble instructional materials, such as structured study notes or mini-lectures, particularly when dealing with complex or abstract concepts.

In contrast, explanations favored by the LLM judge are generally shorter and more compact, with fewer explicit structural markers. These responses emphasize fluency, lexical diversity, and local coherence, indicating that the LLM evaluator prioritizes well-formed prose over global organizational clarity. As a result, LLM-preferred explanations often read as concise summaries, whereas human-preferred explanations function as learning-oriented artifacts designed to support comprehension.

Taken together, these findings suggest that humans and LLMs operationalize the notion of "quality" differently. While both value clarity and relevance, humans place greater emphasis on structural organization and completeness, whereas LLM judges tend to favor brevity and stylistic polish.

4.3 Implications for Prompt Design

The observed evaluation patterns have important implications for prompt design. Across domains and evaluation paradigms, prompt style emerges as a primary driver of explanation quality. Prompts that encourage clarity, accessibility, and explicit structure consistently outperform those that emphasize formality or academic tone alone. In particular, conversational prompts with guided structure benefit both human and LLM evaluations, suggesting that effective explanations strike a balance between approachability and organization.

Iterative prompt refinement further amplifies these benefits. Prompts revised based on evaluation feedback consistently achieve higher rankings, demonstrating that prompt quality can be systematically improved through iterative design. Moreover, the stability of prompt rankings across domains indicates that effective prompt strategies generalize well, reducing the need for domain-specific customization. For practical applications, these results suggest prioritizing user-oriented prompt designs and avoiding overly rigid academic formulations when the goal is to maximize perceived explanation quality.

4.4 Limitations

Despite the promising results, several limitations should be considered when interpreting our findings.

First, the alignment between human judgments and LLM-as-Judge evaluations, while substantial, remains imperfect. The observed Spearman correlation of approximately 0.73 indicates reasonable agreement but also leaves room for meaningful divergence. In particular, scatter plots reveal clear outliers in which automated rankings differ markedly from human preferences. These cases suggest that reliance on LLM-as-Judge alone may misrank certain prompts or fail to capture nuanced human judgments, especially when differences in explanation quality are subtle. Consequently, LLM-based evaluation should be viewed

as a complementary tool rather than a standalone substitute for human assessment.

Second, the evaluation framework exhibits a stylistic bias rooted in differing notions of explanation quality. Human evaluators tend to prefer explanations that are more structured, comprehensive, and explicitly organized, whereas the LLM judge shows a systematic preference for shorter, more concise, and fluent responses. This divergence implies that high LLM-as-Judge scores do not necessarily correspond to explanations that humans find most clear or pedagogically effective. As a result, automated evaluation may overvalue stylistic polish and brevity at the expense of structural clarity and instructional completeness.

Third, the scale and diversity of human evaluation in this study are limited. Human annotations were conducted by a relatively small group with similar educational backgrounds, which may constrain the generalizability of the findings. A larger and more diverse pool of evaluators could yield different estimates of human–LLM agreement and potentially alter conclusions regarding preferred explanation styles and prompt effectiveness.

Finally, our findings are conditioned on a specific evaluation setup, including a single LLM-as-Judge configuration and a fixed set of technical domains. Although prompt rankings show strong stability across the domains studied (with Kendall's τ values around 0.90), this robustness is not guaranteed to extend to other subject areas or alternative judge models. Different domains, audiences, or evaluator architectures may prioritize different aspects of explanation quality. Therefore, calibration and validation with human feedback remain essential before deploying LLM-as-Judge evaluation frameworks in broader or high-stakes settings.

4.5 Future Improvement

Future work could address these limitations by expanding the pool of human evaluators to include participants with more diverse backgrounds and expertise levels. Using different model families for generation and evaluation would help disentangle shared biases and provide a more robust assessment of LLM-as-Judge reliability. Incorporating ranking-sensitive metrics such as Rank-Biased Overlap (RBO) or Normalized Discounted Cumulative Gain (NDCG) could also offer finer-grained insight into partial agreement between human and LLM rankings.

Beyond preference-based evaluation, future studies could incorporate learning-based metrics, such as comprehension quizzes or recall tasks, to directly measure the educational effectiveness of explanations. Finally, refining the LLM-as-Judge prompt to explicitly reward structural clarity and pedagogical usefulness may help narrow the remaining gap between automated and human evaluation.

5 Conclusion

In this project, we proposed and evaluated a structured framework for assessing the instruction-following and explanation quality of large language models using both Human-Judge and

LLM-as-Judge paradigms. By systematically varying prompt styles and applying iterative prompt refinement, we examined how instructional framing influences the clarity, usefulness, and perceived quality of model-generated explanations across multiple technical domains.

Our results demonstrate that prompt design plays a central role in shaping explanation quality. Prompts that balance accessibility with explicit structure consistently outperform highly formal or purely academic formulations, and iterative refinement based on evaluation feedback leads to measurable improvements. Across domains, prompt rankings remain stable, suggesting that effective prompt strategies generalize well beyond domain-specific content.

