PROMPT LOTTERY IN LLM-AS-A-JUDGE: MEASURING AND MITIGATING SENSITIVITY TO SYNTACTIC REPHRASING

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM-as-a-Judge has become a standard for evaluating large language models, yet most approaches implicitly assume that a single judge prompt functions as a stable evaluation instrument. Prior work has primarily explored deliberate prompt modifications such as scoring rubrics, role descriptions, or chain-of-thought reasoning in order to optimize accuracy. In contrast, we investigate whether validated prompt variants intended to preserve evaluation semantics while differing in surface realization (*paraphrastic rephrasings*) yield consistent judgments under a fixed evaluation task. We introduce *Prompt Lottery*, a phenomenon in which paraphrastic rephrasings produce different LLMaJ judgments. Across a large-scale study spanning diverse evaluation tasks, judge families, and model backbones, we find that these variations can alter evaluation outcomes and, in some settings, reverse relative model rankings. A user study with 10 human annotators shows that LLMaJ judges exhibit 6-9$\times$ larger agreement degradation under prompt rephrasings than humans, falling below the generally accepted threshold in usability studies. This indicates that single-prompt evaluation is insufficiently robust for model ranking. Finally, we show that multi-prompt aggregation substantially improves robustness by emphasizing judgments that remain consistent across prompt variants.

1 INTRODUCTION

Since their introduction as a reinforcement learning technique in Constitutional AI (Bai et al., 2022), LLM-as-a-judge (LLMaJ) use has become central to evaluating Large Language Models. The basic premise is that given an (LLM-generated) input, a set of instructions (prompt), and sometimes a reference answer, a backbone model produces a score as a proxy for the quality of input. Early work by (Zheng et al., 2023) reported high agreement between human and model judgments in controlled settings, establishing LLMaJ as a standard evaluation paradigm for multiple tasks (Liu et al., 2023b).

Despite known LLMs' prompt phrasing sensitivity (Wang et al., 2024; Liu et al., 2023a; Sclar et al., 2024; Webson & Pavlick, 2022), current research optimizes LLMaJ via rubric refinement, bias mitigation, or multi-agent aggregation (Ye et al., 2024; Chan et al., 2024; Mizrahi et al., 2024; Zhuge et al., 2025) rather than semantics of evaluation instructions. Unlike prior work on prompt enhancement, we isolate variance arising from paraphrastic prompt variation. We term this the Prompt Lottery: the phenomenon whereby evaluation outcomes depend on prompt formulation choices that, while intended to be semantically equivalent, produce systematically different (unpredictable to practitioners) judgments.

Variant A (P_1)

```
{dialogue}
Above is a dialogue context and the
corresponding response.
Question:  Is the response understandable?
###
Your Answer (respond with ONLY 'Yes' or 'No')
```

Variant B (P_2)

```
{dialogue}
Is the response in the above conversation
clear and easy to comprehend?
###
Answer with 'Yes' or 'No'.
```

An LLMaJ's sensitivity to prompt variation is an indicator of evaluator robustness. If a researcher concludes that a new model outperforms a baseline using a single prompt, but an equally valid

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

rephrasing reverses that conclusion, it reveals single-prompt evaluation to be fragile and misleading. Paraphrastic robustness can be viewed as a form of evaluation invariance: prompt variants intended to preserve semantic meaning should yield stable judgments. Violations of this property suggest that evaluation outcomes may depend partly on prompt formulation rather than only on the evaluated response. Prior work has quantified these effects in human annotation and developed strategies to ensure judgments reflect task semantics rather than syntactic framing (Artstein & Poesio, 2008). These mitigation strategies are grounded in a core requirement: semantically equivalent instructions should yield consistent judgments, a requirement central to principled evaluation.

To test this requirement, we conduct a large-scale evaluation across 825 conditions (spanning 12 datasets and 6 prompt families), followed up by a user study, assessing score variations by replacing human annotators with different model backbones. The results show that LLMaJ evaluators exhibit larger variance, up to 7$\times$ compared to the variance of human annotators in our study, as well as previously observed variance among human raters and generally acceptable level for human user studies, degrading from near-perfect to fair agreement. We demonstrate that this LLMaJ degradation is consequential, as it lowers human alignment and triggers model rank reversals in objective tasks. This prompts us to investigate how to improve the robustness of LLMaJ and we introduce an efficient aggregation strategy that restores reliability while reducing inference costs compared with naive aggregation (2 calls instead of 5 on stable instances), while recovering 86% of full-aggregation decisions.

2 Related Work

Quality human annotation relies on abstracting from surface-level variations to capture underlying semantic intent. This objective is rooted in the distinction between linguistic forms and their core semantic content (Chomsky, 1965), and is enacted through rigorous annotation protocols designed to maximize inter-coder agreement (Artstein & Poesio, 2008; Pavlick & Kwiatkowski, 2019). As human judgments can be susceptible to the Framing Effect in heuristic decision-making (Tversky & Kahneman, 1981), controlled evaluations utilize coding instructions to remove these biases and ensure reliability (Krippendorff, 2018). If LLMaJ pipelines are to replace human evaluators, they must satisfy this same invariance property. A more detailed theoretical analysis of formal semantics and exploring why invariance should or should not hold for LLMs is provided in Appendix C.

LLMaJ emerged as time and resource saving approach, and is now widely used for both reference-based (ground truth matching) and reference-free evaluation (free observation) tasks, often employing pairwise ranking or rubric-based scoring (Zheng et al., 2023; Liu et al., 2023b). Modern frameworks incorporate agentic reasoning or multi-step pipelines to approximate human judgment processes (Zhuge et al., 2025). Despite success, their sensitivity to prompt design raises questions about their stability as evaluation instruments. Recent surveys structure LLMaJ research along four axes: (i) target attributes, (ii) evaluation protocols, (iii) alignment with human judges, and (iv) bias analysis (Gu et al., 2024; Gao et al., 2025; Chang et al., 2024). While these emphasize scaling and aggregation strategies, they largely overlook the effect of initial evaluator prompt formulation. Our work complements this line by focusing on stability under syntactic variation rather than optimization under aggregation.

(Zheng et al., 2023) show that targeted prompt modifications, such as a judge persona, score delimiters, and chain-of-thought reasoning, improve human alignment and reduce position and verbosity biases. Beyond this several works reduce judge bias through aggregation: model ensemble multiple back-models (Williams et al., 2025), while (Li et al., 2025) generates augmented instructions and verifies consistency using a single model. Hierarchical tie-breaking resolves disagreements via a secondary model (Badshah et al., 2025), and (Zhu et al., 2026) rotate models over a fixed prompt to cancel systematic biases. (Haldar & Hockenmaier, 2025) have also identified substantial self-inconsistency in LLM judges across repeated evaluations, even under fixed prompts. Closest settings to ours are, (Mizrahi et al., 2024) who advocate multi-prompt evaluation using format perturbations (e.g., quotes, line breaks) on fixed instructions, (Deng et al., 2026) extend this to rubric scoring with synonym and adversarial variations, and (Chiang & Lee, 2023) who measure human rating averages across instructions. In contrast, we examine *paraphrastic variation*, where semantic equivalence is guaranteed by construction rather than formatting. While (Bellibatlu et al., 2026) proposed a reliability score under paraphrastic rephrasings, they only evaluate models relative to each other and do not assess the system-wide robustness of LLMaJs.

Table 1: Overview of judge families, evaluation tasks, and datasets.

Judge Family	Dimensions	Scoring Task	Datasets	References
Dialogue Quality	*Interesting, Specific, Understandable, Quality*	**Binary (Yes/No):** Presence of quality traits in conversation context	PersonaChat, PersonaUSR, DailyDialog, Topical-Chat, ConTurE	(Mehri & Eskenazi, 2020b; Zhang et al., 2018; Mehri & Eskenazi, 2020a; Li et al., 2017; Gopalakrishnan et al., 2019; Ghazarian et al., 2022)
QA Correctness	*Correctness*	**Likert 1–5:** Completeness and accuracy vs. ground-truth	HotpotQA, SQuAD v2, NQ-Open, TopiOCQA, Natural Questions	(Zheng et al., 2023; Yang et al., 2018; Rajpurkar et al., 2018; Lee et al., 2019; Adlakha et al., 2022; Kwiatkowski et al., 2019)
QA Groundedness	*Groundedness*	**Likert 1–5:** Faithful inference from evidence	HotpotQA, SQuAD v2, NQ-Open, TopiOCQA, Natural Questions	(Es et al., 2024; Yang et al., 2018; Rajpurkar et al., 2018; Lee et al., 2019; Adlakha et al., 2022; Kwiatkowski et al., 2019)
General Response Quality	*Helpfulness, Accuracy, Depth*	**Likert 1–10:** Holistic assessment of open-ended responses	LLM-Arena, PersonaChat, PersonaUSR, DailyDialog, Topical-Chat, ConTurE	(Zheng et al., 2023; Zhang et al., 2018; Mehri & Eskenazi, 2020a; Li et al., 2017; Gopalakrishnan et al., 2019; Ghazarian et al., 2022)
Coherence, Consistency, Fluency	*Summary Quality*	**Likert 1–5:** Summarization quality assessment	SummEval	(Fabbri et al., 2021)
Semantic Alignment	*Content & Style Equivalence*	**Likert 1–5:** Information overlap between reference and prediction	UPHELD	(Subasic et al., 2026)

3 Experimental Setup

In total we evaluate 33 configurations (12 datasets, 6 judge families) with 5 backbone models, drawn from both proprietary and open providers, and 5 prompt variants each, yielding 825 experimental conditions, as not all judge families apply to every dataset (e.g., summarization judge is used only for summarization tasks). We evaluate five backbone LLMs: GPT-4o, gpt-oss-120B, Llama 3.3 (70B Instruct), Kimi-K2.5(-Instruct), and Qwen3(-VL-30B-A3B-Instruct). For each judge family, we use the original prompt and four semantically equivalent variations (Section 3.1), yielding 25 distinct configurations per dataset. (Li et al., 2026) argue that non-zero temperatures capture fuller evaluative potential; however, our focus is the model's capacity for semantic consistency. We therefore use temperature $T = 10^{-5}$ to ensure results reflect core interpretive stability rather than "wisdom-of-the-crowd" effects, though GPU kernel non-determinism may remain (He & Lab, 2025).

We utilize judge families grounded in established evaluation frameworks (Table 1), encompassing G-Eval style multidimensional rubrics (Liu et al., 2023b) and specialized metrics. These span binary classification and Likert-scale scoring across objective (Correctness, Groundedness) and subjective (Dialogue Quality, General Response Quality) axes, demonstrating the Prompt Lottery's impact across the wide spectrum of LLM evaluation. We prioritized datasets with accessible codebases and prompt templates to ensure fidelity to original conditions and eliminate confounding from manual reconstruction. The prompts and paraphrastic variants are provided in Appendix H.

We examine whether paraphrastic rephrasing induces significant variance in judgment scores using four metrics: *Overlap* (rate of identical judgments), *Average Pairwise Cohen's* κ, *Fleiss'* κ, and *Krippendorff's* α (agreement across all rephrased prompts). We report all metrics but focus our analysis on κ scores for direct comparison with prior work on human judgment under instruction variation, while also noting high correlation with Krippendorff's α and Fleiss' κ.

3.1 Paraphrastic Prompt Generation and Validation

We generated four paraphrastic variants using Claude 3 Opus. These were reviewed by human annotators for preservation of intended task meaning and verified using Gemini 1.5 Pro for syntax validity. To characterize the variations qualitatively, we post hoc identified six main recurring, but not mutually exclusive dimensions of surface variation, including role framing, instruction syntax, contextualization, and output specification (Table 10); A within-family permutation test of these dimensions against cross-model Cohen's κ (50 judge prompts, 46 experiments, 10,000 permutations per contrast, Holm-adjusted) found no dimension that consistently produced higher agreement across

models (Appendix B contains full details). First, three independent LLMs (Kimi-K2.5, Gemini 1.5 Pro, and GPT-4o) evaluated pairwise prompt similarity, yielding average scores of 81%, 84%, and 79%, respectively. Second, sBERT semantic similarity (Reimers & Gurevych, 2019) averaged 0.93 across all prompt pairs, exceeding a commonly used high-similarity threshold of 0.80. In contrast, lexical overlap remained comparatively low (Jaccard: 0.38 ± 0.01). High semantic similarity and low lexical overlap suggest that prompt variants preserved intended evaluation content while differing substantially in wording, consistent with successful paraphrastic rephrasing.

To strengthen these findings, we conducted a human validation study focused on task equivalence rather than surface-level similarity. Each pair of generated prompts (including the original prompt) was shown to three annotators, who rated whether the two prompts require the same evaluation task. Annotators consistently judged the prompts to be highly equivalent, with an average POMP (Cohen et al., 1999) score of 76/100. Inter-annotator agreement was substantial (Cohen's $\kappa = 0.78$; Krippendorff's $\alpha = 0.81$), indicating that this perception of equivalence was consistent across raters.

3.2 Annotator Alignment Study and Expected Agreement Levels

While humans perceive the instruction prompts we created as equivalent tasks, prior work on annotation guideline rephrasing suggests this may not guarantee identical annotation behavior (Benikova et al., 2014). To test this, we conducted a user study in which we recruited 10 annotators (at fair market rate) via an freelancer agency. We selected annotators with high success rates, indicating professionalism and skill. They were presented with randomized pairs of instructions for scoring answers, the answers, and context (e.g. conversation so far, or a question). The instructions were identical to the paraphrased prompt variants. Annotators provided individual ratings, and we ensured annotators never rated the same data point with different instructions back-to-back. In total we collected the annotation on 650 data points, and each annotator sequentially rated examples, with each example annotated by at least three annotators. Based on these annotations, we calculated agreement metrics for humans. On average across prompts and datasets, we observed Overlap, Average Pairwise Cohen's κ, Fleiss' κ, and Krippendorff's α (all normalized to 0–1) of 0.78, 0.71, 0.80, and 0.63, respectively. Previous work findings are consistent with our study: (Benikova et al., 2014) found that paraphrasing guidelines reduced inter-annotator agreement from $\kappa = 0.81$ to 0.72 ($\Delta\kappa \approx 0.09$), despite annotators confirming task equivalence. Similarly, (Wiebe et al., 2005) documented κ reductions of ≈ 0.11.

If LLMaJ is to replace and/or improve the performance of human annotators, a reasonable expectation is that it achieves the same level of robustness as humans. This sets the first baseline "human robustness" at $\kappa = 0.80(\Delta 0.07)$ based on the user study described above. However, since LLMs exhibit inherent instability due to floating-point operations, we must establish a stability baseline: what are the reasonable expectations if we do not change the prompts? To establish this stability baseline, we conducted five identical runs over 100 samples per model using fixed prompts. Results showed 89.8% overlap, with Cohen's $\kappa = 0.86$ and Fleiss' $\kappa = 0.89$, indicating "almost perfect agreement" (Cohen, 1960). In contrast, the same runs with $T = 0.7$ produce lower agreement (Fleiss' $\kappa = 0.39$) This establishes the expected upper bound when the Prompt Lottery is absent and confirms that subsequent instabilities stem from prompt sensitivity rather than stochastic artifacts. The final, most permissive baseline is set at $\kappa > 0.61$, a generally accepted agreement threshold in usability studies per (Artstein & Poesio, 2008; Barth et al., 2017).

4 Experiments

We set up experiments to: (1) quantify judge alignment and the existence of the "Prompt Lottery", (2) explore its provenance, (3) its impact on judgment quality, and (4) present mitigation strategies.

4.1 Agreement Collapses Under Prompt Variation

We observe that the agreement degraded from near-perfect (stability and human baseline) to fair (κ: $0.89 \rightarrow 0.24$), a drop substantially larger than the 0.09–0.11 degradation we observe in the human studies (similar in literature and our study). Differences are statistically significant at $p < 0.05$ in aggregate, with 75% of individual dataset-judge configurations showing significant differences. This magnitude of sensitivity points to evaluation reliability issues in LLMaJ evaluation setups compared

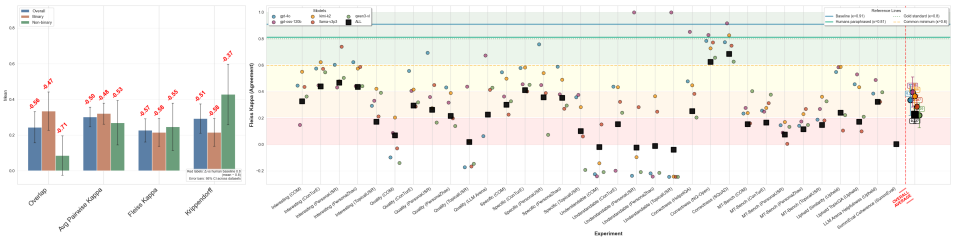

Figure 1: On the left bar-chart: agreement metrics demonstrate low consistency across syntactic variants both across models and prompts. On the right we show Fleiss' κ values and degradation below baselines across judge families for each model. Appendix E provides details.

to human annotation protocols. We also observe that on average all but two cases (NQ-Open Correct and SQUAD Correct) fall below what would be a general agreement level for valid annotation studies ($\kappa < 0.61$). This stands for all metrics (Figure 1 right), with the highest average agreement, measured by Krippendorff's α for non-binary annotations still falls well below both human and theoretical baselines. The agreement is fairly stable, with a 95% confidence interval half-width of 0.065 on average.

4.2 Instability Depends on Task Type and Backbone

The Binary Paradox. Binary tasks show higher raw overlap (0.32 vs. 0.14 for non-binary), suggesting LLMs frequently output identical scores by chance. However, chance-corrected reliability (Fleiss' $\kappa = 0.214$, $\alpha = 0.217$) is lower than for non-binary tasks ($\kappa = \alpha = 0.28$), indicating the output space constraint masks underlying instability rather than indicating true robustness.

Model-Dependent Sensitivity. While all models exhibit the Prompt Lottery, magnitude varies. Kimi-K2.5 and GPT-4o exhibited greater robustness than Qwen3-VL and Llama 3.3, though all models showed substantial degradation from baseline. [1] Objective judges (QA Correctness, Groundedness) show higher robustness (κ=0.47–0.69) compared to subjective and open-ended judges (κ=0.28–0.39), suggesting concrete evaluation criteria are inherently more stable under surface variation.

Newer models. Reasoning-augmented models introduce internal stochasticity orthogonal to temperature control (tree search, best-of-N reasoning). This makes temperature-controlled comparison with direct- scoring models problematic. We show that the Prompt Lottery is a universal property of LLMaJ evaluation across both older and frontier models, rather than an artifact of a specific model generation. We compared GPT-4o and GPT-5.1 scores across identical tasks (5 prompts each, 3 datasets, 400 examples). Inter-model agreement (Cohen's κ) shows a consistent pattern: objective tasks (κ=0.82), dialogue (κ=0.68), summarization (κ=0.25). This suggests that both models struggle equally on summarization despite architectural differences, suggesting the bottleneck is task-inherent difficulty which is not solved by more powerful models.

Table 2: Correlation between prompts' $sBERT$ Similarity and Cohen's Kappa.

Similarity	Pearson r	p	Spearman ρ	p
Average	-0.11	0.157	-0.15	0.022
Min	-0.08	0.25	-0.06	0.37
Max	0.05	0.44	-0.07	0.30

Table 3: Pearson correlation (r) between the default (p_1) and the top-performing prompts.

Dataset	**Top-1**	**Max/p_1 r**	**Rank**
Upheld	P_5	0.44/0.42	4 / 5
Topical-Chat	P_4	0.49/0.47	4 / 5
LLM-Arena	P_5	0.49/0.49	3 / 5

4.3 No Prompt Wins Consistently—It's a Lottery

If LLMaJ relied primarily on semantic content (rather than surface form), high semantic similarity should predict high agreement. However, we observe sBERT similarities of $0.92-0.95$ yet agreement (κ) ranges from 0.0 to 0.7, suggesting LLMs are sensitive to non-semantic surface features. The

[1] Kimi-K2.5: $\kappa = 0.47$; GPT-oss-120B: $\kappa = 0.44$; GPT-4o: $\kappa = 0.42$; Llama 3.3: $\kappa = 0.34$; Qwen3-VL: $\kappa = 0.33$

correlation between semantic similarity and scoring stability was weak and negative overall $r = -0.11, p = 0.157$ (Table 2), indicating that increasing semantic similarity does not improve judgment consistency. This relationship remained weak and inconsistent even when restricting analysis to the most- and least-agreeable prompt subsets, suggesting that semantic similarity is insufficient to explain evaluation stability. Critically, no single prompt variant consistently outperforms others across datasets (Table 3). The top-performing variant differs per task (P_5 on Upheld, P_4 on Topical-Chat, P_5 on LLM-Arena), and average absolute correlation differences are minimal (δ r = 0.00–0.02). This absence of a universal "best" prompt suggests instability does not stem from systematic prompt content differences (e.g., one being semantically clearer or more complete), but rather from LLM sensitivity to superficial syntactic variation that is fundamentally unpredictable across contexts

If default prompts are well optimized, they should achieve the highest alignment with human judgments and the observed lack of agreements has no bearing; the conclusion would be that the LLM practitioners have already optimized the prompts to account for variations. If alternative paraphrastic variants consistently outperform, this would suggest sensitivity to prompt formulation, and that there is an element of chance associated with single prompt LLMaJ setups. To measure this we followed an established protocol in (Liu et al., 2023b) and calculated Pearson correlation (r) between human and paraphrased prompts judgments. For datasets for which human annotations are publicly accessible (Upheld, Topical-chat, LLM-Arena) the results in Table 3 show that the latter is the case. Paraphrastic variations constantly match or outperform P_1 (default) prompts. While the default prompt ranked 3rd-4th across datasets, absolute differences in correlation were small (Δr = 0.00-0.02), and only 2 of 3 comparisons reached significance ($p < 0.05$, uncorrected). Bootstrap rank CIs for human alignment overlap fully across all five prompts, and the default prompt is ranked first in only 5–24% of resamples depending on the human aggregation. The results in Section 4 suggest that this alignment may be an artifact of aggregation. The fact that divergent individual judgments can produce similar aggregate correlations may obscure instability in the evaluation process. This may reconcile our findings with earlier reports of instruction robustness in LLM evaluation (Chiang & Lee, 2023).

4.4 Do Prompts Reverse Model Rankings?

While the Prompt Lottery substantially alters absolute scores and agreement with humans seems to be largely unaffected, a critical question is whether it also reverses *relative model rankings*. This distinction matters: score variance may reduce confidence in conclusions, but ranking reversals directly contradict them. To test whether this instability affects comparative evaluations, we evaluated four inference models (GPT-4o, gpt-oss-120B, Kimi-K2.5, and Qwen3p6-Plus) under five paraphrastic prompt variants using a fixed backbone model. All ranking stability statistics were computed over 100 evaluation samples per task and averaged across three independent folds to reduce sampling variance and isolate prompt-induced ranking instability.

Ranking reversals are strongly task-dependent. Objective evaluation (SQuAD v2 correctness) exhibits severe instability ($\tau = 0.176$; model rank positions change in 43.3% of prompt comparisons). In contrast, subjective dialogue evaluation (MTBench) remains perfectly stable (0% reversals). Intermediate behavior appears in SummEval and persona-based dialogue evaluation, where rankings remain partially stable but top-performing systems still change depending on prompt wording. SQuAD v2 exhibits severe instability: GPT-4o and Qwen3p6-Plus exchange top rankings across prompt variants, with GPT-4o ranging from 1st to 3rd and Qwen3p6-Plus from 1st to 4th. The identity of the "best" model thus depends substantially on prompt formulation. Figure 10 in Appendix F visualizes these dynamics.

Table 4: Ranking stability across paraphrastic prompts. Lower Kendall's τ and higher absolute reversal rates indicate stronger Prompt Lottery effects.

Task	**Mean τ**	**Top-1 Flips**	**Flip %**
MTBench (DailyDial.)	1.00	0/4	0.0%
MTBench (Persona)	0.80	2/4	10.0%
QA Correct. (SQuAD v2)	0.17	3/4	43.3%
Quality (SummEval)	0.64	2/4	40.0%
G-Eval (Topical-Chat)	0.83	1/4	10.0%

5 MITIGATING VIA AGGREGATION

Prior work has implicitly addressed judge instability via LLMaJ aggregation (Zhao et al., 2025; Mizrahi et al., 2024). Motivated by this, we evaluate which aggregation strategy across paraphrastic prompt variants best aligns with human judgments using the three datasets with public human annotations. We compare **Average**, **Majority** (mode), **Plurality** (requiring $> 50\%$ consensus) , and **Plurality+Avg. (PA)** (using average when plurality is not reached) using Pearson correlation (r) with human scores. Table 5 shows Plurality consistently achieves the strongest human alignment, improving over both averaging and majority voting. This points to the possible strategy, often used in human studies, where data points are discarded if they do not reach a certain threshold of agreement.

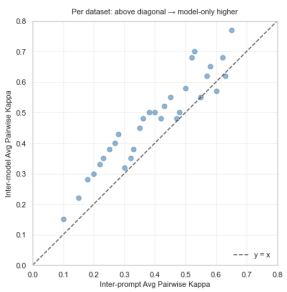

Figure 2: Model vs prompt.

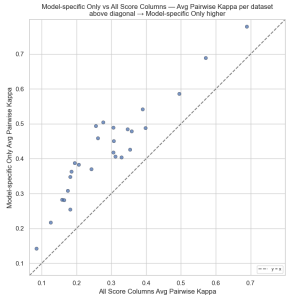

Figure 3: Model vs hybrid.

Table 5: Pearson correlation (r) between human scores and LLMaJ aggregation strategies ($p < 0.05$).

Dataset	Maj.	Avg.	Plur.	PA.
Upheld	0.47	0.54	0.63	0.61
→95% CI	[.42,.57]	[.48,.62]	[.52,.69]	[.48,.62]
Topical-Chat	0.44	0.55	0.62	0.51
→95% CI	[.36,.48]	[.50,.59]	[.57,.69]	[.43,.53]
LLM-Arena	0.44	0.54	0.55	0.52
→95% CI	[.32,.48]	[.47,.57]	[.47,.59]	[.45,.53]

5.1 WHAT SOURCE OF AGGREGATION MATTERS?

Having established that aggregation improves alignment with human judgments, we next ask which source of diversity contributes most to these gains. Prior work has largely emphasized model diversity (e.g., through ensembles or debate-style systems), while our findings suggest that prompt variation itself may provide a complementary source of robustness. To examine this, we compare three aggregation strategies: (1) **Inter-Model** (varying models, fixed prompts), (2) **Inter-Prompt** (varying prompts fixed models), and (3) **Hybrid** (varying both models and prompts). Figures 2 and 3 show that inter-prompt aggregation achieves stronger reliability compared to the other two, suggesting that prompt variation may contribute more strongly to robustness than model diversity alone. Our aggregation strategies assume the five paraphrased-prompt judgments are conditionally independent given the true label. Because all five variants share the same backbone model and task, this assumption may not hold in practice; dependence-aware aggregation methods Balasubramanian et al. (2026) show that independence-assuming rules can be suboptimal.

5.2 SELECTIVE AGGREGATION: TASK-DEPENDENT COST SAVINGS

We analyze when the default prompt aligns with broader consensus. Binary tasks show high stability and matches plurality consensus in 94.7% of cases, making multi-prompt aggregation marginally useful due to limited output space. Non-binary tasks exhibit the opposite, and across all configurations, strict plurality consensus ($\geq 50\%$ agreement) emerges in only 58.73% of cases, indicating that many examples fail to converge even after evaluating five prompt variants. Moreover, the default prompt aligns with the plurality decision in only 44.0% of cases, implying that more than half of single-prompt evaluations disagree with the consensus whenever prompt variation is considered. These results suggest that single-prompt evaluation can diverge substantially from aggregated judgments. Consensus statistics in non-binary tasks suggest many examples either converge early or fail to converge entirely: 37.68% reach plurality by the third judge, 21.05% require all five, and 41.27% never converge. This distribution suggests that substantial inference cost is spent on examples that either converge early or fail to converge entirely. We formulate selective aggregation as a binary classification task: given a prompt and an input evaluated, predict whether the initial judgment will match the plurality consensus (no aggregation needed, label=0) or whether additional prompt variants should be evaluated (label=1). Labels derive from observed plurality outcomes in our experiments.

We evaluated multiple classifiers for selective aggregation, including transformer encoders, classical models, and zero-/few-shot LLM baselines. Architecture selection (Table 6) was performed on a

Table 6: Model ablation results on a smaller dataset with prompts not in the training data.

Model	Bal. acc	TNR
DeBERTa	0.836	0.854
RoBERTa	0.914	0.845
Logistic	0.862	0.751
SVM	0.854	0.742
Llama- zero-shot	0.489	0.042
5-shot	0.451	0.322

Table 7: Cost-benefit strategy comparisons. Worst-case are scores with no plurality consensus. C stands for compute power. $Plu.\,\%$ stands for the data points with plurality decision.

Strategy	C	$Plu.\,\%$	**Worst-case**
Single LLMaJ	$1\times$	44%	–
Task heuristic	~ 4.5	74%	69%
DeBERTa-gate	$\sim 2\times$	86%	65%
Full aggregation	$5\times$	100%	79%

smaller default splits; and the selected architecture was then retrained on $N = 369{,}625$ samples with hyper-parameters selected via Optuna (Akiba et al., 2019) over 35 trials, and evaluated on a held-out test set of $n = 73{,}925$ samples. Logistic Regression and SVM serve as baselines to assess whether prompt stability can be captured with lightweight decision boundaries, while zero-/few-shot LLM variants represent a potentially more expressive alternative that may capture semantic signals at higher computational cost. While RoBERTa (Liu et al., 2019) achieved the highest balanced accuracy, we selected DeBERTa-V3 (He et al., 2023) because our objective is to maximize true negative rate (TNR), rather than overall performance. The final classifier achieves TPR 85.06% and TNR 69.42%, deliberately optimizing in this way as false positives incur only inference cost, while false negatives corrupt evaluation quality. To assess generalization, we evaluate on a held-out dataset with Prometheus judges (Kim et al., 2024) whose prompts were absent from training. Performance suggests the classifier captures signals that generalize beyond the training prompts and datasets with TNR stable at 66.22%. More details on model training is reported in Appendix G.

In our selective aggregation framework, cost savings occur when stable instances are correctly identified as negatives, allowing aggregation to be skipped. False positives unnecessarily trigger additional LLM calls and reduce efficiency gains. We compare against a **task-type heuristic**: always aggregate non-binary tasks (44% instability), never aggregate binary tasks (94.7% stability). While this strategy is cheaper than full aggregation for stable tasks, it wastes compute on non-binary instances and misses unstable binary cases. Our DeBERTa approach provides instance-level granularity that the heuristic lacks. As shown in Table 7, DeBERTa-gating reduces LLM calls by approximately 60% compared to naive aggregation ($2 \leftarrow 5\times$ compute) while recovering 86% of full aggregation decisions. Performance declines to 65% only in difficult no-consensus cases, representing the maximum observed degradation; typical performance is substantially higher. Full aggregation achieves 79% even in worst-case settings because all prompt variants are evaluated directly.

5.2.1 Derivation of LLM Call Reduction Under Selective Aggregation

Step 1: Expected calls under full aggregation. From the empirical distribution of convergence times reported in Section 5, non-binary tasks require the following number of judge calls to determine the plurality outcome. Plurality is reached at three judge calls for 37.68% cases, at five calls 21.05%, and 41.27% of cases require more than five calls to reach plurality.

$$\bar{t}_{\text{full}} > 0.3768 \times 3 + 0.2105 \times 5 + 0.4127 \times 5 = 4.247 \tag{1}$$

Step 2: Expected calls. The classifier achieves a TNR of 69.42%. The remaining go through the aggregation pipeline.

$$\bar{t}_{\text{gated}} = \underbrace{\text{TNR} \times 1}_{\text{single-prompt-cases}} + \underbrace{(1 - \text{TNR}) \times \bar{t}_{\text{full}}}_{\text{aggregated-cases}} = 0.6942 + 0.3058 \times 4.247 \approx 1.99 \tag{2}$$

Step 3: Call reduction. Relative to full aggregation ($\bar{t}_{\text{full}} = 4.247$):

$$\Delta_{\text{vs. full}} = 1 - \frac{\bar{t}_{\text{gated}}}{\bar{t}_{\text{full}}} > 1 - \frac{1.99}{4.247} \approx 53\% \tag{3}$$

Relative to the naïve five-prompt baseline ($\bar{t}_{\text{naive}} = 5.0$) the $\Delta_{\text{vs. naive}}$ would be $\approx 60\%$.

6 Limitations and Final Remarks

We find that LLMaJ outputs are sensitive to paraphrastic prompt variation, affecting absolute scores and relative model rankings. We further show that this instability can reverse relative rankings of evaluated models, particularly in objective QA settings, demonstrating that prompt choice may alter both absolute and comparative evaluation conclusions. We show that plurality aggregation improves agreement with humans compared to single-prompt evaluation. By computing consensus across semantically equivalent prompts, the plurality metric filters pragmatic noise from sub-optimal variations. This significantly improves alignment with human judgments, confirming that LLMaJ reliability is maximized when multiple equivalent prompts converge. While aggregation improves robustness, it increases computational cost. Our selective aggregation approach makes this practical by predicting judgments requiring additional evaluations. Using a DeBERTa classifier, on stable instances, we reduce LLM calls from 5 to 1.99 transforming multi-prompt evaluation into a viable cost-effective default strategy.

Scope and Limitations. Our findings are limited to paraphrastic variations that preserve semantic content; we do not test deliberate semantic modifications (e.g., changing evaluation criteria), which are addressed elsewhere. Semantic equivalence is validated via sBERT similarity (0.93) and human task equivalence ($\kappa = 0.78$) rather than formal logical entailment. Additionally, our main experiments use five model families, with limited coverage of reasoning-augmented models; a preliminary comparison of GPT-4o and GPT-5.1 (400 examples, 3 datasets) suggests Prompt Lottery persists across architectures. However, the magnitude of degradation ($\kappa : 0.89 \rightarrow 0.24$) substantially exceeds human variability ($\Delta\kappa = 0.09$), and findings persist across all experimental conditions, suggesting Prompt Lottery effects are robust despite these constraints. Our mitigation strategy is evaluated against single-prompt (lower bound) and full aggregation (upper bound) baselines. We do not compare against budget-matched alternatives (e.g., single-prompt self-consistency sampling) or orthogonal robustness techniques (ensembles, debate, instruction tuning), so the gains in Section 5 may partly reflect repeated judging rather than prompt diversity per se; this analysis is reserved for future work.

From a theoretical standpoint, these results reveal a departure from *semantic invariance*, a foundational principle from formal semantics predicting that rational judges should produce equivalent outputs for semantically equivalent inputs. This pragmatic behavior, while potentially adaptive for dialogue understanding, creates brittleness in evaluation. Taken together, our findings support the central methodological message of this paper: evaluation should move beyond the single-prompt paradigm and toward systematic multi-prompt assessment. Robustness to prompt formulation is not an optional refinement, but a prerequisite for trustworthy LLMaJ evaluation and reliable comparative model ranking. For practitioners, we summarize these findings in Table 8; for researchers, it means treating prompt sensitivity as a first-class concern in evaluation methodology, not an afterthought.

Reproducibility Statement. All resources sufficient to reproduce result reported are added as supplementary material. We withhold public release until after review to preserve anonymity. Exact reproduction of raw judge scores may not be possible for closed or updated endpoints.

Ethical Statement. We recognize that LLMaJs risk reducing data annotation roles, and that this work may lead towards putting more stress on those roles. Likewise, this paper points to a theoretical and empirical issue with replacing human annotators with purely automatic judges and cautions against the use of fast and often unsound methodologies in order to cut human labor costs.

Table 8: Takeaways for robust LLMaJ evaluation.

Factor	Key Finding
Backbone model choice	Largest source of pairwise agreement variance
Prompt rephrasing	Significant variance despite semantic equivalence
Binary evaluation tasks	High stability, skip aggregation
Non-binary evaluation tasks	High instability, aggregate
Plurality aggregation	Strong robustness improvement
Selective aggregation	Major cost reduction ($\sim 60\%$)

AI USE DISCLOSURE

In this work, we used generative AI tools for:

- Polishing the writing, fixing LaTeX formatting, correcting typos, improving unclear or ambiguous sentences, and formatting tables, figures, and equations.
- Drafting select sections based on initial human outlines and concepts, followed by iterative human review, correction, and refinement.
- Implementing certain calibration methods using AI coding assistants (Claude Code and Cursor), followed by thorough human code review and test-case verification.
- Help with literature review searches (paired with standard search engines) to identify related works within the rapidly growing body of literature.
- Data analysis and chart generation assistance using Cursor and Google NotebookLM.
- As a part of paraphrastic rephrasing of evaluation prompts.
- Parts of the pipeline for interacting with LLM APIs was developed using AI assistants.

We did not use generative AI tools for:

- Setting up the two user studies we report on in the paper. And we have extensively checked annotator inputs for not permitted LLM use by the annotators.
- Conceptualizing or designing the research ideas and experimental frameworks.
- Autonomous agentic workflows where AI generated content from broad goals without direct human oversight.
- Generating synthetic data (all datasets used are publicly available).
- Fully autonomous section writing. No text was generated without initial human input and extensive subsequent editing of AI suggestions or edits.

We take full responsibility for the content of this work, including all text, claims, and artifacts produced with the aid of generative AI tools.

REFERENCES

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. TopiOCQA: Open-domain conversational question answering with topic switching. *Transactions of the Association for Computational Linguistics*, 10:468–483, 2022. doi: 10.1162/tacl_a_00471.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

Ron Artstein and Massimo Poesio. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008. doi: 10.1162/coli.07-034-R2. URL https://aclanthology.org/J08-4004/.

Sher Badshah, Moamen Moustafa, and Hassan Sajjad. CLEV: LLM-based evaluation through lightweight efficient voting for free-form question-answering. In Kentaro Inui, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dhirendra Pratap Singh (eds.), *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 1513–1531, Mumbai, India, December 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. ISBN 979-8-89176-303-6. doi: 10.18653/v1/2025.findings-ijcnlp.93. URL https://aclanthology.org/2025.findings-ijcnlp.93/.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, and others... Constitutional AI: Harmlessness from ai feedback, 2022. URL https://arxiv.org/abs/2212.08073.

Krishnakumar Balasubramanian, Aleksandr Podkopaev, and Shiva Prasad Kasiviswanathan. Dependence-aware label aggregation for llm-as-a-judge via ising models. *arXiv preprint arXiv:2601.22336*, 2026.

Jürgen Barth, Wout de boer, Jason Busse, Jan Hoving, Sarah Kedzia, Rachel Couban, Katrin Fischer, David von Allmen, Jerry Spanjer, and Regina Kunz. Inter-rater agreement in evaluation of disability: Systematic review of reproducibility studies. *BMJ: British medical journal*, 356, 01 2017. doi: 10.1136/bmj.j14.

Rohith Reddy Bellibatlu, Edward Raff, and Wenbin Zhang. Judgesense: A benchmark for prompt sensitivity in llm-as-a-judge systems. *arXiv preprint arXiv:2604.23478*, 2026.

Darina Benikova, Chris Biemann, and Marc Reznicek. NoSta-D German Named Entity Corpus: Building a German NER Corpus for GermEval 2014. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, pp. 2328–2333, 2014.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better LLM-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=FQepisCUWu.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Weiqiang Xu, and others... A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.

Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15607–15631, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.870. URL https://aclanthology.org/2023.acl-long.870/.

Noam Chomsky. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA, 1965.

Alonzo Church. A formulation of the simple theory of types. *Journal of Symbolic Logic*, 5(2):56–68, 1951. doi: 10.2307/2266170.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Patricia Cohen, Jacob Cohen, Leona S Aiken, and Stephen G West. The problem of units and the circumstance for pomp. *Multivariate Behavioral Research*, 34(3):315–346, 1999.

Haotian Deng, Chris Farber, Jiyoon Lee, and David Tang. Rubric-conditioned llm grading: Alignment, uncertainty, and robustness. *arXiv preprint arXiv:2601.08843*, 2026.

Daniel C. Dennett. *The Intentional Stance*. MIT Press, Cambridge, MA, 1987.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. RAGAS: Automated evaluation of retrieval augmented generation. In *Proceedings of EACL (Demonstrations)*, 2024.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021. doi: 10.1162/tacl_a_00373.

Gottlob Frege. *Die Grundlagen der Arithmetik: eine logisch mathematische Untersuchung über den Begriff der Zahl*. Wilhelm Koebner, Breslau, 1884. Translated as *The Foundations of Arithmetic* by J.L. Austin (1950), Philosophical Library.

Mingqi Gao, Xinyu Hu, Xunjian Yin, Jie Ruan, Xiao Pu, and Xiaojun Wan. Llm-based nlg evaluation: Current status and challenges. *Computational Linguistics*, 51(2):661–687, 2025.

Sarik Ghazarian, Behnam Hedayatnia, Alexandros Papangelis, Yang Liu, and Dilek Hakkani-Tur. What is wrong with you?: Leveraging user sentiment for automatic dialog evaluation, 2022. URL https://arxiv.org/abs/2203.13927.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pp. 1891–1895, 2019. doi: 10.21437/Interspeech.2019-3079. URL http://dx.doi.org/10.21437/Interspeech.2019-3079.

H. Paul Grice. *Logic and Conversation*, volume 3. Academic Press, New York, 1975.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, and others... A survey on llm-as-a-judge. *The Innovation*, 2024. URL https://arxiv.org/abs/2411.15594.

Rajarshi Haldar and Julia Hockenmaier. Rating roulette: Self-inconsistency in llm-as-a-judge frameworks. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *EMNLP 2025 - 2025 Conference on Empirical Methods in Natural Language Processing, Findings of EMNLP 2025*, EMNLP 2025 - 2025 Conference on Empirical Methods in Natural Language Processing, Findings of EMNLP 2025, pp. 24986–25004. Association for Computational Linguistics (ACL), 2025. doi: 10.18653/v1/2025.findings-emnlp.1361.

Horace He and Thinking Machines Lab. Defeating nondeterminism in llm inference, 2025. URL https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=sE7-XhLxHA.

Dell H. Hymes. Models of the interaction of language and social life. *Journal of Social Issues*, 23(2): 8–28, 1972.

Seungone Kim, Jay Shin, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Ryan Shin, and others... Prometheus: Inducing fine-grained evaluation capability in language models. In *International Conference on Learning Representations*, volume 2024, pp. 29927–29962, 2024.

Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, Thousand Oaks, CA, 4th edition, 2018.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019. Association for Computational Linguistics. URL https://aclanthology.org/P19-1612.

Jiajie Li, Huayi Zhang, Peng Lin, Jinjun Xiong, and Wei Xu. Auto-prompt ensemble for llm judge, 2025.

Lujun Li, Lama Sleem, Yangjie Xu, Yewei Song, Aolin Jia, Jerome Francois, and Radu State. The necessity of setting temperature in llm-as-a-judge. 2026. URL https://arxiv.org/abs/2603.28304.

Yankai Li, Hang Su, Xiaodong Shen, Wenhao Li, Zheng Cao, and Shijin Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*, 2017. URL https://arxiv.org/abs/1710.03957.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023a. doi: 10.1145/3560815.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized bert pretraining approach. 2019.

Shikib Mehri and Maxine Eskenazi. Unsupervised evaluation of interactive dialog with DialoGPT. In Olivier Pietquin, Smaranda Muresan, Vivian Chen, Casey Kennington, David Vandyke, Nina Dethlefs, Koji Inoue, Erik Ekstedt, and Stefan Ultes (eds.), *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 225–235, 1st virtual meeting, July 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.sigdial-1.28. URL https://aclanthology.org/2020.sigdial-1.28/.

Shikib Mehri and Maxine Eskenazi. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of ACL*, pp. 681–707, 2020b.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? a call for multi-prompt LLM evaluation, 2024. URL https://aclanthology.org/2024.tacl-1.52/.

Richard Montague. *Formal Philosophy: Selected Papers*. Yale University Press, New Haven, CT, 1974.

Allen Newell and Herbert A. Simon. *Computer Science as Empirical Inquiry: Symbols and Search*, volume 19. 1976. doi: 10.1145/360018.360022.

Barbara Hall Partee. Compositionality. In *Varieties of Formal Semantics*, pp. 281–312. Foris Publications, Dordrecht, 1984.

Ellie Pavlick and Tom Kwiatkowski. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694, 11 2019. doi: 10.1162/tacl_a_00293.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Melbourne, Australia, 2018. Association for Computational Linguistics. URL https://aclanthology.org/P18-2124.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. URL https://arxiv.org/abs/1908.10084.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, volume 2024, pp. 25055–25083, 2024.

Dan Sperber and Deirdre Wilson. *Relevance: Communication and Cognition*. Harvard University Press, Cambridge, MA, first edition, 1986.

Ilija Subasic, Andrew Rabinovich, and Zhao Chen. Evaluating language models in realistic conversational contexts. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Albert Webson and Ellie Pavlick. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2300–2344, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.167. URL https://aclanthology.org/2022.naacl-main.167/.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2):165–210, 2005.

Gwydion Williams, Samuel Rutunda, Floris Nzabakira, and Bilal A Mateen. Human evaluators vs. llm-as-a-judge: Toward scalable, real-time evaluation of genai in global health. *medRxiv*, 2025. doi: 10.1101/2025.10.27.25338910. URL https://www.medrxiv.org/content/early/2025/10/28/2025.10.27.25338910.

Zhilin Yang, Peng Qi, Saizheng Zhang, Edward Benson, William W. Cohen, and Ruslan Salakhutdinov. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018. Association for Computational Linguistics. URL https://aclanthology.org/D18-1259.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. Justice or prejudice? quantifying biases in llm-as-a-judge. 2024. URL https://arxiv.org/abs/2410.02736.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*, 2018. URL https://arxiv.org/abs/1801.07243.

Jitian Zhao, Changho Shin, Tzu-Heng Huang, Satya Sai Srinath Namburi GNVV, and Frederic Sala. From many voices to one: A statistically principled aggregation of llm judges. In *NeurIPS 2025 Workshop: Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*, 2025. URL https://openreview.net/forum?id=XdcofpTCyq.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023.

Ziyi Zhu, Olivier Tieleman, Alexey Bukhtiyarov, and Jinghong Chen. Cyclicjudge: Mitigating judge bias efficiently in llm-based evaluation, 2026. URL https://arxiv.org/abs/2603.01865.

Mingchen Zhuge, Changsheng Zhao, Dylan R. Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, Yangyang Shi, Vikas Chandra, and Jürgen Schmidhuber. Agent-as-a-judge: Evaluate agents with agents. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 80569–80611. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/zhuge25a.html.

A Mapping of Judge Families to Dataset and Tasks

Table 9: Overview of the 33 experimental configurations used in this study. Each configuration pairs a dataset with a judge family and is evaluated with $5 \times 5 = 25$ distinct LLMAJ triples (5 back-models $\times$ 5 prompt variants). *Binary?* indicates whether the judge produces a binary (Yes/No) rather than a Likert-scale score. Judge-family abbreviations—**DQ**: Dialogue Quality; **GQ**: General Quality (MT-Bench style); **QC**: QA Correctness; **QG**: QA Groundedness; **SA**: Semantic Alignment;

Dataset	**Judge family (dimension)**	**Binary?**	**Configs**	n
PersonaChat	DQ — Interesting	✓	1	149
	DQ — Quality	✓	1	149
	DQ — Specific	✓	1	149
	DQ — Understandable	✓	1	149
	GQ — MT-Bench		1	149
PersonaUSR	DQ — Interesting	✓	1	60
	DQ — Quality	✓	1	60
	DQ — Specific	✓	1	60
	DQ — Understandable	✓	1	60
	GQ — MT-Bench		1	60
Topical-Chat	DQ — Interesting	✓	1	60
	DQ — Quality	✓	1	60
	DQ — Specific	✓	1	60
	DQ — Understandable	✓	1	60
	GQ — MT-Bench		1	60
ConTurE	DQ — Interesting	✓	1	1,013
	DQ — Quality	✓	1	1,013
	DQ — Specific	✓	1	1,013
	DQ — Understandable	✓	1	1,013
	GQ — MT-Bench		1	1,013
DailyDialog	GQ — MT-Bench		1	100
HotpotQA	QC — Correctness		1	1,000
	QG — Groundedness		1	1,000
SQuAD v2	QC — Correctness		1	1,000
	QG — Groundedness		1	1,000
NQ-Open	QC — Correctness		1	1,000
	QG — Groundedness		1	1,000
SummEval	Coherence, Consistency, Fluency		3	160
Topical QA	SA — Content		1	1,000
LLM-arena	SA — Content		1	1,000
Upheld	SA — Content		1	1,000
12 datasets			**33**	

B Dimensions of Paraphrastic Variation

After examining paraphrastic variations post hoc across default prompts from all judge families, we distilled six representative dimensions. The six paraphrastic dimensions (Appendix B) were identified by post hoc inspection of the generated variants rather than imposed as generation constraints. These are shown in Table 10.

Table 10: Qualitative Analysis of Judge Prompts

Dimension	**Description**	**Examples**
1. Role Framing	How the judge's identity/authority is established.	"You are an expert evaluator" vs. "Act as a judge" vs. "You evaluate"
2. Instruction Syntax	Imperative vs. declarative vs. modal phrasing.	"Compare..." vs. "You should compare..." vs. "Your task is to compare..."
3. Criteria Presentation	How evaluation standards are formatted.	Bullet list vs. numbered list vs. prose vs. scale labels
4. Output Format Specification	Strictness of format requirements.	"strictly following this format" vs. "using exactly:" vs. "in the format:"
5. Task Contextualization	How much context/setup precedes the core task.	Verbose explanation vs. direct instruction
6. Linguistic Register	Formality level and vocabulary choices.	"assess" vs. "evaluate" vs. "rate" vs. "determine" / "brief" vs. "short" vs. "concise"

Each judge prompt was coded on the six dimensions using GPT-5.0 and matched to mean pairwise Cohen's κ among models scoring the same items with that prompt (50 prompts, 46 experiments, without G_{15}). Feature labels were permuted *within* each experiment–family block (10,000 draws per contrast), so the task, items, and model set stayed fixed; associations were averaged with equal weight across families and tested with Holm correction over the seven primary contrasts. Table 11 summarizes the results after residualizing out whether the prompt included a reference answer (which is not one of the six dimensions). Role, syntax, criteria layout, length, verb choice, and brevity all matched the permutation null; their eta-squared values (0.26–0.48) were comparable to the chance baseline for a random grouping of five paraphrases (null means 0.26–0.48). Output-format strictness was the only univariate slope that rejected the null ($+0.076$ κ per step on a four-level code from loose to "strictly following"; Holm $p = 0.0007$), but this did not survive as a format effect: the MT-Bench family written to span those phrases had slope $+0.005$, and holding the other coded features fixed reduced the slope to $+0.031$ ($p = 0.17$). Two listed framings ("You evaluate" and "you should compare") do not occur in this prompt set. Overall, none of the six dimensions identifies wording that consistently raises cross-model score agreement.

Table 11: Within-family permutation tests of paraphrastic prompt dimensions against cross-model Cohen's κ. Permutation p-values are two-sided for slopes and one-sided for eta-squared; Holm adjustment is over the seven contrasts.

Dimension	**Estimand**	**Value**	**Perm. p**	**Holm p**	**Families**
Role framing	Mean within-family η^2	0.34	0.34	1.00	8
Instruction syntax	Mean within-family η^2	0.32	0.78	1.00	8
Criteria presentation	Mean within-family η^2	0.26	0.45	1.00	7
Output-format strictness	Slope (κ per step)	$+0.076$	< 0.001	0.0007	4
Task contextualization (length)	Slope (κ per 100 words)	-0.044	0.63	1.00	10
Linguistic register (verb)	Mean within-family η^2	0.48	0.51	1.00	10
Linguistic register (brevity)	Mean within-family η^2	0.42	0.77	1.00	2

C Semantic Invariance in Language Understanding: Theoretical Foundations

A foundational principle in formal semantics is **semantic invariance**: two linguistic expressions with identical truth conditions should produce equivalent outputs in rational language-processing systems. This principle—rooted in (Frege, 1884), (Church, 1951), and (Montague, 1974)—motivates the expectation that LLM-as-a-Judge systems should be robust to paraphrastic (semantically equivalent) prompt variations. We briefly develop this principle and explain why the Prompt Lottery phenomenon represents a departure from these theoretical expectations.

C.1 Theoretical Basis for Semantic Invariance

Three theoretical commitments support the expectation of semantic invariance in language understanding systems:

Compositionality. Frege's principle of compositionality (Frege, 1884; Partee, 1984) states that the meaning of a complex expression depends on the meanings of its parts and the rules combining them. Under this view, two prompts with identical component meanings and combination rules should yield identical results. For example, a prompt asking "Is the response understandable?" and one asking "Is the response clear and easy to comprehend?" share the same propositional content, differing only in salience and cognitive framing.

Representation Independence. Church's thesis in semantics Church (1951) asserts that semantic content is independent of representation. A rational agent should produce equivalent outputs given two different encodings of the same semantic content. This principle motivated Montague's formalization of natural language semantics Montague (1974), where truth conditions are computed independently of surface form.

Rational Agent Theory. (Dennett, 1987) and (Newell & Simon, 1976) establish that a rational agent's outputs should depend on the content of its beliefs and goals, rather than their specific representation. Consequently, representation-invariant systems exhibit higher rationality and generalization.

C.2 Why Semantic Invariance Might Fail: Pragmatics

However, pragmatics—the study of meaning in context—provides complementary explanations for why semantic invariance should not be expected in real-world language processors:

Gricean Implicature. Grice's Cooperative Principle Grice (1975) characterizes communication as the cooperative coordination of speaker intent. Deviations from Gricean maxims (quantity, quality, manner, and relevance) generate implicatures: listener inferences about unspoken speaker intent. A prompt that elaborates "clear *and easy* to comprehend" implicates that comprehensibility is non-obvious, which subtly raises evaluation thresholds. LLMs sensitive to implicature will diverge on such variants despite identical semantic content.

Cognitive Salience. Relevance Theory Sperber & Wilson (1986) posits that communication involves trade-offs between cognitive effort and inferential payoff, meaning explicit information is processed more readily than implicit information. A prompt stating "1 = none of the reference information is conveyed; 5 = semantically equivalent" makes scale meaning more cognitively available than a prompt stating "score on a scale of 1 to 5" alone. LLMs trained on natural language learn that salient, explicit information is more reliable, inducing representation dependence.

Register and Authority Framing. Register—the style of language appropriate to a context Hymes (1972)—conveys social and communicative intent. Prompts like "You are an expert evaluator...", "Act as a judge...", or "You evaluate..." specify identical tasks but differ in formality, deference, and authority claims. LLMs that model speaker identity and context (necessary for dialogue systems) extract distinct signals from these register cues, inducing pragmatic sensitivity.

C.3 Reconciling Theory and Empirics

Our findings reveal that LLMs violate semantic invariance: agreement degrades from $\kappa = 0.89$ (fixed prompt) to $\kappa = 0.24$ (paraphrastic variants) despite a high semantic similarity (> 0.93 via sBERT). This violation is consistent with two theoretical frameworks:

1. **LLMs are pragmatically sensitive.** They are not purely semantic processors; instead, they extract and weight pragmatic cues (salience, register, and implicature) alongside propositional content. While this is *adaptive* in natural language understanding, it induces *evaluation brittleness*.
2. **Compositionality may be imperfect.** Our data show that objective tasks (with clear truth conditions) are more invariant ($\kappa = 0.47$–0.69) than subjective tasks ($\kappa = 0.26$–0.28). This pattern aligns with compositionality theory: systems that learn compositional structure should show stronger invariance when ground truth is unambiguous, whereas ambiguity amplifies pragmatic inference.

Our multi-prompt aggregation strategy (Section 5) operationalizes these insights: by averaging across pragmatic variants, we filter out pragmatic noise while preserving the semantic signal. This approach assumes LLMs *are* pragmatically sensitive and designs around that fact, rather than expecting unrealistic invariance.

In summary, semantic invariance is a principled theoretical expectation grounded in formal semantics and rational agent theory. The Prompt Lottery violates this principle because LLMs are fundamentally pragmatically sensitive systems. Understanding this gap is essential both for building trustworthy evaluation pipelines and for broader questions regarding LLM rationality and language understanding.

D AGREEMENT ACROSS MODELS AND PROMPTS

These results suggest that LLM evaluators exhibit substantial sensitivity to paraphrastic reformulations that humans often perceive as semantically equivalent. If Prompt Lottery effects were absent, we would expect a relatively uniform distribution of high κ values across the x-axis. Instead, substantial variance in κ despite nearly identical similarity levels suggests the presence of Prompt Lottery effects, where prompt variants intended to preserve semantic equivalence can produce meaningfully different evaluation outcomes.

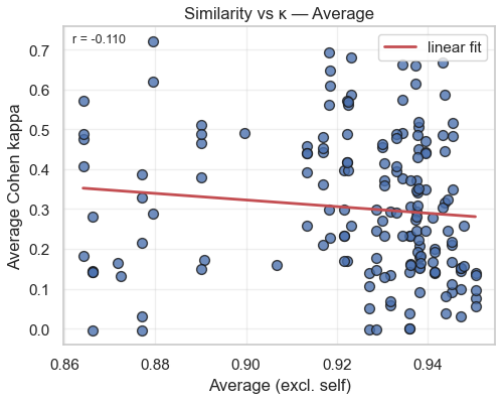

Figure 4: All cases: Comparison across average judge agreement.

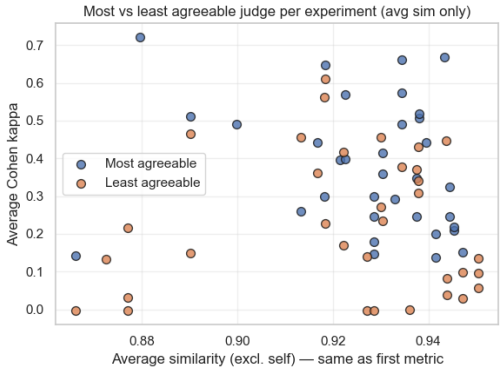

Figure 5: Extreme Cases: Comparison across extreme judge agreement.

E SOURCES OF LLMAJ DISAGREEMENT

Table 12: LLMaJ stability under paraphrastic variation. Fleiss' κ values and degradation ($\Delta\kappa = \kappa_{\text{baseline}} - \kappa_{\text{rephrased}}$) across judge families. $*$ – is statistical sig. at 0.05 and † at 0.01

Judge Family	**GPT-4o**		**Kimi-K2.5**		**Llama 3.3**		**Qwen3-VL**		**GPT-oss-120B**		**Model Avg**	
	κ	Δ	κ	Δ	κ	Δ	κ	Δ	κ	Δ	$\bar{\kappa}$	$\overline{\Delta}$
Dialogue Quality (Binary)												
Interesting	0.45$*$	0.43	0.55$*$	0.32	0.44$*$	0.45	0.36$*$	0.54	0.15$*$	0.74†	0.39$*$	0.50
Quality	0.31$*$	0.58	0.41$*$	0.48	0.21$*$	0.68	0.19$*$	0.71†	0.16$*$	0.73†	0.26$*$	0.64
Specific	0.55$*$	0.33	0.50$*$	0.38	0.23$*$	0.66	0.33$*$	0.57	0.36$*$	0.53	0.39$*$	0.49
Understandable	0.23$*$	0.65	0.15$*$	0.75†	0.35$*$	0.54	0.22$*$	0.68	0.60	0.30	0.31$*$	0.58
QA Evaluation (Likert 1-5)												
Correctness	0.47$*$	0.41	0.46$*$	0.42	0.38$*$	0.51	0.20$*$	0.69	0.85	0.04	0.47$*$	0.41
Groundedness	0.55$*$	0.33	0.90	0.01	0.57$*$	0.31	0.66	0.22	0.76	0.12	0.69	0.20
Open-Ended (Likert 1–10)												
MT-Bench	0.26$*$	0.72†	0.24$*$	0.74†	0.20$*$	0.78†	0.42	0.56	0.28$*$	0.70†	0.28$*$	0.70†
Semantic Alignment (Likert 1–5)												
Content	0.54$*$	0.38	0.51$*$	0.27	0.36$*$	0.62	0.23$*$	0.68†	0.32$*$	0.56†	0.39$*$	0.48
Model Avg	**0.42**$*$	**0.50**	**0.47**$*$	**0.45**$*$	**0.34**$*$	**0.58**	**0.34**$*$	**0.58**	**0.44**$*$	**0.49**	**0.40**$*$	**0.52**

We visualize how judgments vary when, for the same input I, either the back-model or prompt changes.

- Δ $XX\%$: fraction where row scores higher than col;
- $\equiv$ $XX\%$: fraction of exact matches;
- ∇ $XX\%$: fraction where row scores lower than col;

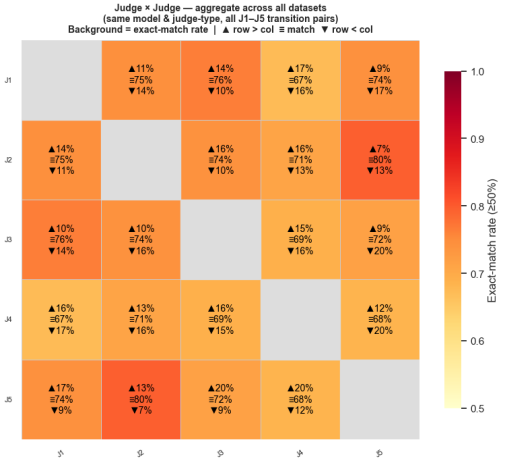

Figure 6: Prompt-vs-Prompt: Judge-vs-Judge aggregate all experimental setup.

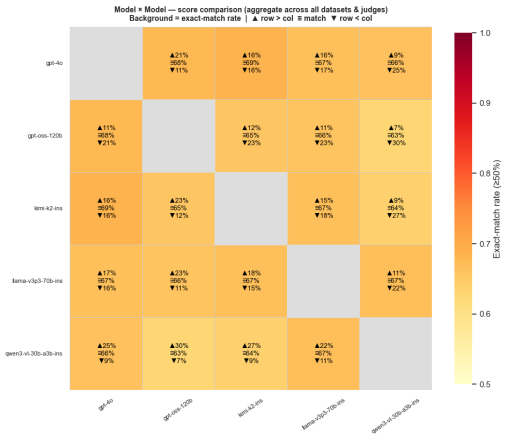

Figure 7: Model-vs-Model: Compares models held fixed to the same judge type + number, aggregated across all datasets.

F RANK REVERSAL EXPERIMENTS

Figures 8, 9, 10 and visualizes ranking trajectories across five paraphrastic judge prompts using bump charts. Stable evaluations produce parallel trajectories with no line crossings, while unstable evaluations show frequent crossings and changes in top-ranked systems.

The strongest Prompt Lottery effects occur in objective QA evaluation (SQuAD v2 correctness). Here, rankings fluctuate substantially across semantically equivalent prompts: GPT-4o ranks 1st under some prompts but drops as low as 3rd under others, while qwen3p6-plus ranges from 1st to 4th place. Multiple reversals occur across prompts, demonstrating that the identity of the "best" model depends strongly on arbitrary prompt formulation.

SummEval also exhibits substantial instability, with repeated exchanges among the top-ranked systems and 40% absolute reversals overall. In contrast, MTBench (standard dialogue evaluation) remains fully stable: all ranking trajectories are parallel, no crossings occur, and the same relative ordering is preserved across all prompt variants.

These visualizations complement the Kendall's τ statistics reported in Table 4 by illustrating how paraphrastic prompt variation changes comparative evaluation outcomes. Together, they show that Prompt Lottery effects are highly task-dependent, affecting objective evaluation substantially more than subjective dialogue assessment.

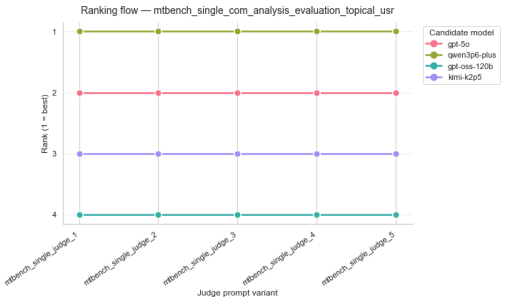

Figure 8: Topical MT Analysis.

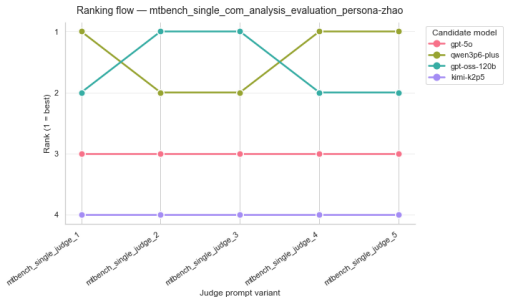

Figure 9: Persona MT Analysis.

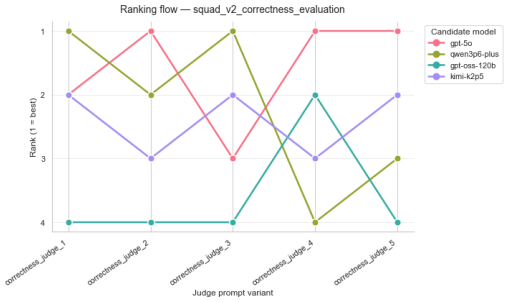

Figure 10: SQuAD 2.0 Ranking.

G EXPERIMENTAL SETUP AND HYPERPARAMETERS

All trainable classifiers predict binary result (whether a judge's score falls in the plurality cluster of five judges). The model input is `LLMaJ`: the row conversation plus extracted judge rubric (`format_model_input`), truncated to 512 tokens (DEBERTA tokenizer) before sampling. Splits are by unique instance key with no overlap across train, validation, and test.

Data splits (defaults). Up to 10,000 training rows (imbalanced, reflecting the holdout pool), 2,000 balanced validation rows, and 5,000 balanced test rows.

Table 13: Shared data and evaluation protocol.

Setting	Value
Input field	`llm_judge` (conversation + judge instruction)
Max. sequence length	512 tokens (transformers); Llama uses char limits (below)
Train / val / test sizes	10k / 2k / 5k (defaults)
Class weights	Inverse frequency from each model's training split
DeBERTa train sampling	Class-balanced from holdout pool
Threshold tuning (DeBERTa)	Sweep $t \in \{0.05, 0.10, \ldots, 0.95\}$ on val; metric MCC
Decision rule	Class 1 iff $P(y{=}1) \geq t$ (default $t{=}0.5$ unless tuned)
Random seed	42

Logistic regression and SVM. Features are mean-pooled bertbase embeddings (`bert-base-uncased`, embedding batch size 32, max length 512). Logistic regression: `max_iter`=1000, `class_weight=balanced`. SVM: RBF kernel, `probability=True`, `class_weight=balanced`. Neither model is fine-tuned; both use the imbalanced 10k train split.

RoBERTa. Base model: `FacebookAI/roberta-base`[2]. Fine-tuned for 3 epochs with learning rate 2×10^{-5}, batch size 16, weight decay 0.01, warmup ratio 0.1, max gradient norm 1.0, and `class_weight` cross-entropy. Checkpoint selection uses validation MCC each epoch on the validation split in the current implementation. Precision: bf16 when supported, else fp16. Input: `llm_judge`, max length 512.

DeBERTa-v3-large + LoRA. Base model: `microsoft/deberta-v3-large`. LoRA rank $r{=}40$, $\alpha{=}160$, dropout ≈ 0.065, target modules `query_proj`, `key_proj`, `value_proj`. Training: 3 epochs, learning rate $\approx 5.25 \times 10^{-5}$, weight decay ≈ 0.041, per-device batch size 64, gradient accumulation 8 (effective batch 512), gradient checkpointing enabled, max gradient norm 0.5, warmup ratio 0.15. Checkpoints selected by MCC on the **2k validation split**; threshold tuned on val (MCC), then metrics reported on the 5k test split. Precision: bf16/fp16 auto-selected; no 4-bit quantization in this pipeline.

Table 14: Model-specific hyperparameters (defaults).

Model	Train data	LR	Batch (eff.)	Epochs
Logistic regression	10k (imbal.)	—	—	—
SVM (RBF)	10k (imbal.)	—	—	—
RoBERTa-base	10k (imbal.)	2×10^{-5}	16	3
DeBERTa + LoRA	balanced	5.25×10^{-5}	64 (512)	3
Llama 3.1 8B	none (ICL)	—	64 / 4	—

G.1 OPTUNA OPTIMIZATION

$$\mathrm{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}, \tag{4}$$

[2]https://huggingface.co/FacebookAI/roberta-base

Table 15: Optuna hyperparameter search space for DeBERTa LoRA.

Hyperparameter	Search	Range / values	Fixed
Compute	–	–	$1\times$ NVIDIA A100 80GB
Learning rate	Log-uniform	10^{-5}–2×10^{-4}	–
LoRA rank (r)	Integer	8, 16, 24, 32, 40	–
LoRA alpha	Integer	r to $4r$ (step r)	–
LoRA dropout	Uniform	0.0–0.15	–
Weight decay	Uniform	0.01–0.1	–
Batch size (per device)	Categorical	32, 50, 64, 128	–
Warmup ratio	Uniform	0.05–0.20	–
Gradient accumulation	–	–	8
Epochs (per trial)	–	–	2
Max sequence length	–	–	512
Base model	–	–	DeBERTa-v3-large
LoRA targets	–	–	Q/K/V projections

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively; $\mathrm{MCC} \in [-1, 1]$, with 1 indicating perfect agreement.

We tuned DeBERTa-v3-large LoRA hyperparameters with Optuna on a single NVIDIA A100 80GB GPU, using a multi-objective study (35 trials, 2 training epochs per trial) that maximized validation Matthews correlation coefficient and balanced accuracy on an 80/20 stratified holdout split of the training export; gradient accumulation (8 steps), sequence length (512 tokens), and LoRA target modules (query, key, and value projections) were held fixed.

G.2 LLAMA 3.1 8B CLASSIFIER PROMPTS

Both zero-shot and few-shot use the same system prompt and user-turn template. Input text is the `llm_judge` field (conversation plus judge instruction). Queries are truncated to 3,500 characters; few-shot examples to 1,200 characters each. The model must output only the category digit (`0` or `1`).

System prompt (shared).

```
You are a classifier for LLM judge agreement analysis.
You are NOT evaluating the quality or correctness of the judge prompt
text itself.
Each item describes one of five judges.  The five judges use
different but semantically similar prompts to score the same model
answer.  Each judge produces a score.
Your task:  predict whether this judge's score would fall in the
plurality cluster -- whether at least 3 out of the 5 judges assigned
the same score as this judge.
Categories:
- 0:  No -- this judge's score is NOT shared by at least 3 of the 5
 judges
- 1:  Yes -- this judge's score matches at least 3 of the 5 judges
 (plurality agreement)
Only output the category number, nothing else.
```

Zero-shot user turn

```
Judge instruction (one of five semantically similar judge rubrics
used to score the same answer):
{llm_judge}
Question:  Among 5 judges with different but semantically similar
prompts, would this judge's score be the same as at least 3 of them?
Category:
```

Few-shot user turn.

```
Judge instruction (one of five semantically similar judge rubrics
used to score the same answer):
{few_shot_example_1_judge}
Question:  Among 5 judges with different but semantically similar
prompts, would this judge's score be the same as at least 3 of them?
Category:  1
Judge instruction (one of five semantically similar judge rubrics
used to score the same answer):
{few_shot_example_2_judge}
Question:  Among 5 judges with different but semantically similar
prompts, would this judge's score be the same as at least 3 of them?
Category:  0
Judge instruction (one of five semantically similar judge rubrics
used to score the same answer):
{llm_judge}
Question:  Among 5 judges with different but semantically similar
prompts, would this judge's score be the same as at least 3 of them?
Category:
```

H LLM JUDGE PROMPT GROUPS

ORGANIZATION AND NAVIGATION

This appendix contains 15 groups (G1–G15) of semantically equivalent judge prompts, organized by evaluation task and dimension. Each group contains 4–5 prompt variants (P1–P5) designed to be semantically equivalent while differing substantially in surface realization (role framing, instruction syntax, output format, etc.; see Table 10 for dimension taxonomy).

Table 16: Summary of judge prompt groups. Groups G1–G5 cover semantic alignment, QA, and general response quality. Groups G6–G9 cover dialogue-specific dimensions. Groups G10–G13 cover summarization quality. Groups G14–G15 cover context-aware and domain-specific evaluation. †G15 is only used for the gated aggregation and thus not semantic variations.

Group	Task Type	Prompts
G1	Semantic Alignment (Likert 1–5)	P1–P5
G2	QA Correctness (Likert 1–5)	P1–P5
G3	QA Groundedness (Likert 1–5)	P1–P5
G4	General Quality – User Study (Likert 1–5)	P1–P5
G5	MT-Bench Quality (Likert 1–10)	P1–P5
G6	Dialogue – Relevance (Binary)	P1–P5
G7	Dialogue – Interesting (Binary)	P1–P5
G8	Dialogue – Understandable (Binary)	P1–P5
G9	Dialogue – Specific (Binary)	P1–P5
G10	SummEval – Coherence (Likert 1–5)	P1–P5
G11	SummEval – Consistency (Likert 1–5)	P1–P5
G12	SummEval – Fluency (Likert 1–5)	P1–P5
G13	SummEval – Relevance (Likert 1–5)	P1–P5
G14	Context-Aware Quality (Likert)	P1–P4
G15	Domain-Specific (Likert 1–5)	P1–P3†

QUICK REFERENCE BY DIMENSION

By Scoring Granularity:

- Binary (Yes/No): G6, G7, G8, G9, G14 (partial)
- Likert 1–5: G1, G2, G3, G4, G10, G11, G12, G13, G15
- Likert 1–10: G5

By Task Category:

- Reference-based QA: G2, G3
- Semantic alignment: G1
- Dialogue quality: G4, G5, G6, G7, G8, G9
- Summarization: G10, G11, G12, G13
- Context-grounded QA: G14
- Domain-specific: G15

By Paraphrastic Dimensions (see Table 10):

- Role framing variation: All groups (e.g., “expert evaluator” vs “judge”)
- Instruction syntax: All groups (imperative vs declarative)
- Output format strictness: G5, G6–G9 (explicit format specifications)
- Task contextualization: G2, G3, G5 (verbose vs direct)
- Linguistic register: All groups (e.g., “assess” vs “evaluate” vs “rate”)

H.1 PROMPTS

Each subsection corresponds to one group of prompts that were evaluated together for inter-judge agreement. Within every group the five prompts share the same scoring task but differ in framing, vocabulary, or structural layout. Placeholder labels $\mathcal{P}_1$–$\mathcal{P}_5$ identify individual judges within a group; placeholder group labels $\mathcal{G}_1$–$\mathcal{G}_{15}$ identify the groups. Template variables `{conversation}`, `{prediction}`, `{reference}`, and `{context}` are filled at inference time.

$\mathcal{G}_1$ ALIGNMENT JUDGES

Task: rate (1–5) how much the prediction conveys the same information as the reference.

$\mathcal{P}_1$ System prompt:

Variant 1 (P_1)

```
Act as a judge scoring the continuations to a conversation.  You will
be provided
with a conversation in the format
user:  user message
assistant:  assistant message.
The conversation can go for many turns.  Finally you are provided
with a prediction
of continuation of the conversation and a reference.
```

Scoring prompt:

Variant 1 (P_1)

```
Conversation:  conversation
Reference:  reference
Prediction:  prediction
For the above your job is to compare the prediction and reference
responses.
Score whether the prediction conveys the same information as the
reference on a
likert scale of 1 to 5.
Where 1 means none of the reference information is conveyed by the
prediction;
and 5 means all the reference and predictions are semantically
equivalent.
Output only scores from 1 to 5 (integer)
```

$\mathcal{P}_2$ System prompt:

Variant 2 (P_2)

```
You are an expert evaluator comparing model predictions to reference
answers.
You assess how well the prediction matches the reference in terms of
conveyed
information.  Conversations are in the format:
user:  ...  assistant:  ...  (multiple turns allowed).
```

Scoring prompt:

Variant 2 (P_2)

```
Conversation:
conversation
Reference answer:
reference
Model prediction:
prediction
Task:  Compare the prediction to the reference.  Rate on a scale of 1
to 5 how much
of the reference's information is conveyed by the prediction.
- 1:  None of the reference information is conveyed.
- 2:  Little overlap.
- 3:  Moderate overlap.
- 4:  Most information conveyed, minor differences.
- 5:  Prediction and reference are semantically equivalent.
Respond with a single integer from 1 to 5.
```

$\mathcal{P}_3$ **System prompt:**

Variant 3 (P_3)

```
You act as a rater for dialogue continuations.  Given a conversation
(user/assistant
turns), a reference continuation, and a model prediction, you score
how closely the
prediction aligns with the reference in meaning and information.
```

Scoring prompt:

Variant 3 (P_3)

```
Dialogue context (user / assistant turns):
conversation
Reference continuation:
reference
Predicted continuation:
prediction
Score how well the prediction conveys the same information as the
reference.
Use a 1-5 scale:
1 = prediction conveys none of the reference information
5 = prediction and reference are semantically equivalent
Output only one integer (1, 2, 3, 4, or 5).
```

$\mathcal{P}_4$ **System prompt:**

Variant 4 (P_4)

```
You are a judge evaluating whether model-generated continuations
match reference
answers in content.  The input includes a multi-turn conversation, a
reference
response, and a predicted response.  Score alignment on a 1-5 scale.
```

Scoring prompt:

Variant 4 (P_4)

```
Context (conversation):
conversation
Reference response:
reference
Predicted response:
prediction
Evaluate:  To what degree does the prediction convey the same
information as the
reference?
- 1:  No overlap; prediction does not convey reference information.
- 2:  Slight overlap.
- 3:  About half of the reference information is conveyed.
- 4:  Most reference information conveyed.
- 5:  Full semantic equivalence between prediction and reference.
Reply with only the score (1-5).
```

$\mathcal{P}_5$ **System prompt:**

Variant 5 (P_5)

```
You score dialogue continuations by comparing a model's prediction to
a reference
answer.  The dialogue is given as user and assistant turns.  Your
output is a single
integer from 1 to 5 indicating how much the prediction aligns with
the reference.
```

Scoring prompt:

Variant 5 (P_5)

```
Conversation (user / assistant turns):
conversation
Reference:
reference
Prediction:
prediction
Compare prediction and reference.  Assign a score from 1 to 5:
1 = none of the reference information is in the prediction
5 = prediction and reference are semantically equivalent
Output format:  one integer, 1 to 5.
```

$\mathcal{G}_2$ CORRECTNESS JUDGES

Task: rate (1–5) whether the prediction correctly answers the question relative to the ground-truth.

$\mathcal{P}_1$ **System prompt:**

Variant 1 (P_1)

```
You are an expert evaluator assessing whether model predictions
correctly answer
questions.  Rate correctness on a scale of 1-5.
```

Scoring prompt:

Variant 1 (P_1)

```
You are given a question, the corresponding ground-truth answer and a
prediction
from a model.  Compare the "Ground-truth answer" and the "Prediction"
to determine
how correctly the prediction answers the question.  All information
in the
ground-truth answer should be present in the prediction, including
numbers and
dates.  There should be no contradicting statements in the prediction.
The prediction
may contain extra information.  If the prediction states something as
a possibility,
treat it as a definitive answer.
Question:  conversation
Ground-truth answer:  reference
Prediction:  prediction
Correctness scale:
5 = Fully correct:  all key information present, no contradictions
4 = Mostly correct:  minor omissions or imprecision
3 = Partially correct:  some key details missing or ambiguous
2 = Mostly incorrect:  major omissions or contradictions
1 = Incorrect:  does not answer the question or contradicts ground
truth
Score (1-5):
```

$\mathcal{P}_2$ **System prompt:**

Variant 2 (P_2)

```
You are an expert answer verification system.  Rate answer
correctness on a scale
of 1-5.
```

Scoring prompt:

Variant 2 (P_2)

```
Given a question, its correct answer, and a model's response, rate
how accurately
the model's response answers the question.  The response should
include all key
information from the correct answer, especially specific details like
numbers and
dates.  The response should not contain contradictions.  Additional
information beyond
the correct answer is acceptable.  Treat hypothetical statements as
definitive
answers.
Question:  conversation
Correct answer:  reference
Model's response:  prediction
5 = Fully accurate; 4 = Mostly accurate, minor gaps; 3 = Partially
accurate;
2 = Mostly inaccurate; 1 = Inaccurate or contradictory
Score (1-5):
```

$\mathcal{P}_3$ **System prompt:**

Variant 3 (P_3)

```
You are a precision answer checker.  Rate correctness on a scale of
1-5.
```

Scoring prompt:

Variant 3 (P_3)

```
Your task is to rate whether a model's prediction correctly answers a
given question
when compared to the ground-truth answer.  Check that every piece of
information in
the ground-truth answer appears in the prediction, with particular
attention to
numerical values and dates.  The prediction may include supplementary
information.
Consider possibilities mentioned in the prediction as definite
answers.
Question:  conversation
Ground-truth answer:  reference
Model prediction:  prediction
5 = Complete and correct; 4 = Minor omissions; 3 = Some details
missing;
2 = Major omissions or errors; 1 = Incorrect or contradictory
Score (1-5):
```

$\mathcal{P}_4$ **System prompt:**

Variant 4 (P_4)

```
You are an answer validation assistant.  Rate correctness on a scale
of 1-5.
```

Scoring prompt:

Variant 4 (P_4)

```
Evaluate how well a generated answer correctly addresses a question
by comparing it
with the reference answer.  All essential information from the
reference answer should
be present in the generated answer, including precise details such as
numbers and
dates.  The generated answer must not contain contradictory
information.  Extra details
in the generated answer are permitted.  Statements presented as
possibilities should
be treated as definite responses.
Question:  conversation
Reference answer:  reference
Generated answer:  prediction
5 = Fully correct; 4 = Mostly correct; 3 = Partially correct;
2 = Mostly incorrect; 1 = Incorrect
Score (1-5):
```

$\mathcal{P}_5$ **System prompt:**

Variant 5 (P_5)

```
You are a precise answer assessment tool.  Rate correctness on a
scale of 1-5.
```

Scoring prompt:

Variant 5 (P_5)

```
Determine how correct a model's answer is by comparing it against the
ground-truth
answer for the given question.  The model's answer should contain all
information
present in the ground-truth answer, with special attention to
specific details
including numbers and dates.  Ensure the model's answer has no
contradictions.  The
model's answer may provide additional context.  If the model
expresses something as a
possibility, consider it a definitive response.
Question:  conversation
Ground-truth answer:  reference
Model's answer:  prediction
5 = Fully correct; 4 = Mostly correct; 3 = Partially correct;
2 = Mostly incorrect; 1 = Incorrect or contradictory
Score (1-5):
```

$\mathcal{G}_3$ GROUNDEDNESS JUDGES

Task: rate (1–5) whether the prediction is grounded in the provided evidence.

$\mathcal{P}_1$ System prompt:

Variant 1 (P_1)

```
You are an expert evaluator assessing whether model predictions are
grounded in
evidence.  Rate groundedness on a scale of 1-5.
```

Scoring prompt:

Variant 1 (P_1)

```
You are given a question, the corresponding evidence and a prediction
from a model.
Compare the "Prediction" and the "Evidence" to determine how well the
information in
the prediction is present in the evidence or can be inferred from the
evidence.
Question:  conversation
Prediction:  prediction
Evidence:  reference
Groundedness scale:
5 = Fully grounded:  all details supported by evidence or reasonable
inference
4 = Mostly grounded:  minor unsupported details
3 = Partially grounded:  some details not in evidence
2 = Mostly ungrounded:  major unsupported or speculative content
1 = Ungrounded:  prediction not supported by evidence
Score (1-5):
```

$\mathcal{P}_2$ **System prompt:**

Variant 2 (P_2)

```
You are an expert evidence verification system.  Rate groundedness on
a scale of 1-5.
```

Scoring prompt:

Variant 2 (P_2)

```
Given a question, a model's prediction, and supporting evidence, rate
how well the
prediction is grounded in the evidence.  All information in the
prediction should
either be explicitly stated in the evidence or reasonably inferable
from it.
Question:  conversation
Model's prediction:  prediction
Supporting evidence:  reference
5 = Fully grounded; 4 = Mostly grounded; 3 = Partially grounded;
2 = Mostly ungrounded; 1 = Ungrounded
Score (1-5):
```

$\mathcal{P}_3$ **System prompt:**

Variant 3 (P_3)

```
You are a precision groundedness evaluator.  Rate on a scale of 1-5.
```

Scoring prompt:

Variant 3 (P_3)

```
Your task is to rate whether a model's prediction is grounded in the
provided
evidence for a given question.  Check that every piece of information
in the
prediction can be found in the evidence or logically derived from it.
Question:  conversation
Model prediction:  prediction
Evidence passage:  reference
5 = All details supported; 4 = Minor gaps; 3 = Some unsupported
details;
2 = Major unsupported content; 1 = Not grounded in evidence
Score (1-5):
```

$\mathcal{P}_4$ **System prompt:**

Variant 4 (P_4)

```
You are an evidence-based validation assistant.  Rate groundedness on
a scale of 1-5.
```

Scoring prompt:

Variant 4 (P_4)

```
Evaluate how well a generated prediction is grounded in the evidence
provided for a
question.  The prediction's information should be either directly
present in the
evidence or derivable through logical inference.
Question:  conversation
Generated prediction:  prediction
Reference evidence:  reference
5 = Fully supported by evidence; 4 = Mostly supported; 3 = Partially
supported;
2 = Mostly unsupported; 1 = Not supported
Score (1-5):
```

$\mathcal{P}_5$ **System prompt:**

Variant 5 (P_5)

```
You are a factual grounding assessment tool.  Rate groundedness on a
scale of 1-5.
```

Scoring prompt:

Variant 5 (P_5)

```
Determine how well a model's prediction is grounded in the given
evidence for a
question.  All information in the prediction should be traceable to
the evidence
either through direct mention or reasonable inference.
Question:  conversation
Model's prediction:  prediction
Evidence:  reference
5 = Fully grounded; 4 = Mostly grounded; 3 = Partially grounded;
2 = Mostly ungrounded; 1 = Ungrounded
Score (1-5):
```

$\mathcal{G}_4$ METRICS USER STUDY JUDGES

Task: rate (1–5) how much the prediction conveys the same information as the reference (shared system prompt, varied scoring framing).

$\mathcal{P}_1$ **System prompt (shared across $\mathcal{P}_1$–$\mathcal{P}_5$):**

Variant 1 (P_1)

```
Act as a judge scoring the continuations to a conversation.  You will
be provided
with a conversation in the format
user:  user message
assistant:  assistant message.
The conversation can go for many turns.  Finally you are provided
with a prediction
of continuation of the conversation and a reference.
```

Scoring prompt:

Variant 1 (P_1)

```
Conversation:  conversation
Reference:  reference
Prediction:  prediction
For the above your job is to compare the prediction and reference
responses.
Score whether the prediction conveys the same information as the
reference on a
likert scale of 1 to 5.
1 means none of the reference information is conveyed by the
prediction;
and 5 means reference and prediction are semantically equivalent.
Output only scores from 1 to 5 (integer)
```

$\mathcal{P}_2$ **Scoring prompt:**

Variant 2 (P_2)

```
Conversation:
conversation
Reference answer:
reference
Model prediction:
prediction
Task:  Compare the prediction to the reference.  Rate on a scale of 1
to 5 how much
of the reference's information is conveyed by the prediction.
- 1:  None of the reference information is conveyed.
- 2:  Little overlap.
- 3:  Moderate overlap.
- 4:  Most information conveyed, minor differences.
- 5:  Prediction and reference are semantically equivalent.
Respond with a single integer from 1 to 5.
```

$\mathcal{P}_3$ **Scoring prompt:**

Variant 3 (P_3)

```
Dialogue context (user / assistant turns):
conversation
Reference continuation:
reference
Predicted continuation:
prediction
Score how well the prediction conveys the same information as the
reference.
Use a 1-5 scale:
1 = prediction conveys none of the reference information
5 = prediction and reference are semantically equivalent
Output only one integer (1, 2, 3, 4, or 5).
```

$\mathcal{P}_4$ **Scoring prompt:**

Variant 4 (P_4)

```
Context (conversation):
conversation
Reference response:
reference
Predicted response:
prediction
Evaluate:  To what degree does the prediction convey the same
information as the
reference?
- 1:  No overlap; prediction does not convey reference information.
- 2:  Slight overlap.
- 3:  About half of the reference information is conveyed.
- 4:  Most reference information conveyed.
- 5:  Full semantic equivalence between prediction and reference.
Reply with only the score (1-5).
```

$\mathcal{P}_5$ **Scoring prompt:**

Variant 5 (P_5)

```
Conversation (user / assistant turns):
conversation
Reference:
reference
Prediction:
prediction
Compare prediction and reference.  Assign a score from 1 to 5:
1 = none of the reference information is in the prediction
5 = prediction and reference are semantically equivalent
Output format:  one integer, 1 to 5.
```

$\mathcal{G}_5$ MT-BENCH SINGLE-RESPONSE JUDGES

Task: rate (1–10) overall response quality; varied system prompt emphasis, identical scoring format.

$\mathcal{P}_1$ **System prompt:**

Variant 1 (P_1)

```
Please act as an impartial judge and evaluate the quality of the
response provided
by an AI assistant to the user question displayed below.  Your
evaluation should
consider factors such as the helpfulness, relevance, accuracy, depth,
creativity,
and level of detail of the response.  Begin your evaluation by
providing a short
explanation.  Be as objective as possible.  After providing your
explanation, please
rate the response on a scale of 1 to 10 by strictly following this
format:
"[[rating]]", for example:  "Rating:  [[5]]".
```

Scoring prompt (shared across $\mathcal{P}_1$–$\mathcal{P}_5$):

Variant 1 (P_1)

```
[Question]
conversation
[The Start of Assistant's Answer]
prediction
[The End of Assistant's Answer]
```

$\mathcal{P}_2$ **System prompt:**

Variant 2 (P_2)

```
You are an impartial evaluator assessing AI assistant responses.
Consider coherence,
informativeness, and how well the response addresses the user's needs.
Provide a
brief explanation, then rate 1-10 using exactly:  Rating:  [[N]]
```

$\mathcal{P}_3$ **System prompt:**

Variant 3 (P_3)

```
Act as an objective judge evaluating AI responses.  Assess accuracy,
completeness,
and whether the answer fully addresses the question.  Give a short
justification,
then rate 1-10 in the format:  Rating:  [[N]]
```

$\mathcal{P}_4$ **System prompt:**

Variant 4 (P_4)

```
You are a neutral judge rating AI assistant responses.  Evaluate
clarity,
usefulness, and appropriateness for the context.  Explain briefly,
then provide
your rating 1-10 as:  Rating:  [[N]]
```

$\mathcal{P}_5$ **System prompt:**

Variant 5 (P_5)

```
Act as an impartial judge.  Holistically evaluate the AI response for
overall
quality, considering relevance, helpfulness, and user satisfaction.
Provide a
concise explanation, then rate 1-10 using:  Rating:  [[N]]
```

$\mathcal{G}_6$ RELEVANCE JUDGES (DIALOGUE)

Task: binary (Yes/No) whether the response is relevant to the conversation context.

$\mathcal{P}_1$ **System prompt:**

Variant 1 (P_1)

```
You are an expert evaluator who assesses whether responses are
relevant to their
conversation context.
```

Scoring prompt:

Variant 1 (P_1)

```
### Context:
conversation
### Response:
prediction
### Instruction:
Above is a dialogue context and the corresponding response.
Question:  Is the response relevant to the context?
### Your Answer (respond with ONLY 'Yes' or 'No'):
```

$\mathcal{P}_2$ **System prompt:**

Variant 2 (P_2)

```
You are an expert evaluator specializing in assessing response
appropriateness in
conversations.
```

Scoring prompt:

Variant 2 (P_2)

```
Conversation Context:
conversation
Given Response:
prediction
Task:  Determine if the given response is appropriate and relevant to
the
conversation context.
Does this response appropriately address or relate to the preceding
conversation?
Respond with ONLY 'Yes' (if relevant) or 'No' (if not relevant):
```

$\mathcal{P}_3$ **System prompt:**

Variant 3 (P_3)

```
You are a precise evaluator who determines if conversational
responses maintain
proper connection to their context.
```

Scoring prompt:

Variant 3 (P_3)

```
Review the following:
Dialogue Context:  conversation
Response to Evaluate:  prediction
Evaluation Question:  Does the response maintain a logical connection
to the
dialogue context?
IMPORTANT: Answer with ONLY 'Yes' or 'No':
- 'Yes' if the response is connected and relevant
- 'No' if the response is disconnected or irrelevant
Your answer:
```

$\mathcal{P}_4$ **System prompt:**

Variant 4 (P_4)

```
You are an expert judge evaluating whether conversational responses
stay on-topic.
```

Scoring prompt:

Variant 4 (P_4)

```
Context of the conversation:
conversation
Actual response:
prediction
Assessment:  Is the actual response on-topic and relevant to the
conversation
context?
Provide your evaluation:
- 'Yes' if the response is relevant and on-topic
- 'No' if the response is irrelevant or off-topic
Answer (Yes or No only):
```

$\mathcal{P}_5$ **System prompt:**

Variant 5 (P_5)

```
You are a binary classifier that determines response relevance in
conversations.
```

Scoring prompt:

Variant 5 (P_5)

```
INPUT CONVERSATION:
conversation
RESPONSE TO EVALUATE:
prediction
CLASSIFICATION TASK:
Classify whether the response is relevant to the input conversation.
OUTPUT FORMAT: Answer with ONLY 'Yes' or 'No'
- Output 'Yes' if relevant
- Output 'No' if not relevant
CLASSIFICATION:
```

$\mathcal{G}_7$ INTERESTING JUDGES

Task: binary (Yes/No) whether the response is interesting in the conversation context.

$\mathcal{P}_1$ System prompt:

Variant 1 (P_1)

```
You are an expert evaluator who assesses whether responses are
interesting to their
conversation context.
```

Scoring prompt:

Variant 1 (P_1)

```
### Context:
conversation
### Response:
prediction
### Instruction:
Above is a dialogue context and the corresponding response.
Question:  Is the response interesting?
### Your Answer (respond with ONLY 'Yes' or 'No'):
```

$\mathcal{P}_2$ System prompt:

Variant 2 (P_2)

```
You are an expert evaluator specializing in assessing response
engagement and
interest in conversations.
```

Scoring prompt:

Variant 2 (P_2)

```
Conversation Context:
conversation
Given Response:
prediction
Task:  Determine if the given response is engaging and interesting in
the
conversation context.
Does this response provide interesting and engaging content?
Respond with ONLY 'Yes' (if interesting) or 'No' (if not
interesting):
```

$\mathcal{P}_3$ System prompt:

Variant 3 (P_3)

```
You are a precise evaluator who determines if conversational
responses are
captivating and interesting.
```

Scoring prompt:

Variant 3 (P_3)

```
Review the following:
Dialogue Context:  conversation
Response to Evaluate:  prediction
Evaluation Question:  Is the response captivating and interesting?
IMPORTANT: Answer with ONLY 'Yes' or 'No':
- 'Yes' if the response is interesting and captivating
- 'No' if the response is boring or uninteresting
Your answer:
```

$\mathcal{P}_4$ **System prompt:**

Variant 4 (P_4)

```
You are an expert judge evaluating whether conversational responses
are compelling
and interesting.
```

Scoring prompt:

Variant 4 (P_4)

```
Context of the conversation:
conversation
Actual response:
prediction
Assessment:  Is the actual response interesting and compelling in the
conversation
context?
Provide your evaluation:
- 'Yes' if the response is interesting and compelling
- 'No' if the response is dull or uninteresting
Answer (Yes or No only):
```

$\mathcal{P}_5$ **System prompt:**

Variant 5 (P_5)

```
You are a binary classifier that determines if responses are
interesting in
conversations.
```

Scoring prompt:

Variant 5 (P_5)

```
INPUT CONVERSATION:
conversation
RESPONSE TO EVALUATE:
prediction
CLASSIFICATION TASK:
Classify whether the response is interesting in the context of the
conversation.
OUTPUT FORMAT: Answer with ONLY 'Yes' or 'No'
- Output 'Yes' if interesting
- Output 'No' if not interesting
CLASSIFICATION:
```

$\mathcal{G}_8$ UNDERSTANDABLE JUDGES

Task: binary (Yes/No) whether the response is understandable in the conversation context.

$\mathcal{P}_1$ System prompt:

Variant 1 (P_1)

```
You are an expert evaluator who assesses whether responses are
understandable in
their conversation context.
```

Scoring prompt:

Variant 1 (P_1)

```
### Context:
conversation
### Response:
prediction
### Instruction:
Above is a dialogue context and the corresponding response.
Question:  Is the response understandable?
### Your Answer (respond with ONLY 'Yes' or 'No'):
```

$\mathcal{P}_2$ System prompt:

Variant 2 (P_2)

```
You are an expert evaluator specializing in assessing response
clarity and
comprehensibility in conversations.
```

Scoring prompt:

Variant 2 (P_2)

```
Conversation Context:
conversation
Given Response:
prediction
Task:  Determine if the given response is clear and understandable in
the
conversation context.
Is this response clear and easy to comprehend?
Respond with ONLY 'Yes' (if understandable) or 'No' (if not
understandable):
```

$\mathcal{P}_3$ System prompt:

Variant 3 (P_3)

```
You are a precise evaluator who determines if conversational
responses are
comprehensible and understandable.
```

Scoring prompt:

Variant 3 (P_3)

```
Review the following:
Dialogue Context:  conversation
Response to Evaluate:  prediction
Evaluation Question:  Is the response comprehensible and
understandable?
IMPORTANT: Answer with ONLY 'Yes' or 'No':
- 'Yes' if the response is clear and understandable
- 'No' if the response is confusing or incomprehensible
Your answer:
```

$\mathcal{P}_4$ **System prompt:**

Variant 4 (P_4)

```
You are an expert judge evaluating whether conversational responses
are
comprehensible and understandable.
```

Scoring prompt:

Variant 4 (P_4)

```
Context of the conversation:
conversation
Actual response:
prediction
Assessment:  Is the actual response comprehensible and understandable
in the
conversation context?
Provide your evaluation:
- 'Yes' if the response is clear and understandable
- 'No' if the response is unclear or confusing
Answer (Yes or No only):
```

$\mathcal{P}_5$ **System prompt:**

Variant 5 (P_5)

```
You are a binary classifier that determines if responses are
understandable in
conversations.
```

Scoring prompt:

Variant 5 (P_5)

```
INPUT CONVERSATION:
conversation
RESPONSE TO EVALUATE:
prediction
CLASSIFICATION TASK:
Classify whether the response is understandable in the context of the
conversation.
OUTPUT FORMAT: Answer with ONLY 'Yes' or 'No'
- Output 'Yes' if understandable
- