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

Haoyang Yu[1] hay034@ucsd.edu
Jessie Zhang[1] jiz199@ucsd.edu
Jessica Zhang[1] jez004@ucsd.edu
Anduo Wang[1] anw043@ucsd.edu
Mentor: Ryan Lingo[2] ryan_lingo@honda-ri.com
Mentor: Rajeev Chhajer[2] rajeev_chhajer@honda-ri.com

[1]University of California, San Diego [2]Honda Research Institute 99P Lab

Introduction

Large language models (LLMs) are widely evaluated on correctness and reasoning, yet their ability to **understand and follow human instructions** remains underexplored. Since text-based interaction is the primary interface between humans and LLMs, explanation clarity and instruction sensitivity are critical for real-world use.

We propose a structured evaluation framework to study how prompt design influences explanation quality and whether **LLM-as-Judge rankings align with Human-Judge preferences**.

Method Design

We propose a modular evaluation framework to study instruction sensitivity and Human–LLM alignment. The pipeline consists of concept selection, prompt variation, pairwise LLM evaluation, and Elo-based ranking aggregation.

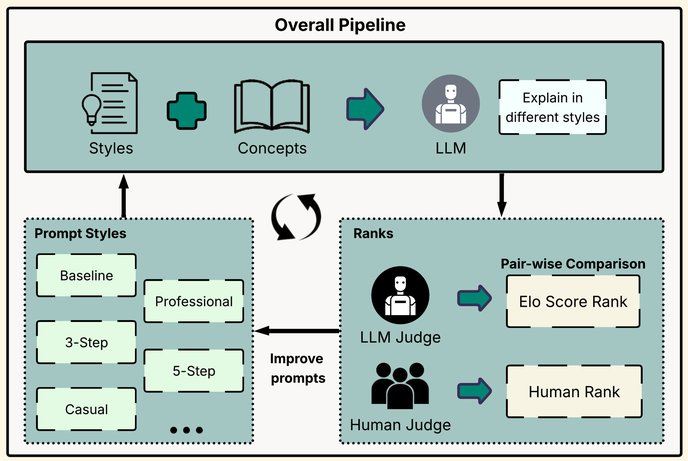


Figure 1. Overall evaluation pipeline. Concepts and prompt styles generate diverse explanations, which are ranked via pairwise LLM-as-Judge and Human-Judge comparisons. Elo scoring aggregates results into leaderboards, and evaluation feedback guides iterative prompt refinement.

Concept Selection. We select 30 high-difficulty concepts by:

- Collecting professional vocabulary dataset from AI, Computer Science, and Statistics glossaries
- Involving LLM-as-Judge scoring system to choose top 10 concepts from each dataset by complexity, familiarity, and explainability to ensure challenging evaluation cases

Prompt Design and Generation. Five base prompts vary in structural depth, example usage, and persona framing. For each concept c_i and prompt p_j:

$$e_{i,j} = f(c_i, p_j)$$

A second round introduces five refined prompts based on observed weaknesses.

- Compare base vs. refined prompts
- Measure instruction sensitivity

Pairwise Evaluation and Ranking. Explanations are compared using reverse-order LLM-as-Judge evaluation:

$$g(e_a, e_b) \to \{A, B, \text{tie}\}$$

Outcomes are aggregated via Elo scoring to produce stable prompt rankings.

- Reverse-order control reduces bias
- Elo converts pairwise judgments into comparable scores

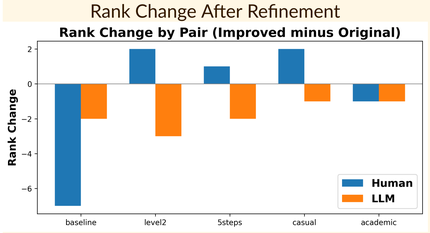
Human Evaluation and Alignment. Human judges rank explanations by clarity, readability, and instructional usefulness.

- Compare Human vs. LLM rankings
- Quantify agreement and divergence patterns

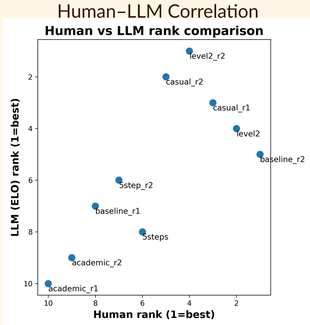
Human vs. LLM Ranking Comparison Analysis

We compare original vs. refined prompt rankings under Human-Judge and LLM-as-Judge evaluations.

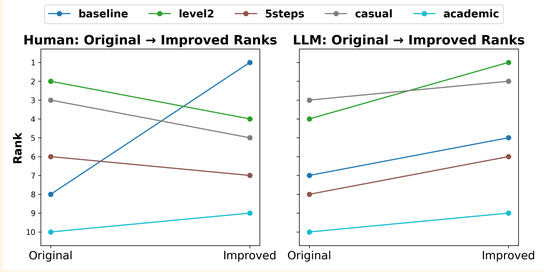
Rank Change After Refinement



Human–LLM Correlation



Rank Transitions (Original → Improved)

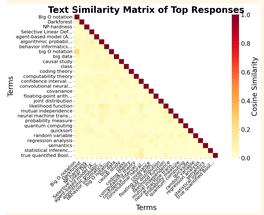


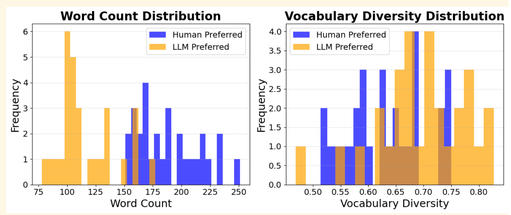
Key Observations

- Most refined prompts improve in rank after regeneration.
- Human and LLM trends are largely consistent across different academic fields.
- Moderate correlation ($\rho \approx 0.73$).

Qualitative Analysis: Human vs. LLM Preferences

We analyze lexical similarity, structural features, and distributional patterns between Human-preferred and LLM-preferred explanations.





- Near-zero cosine similarity indicates term-specific responses.
- The human-preferred answers are longer, with more sentences and more paragraphs. They often use bullet points, numbered lists, and section headers.
- The LLM-preferred answers are shorter and more compact, with fewer structural markers but slightly more vocabulary diversity and longer sentences.

Human judges seem to prefer structured, organized, and complete explanations that are easy to scan. In contrast, the LLM judge prefers short, fluent, and lexical diverse answers

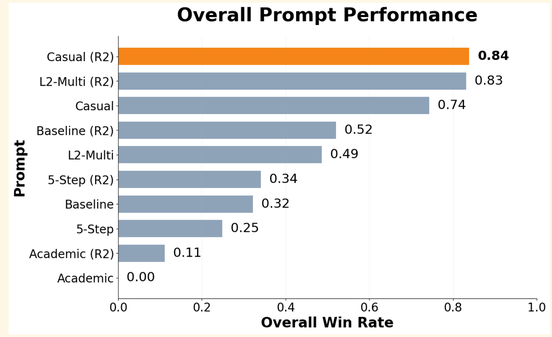
Interpretation

Humans favor structured and comprehensive explanations, whereas the LLM judge prefers concise and fluent responses. This divergence may reflect implicit length sensitivity in the LLM-as-Judge prompt.

Winning Rate Across Prompts and Domains

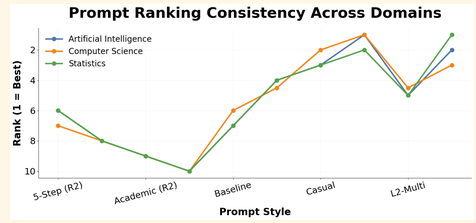
Takeaway. Iterative prompt refinement improves explanation quality and generalizes consistently across domains.

Overall Performance



- Round 2 prompts significantly outperform Round 1 versions.
- `Casual_round2` and `Level2_multi_aspect_round2` achieve the highest win rates.
- Academic-style prompts consistently underperform.

Robustness Across Domains



- Kendall's $\tau \approx 0.85$–0.96.
- Prompt rankings remain highly consistent across AI, CS, and Statistics.
- Domain has minimal impact on relative performance.

Conclusion and Future Work

Prompt design significantly influences the quality of LLM-generated explanations. While LLM-as-Judge provides scalable evaluation, it does not fully align with human preferences.

Our framework offers a systematic method to measure instruction-following behavior, quantify alignment gaps, and guide iterative prompt improvement.

Future Work.

- Expand human evaluator pool for stronger statistical reliability.
- Evaluate across multiple model families to test generalization.
- Incorporate learning outcome metrics (e.g., comprehension testing).

contact

Github

Project Website

Scan to explore the interactive leaderboard, Human–LLM comparison dashboard, and results.