We further show that LLM-as-Judge evaluations align moderately to strongly with human preferences, particularly for identifying top-performing prompts. However, deeper analysis reveals systematic differences in evaluation criteria: human evaluators tend to favor structured, comprehensive explanations, while the LLM judge prioritizes brevity and linguistic fluency. These findings highlight both the potential and the limitations of automated evaluation, emphasizing the need for careful calibration when using LLM-based judges in educational or user-facing settings.

In short, this study contributes empirical evidence that explanation quality is not solely a function of model capability, but is strongly shaped by prompt design and evaluation methodology. Our findings support the use of iterative, user-oriented prompt engineering and position LLM-as-Judge as a scalable but imperfect proxy for human evaluation. We hope this work informs future research on prompt optimization, explanation-centered benchmarks, and more human-aligned evaluation frameworks for large language models.

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*.

A Project Proposal

A.1 Overview

As the AI systems become more advanced, people increasingly rely on Large Language Models (LLMs) not just for simple 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 student, and an expert researcher all need 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 complex concepts.

Given a collection of difficult concepts $C = \{c_1, \ldots, c_n\}$, our goal is to evaluate the quality of explanations produced by multiple LLMs under different prompting strategies. Each concept is explained at three levels of depth (baseline, level 2, level 3). Specifically, we set baseline as simple answers in under 200 words, level 2 as answers with examples and contexts in under 200 words, and level 3 as a detailed explanation with various examples and justification under different aspects. 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.

A.1.1 Problem Definition

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

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

(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:

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

with all pairs being compared, 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.

A.2 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?

A.3 Relation to Prior Work

Existing benchmarks for LLMs are typically categorized into two main groups, each being a significant challenge in terms of examining explanation quality. Benchmarks such as MMLU(**?**) and MT-Bench(**?**) primarily measure a model's factual recall or problem-solving ability using fixed question sets. While we are able to determine whether the answers are correct, these benchmarks, which measure a model's competency rather than its interaction with humans, provide limited insight into how readable or understandable a model's explanations are for real human users with different ages, backgrounds, or expertise levels. They evaluate correctness, but not the usability 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 human preferences, they do not explicitly evaluate educational clarity or whether an

explanation supports learning outcomes. Therefore, a model producing convincing but incorrect explanations could attain a high rank.

To provide a solution to this shortcoming, our project proposes to design a benchmark that emphasizes explanation quality for complex or abstract concepts. While adapting three different degrees of controlling prompts (baseline, intermediate, and highly structured prompts), we will evaluate the output based on the judgments of humans and LLM-as-judge pairwise comparisons. The pairwise preferences will then be pooled through Elo scoring to create a ranking of which prompting strategies produce explanations that are most likely to be considered helpful and comprehensible by humans. Our method, which combines preference with knowledge, allows us to check not only the correctness of the explanation but also its accessibility to a variety of users.

A.4 Data Collection

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

We begin by collecting hundreds of candidate concepts from different sources. Then, we use an LLM to sort them by difficulty, looking for terms that really need abstract thinking, statistical intuition, or the kind of understanding that crosses over into different fields. Each idea receives a difficulty score, and only the most challenging and interesting ones are selected. This way, we focus on the concepts where an informative explanation really stands out, and where the differences between prompt styles or models are more likely to matter.

The four of us in the project team will handle most of the human evaluation. We have all taken college-level statistics or machine learning, so we are familiar with what we are evaluating. For each explanation, we use a clear rubric that covers explainability, how complex the explanation is, and how familiar the concept feels. Although it is a small group, we want a tight pilot study to fine-tune the rubric and see how humans and LLMs align. If we need to verify our results, we will recruit more student evaluators.

The timeline for this project is reasonable since the data pipeline and the main evaluation parts were completed in the first quarter already. In detail, we made those for a smaller setting, such as choosing concepts from the three glossaries, providing explanations based on different prompt templates, performing pairwise LLM-as-judge comparisons with reverse-order controls, and using Elo scoring to combine results for ranking prompts. All these methods were applied to 30 concepts (10 per domain), which led to the production of 90 explanations and 540 pairwise judgments, along with stable and interpretable Elo rankings throughout the different types of prompts.

Scaling this study is practical. The glossaries have hundreds of extra terms included, so there is no need to collect new data. The increase of concepts from 30 to a few hundred will cause linear API usage, but it will still be computationally manageable since the system depends on the hosted LLM APIs only and not on training a model. Moreover, our pilot

experiments demonstrate that the runtime and costs are still in a very economical range.

A.5 Primary Output

The primary output of this project will be a comprehensive research report and an interactive website that showcases our evaluation procedure and results. The report will document the entire methodology we applied, which includes the prompt design, explanation generation, human evaluation methods, LLM-as-judge pairwise comparison, Elo rating calculation, and statistical treatment of agreement and disagreement between humans and models.

For the website, 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, disagreements between humans and LLMs, and how we improve the model explanations by adjusting the prompts. Through these visualizations, it will be possible for people to recognize the different responses of models to different prompt types and the differences in their explanations regarding clarity, accuracy, and usefulness. More importantly, they will be supporting our research questions directly by demonstrating (1) how prompt depth influences explanation quality, (2) where human and LLM-judge evaluations differ, and (3) whether prompt optimization meaningfully improves understandability.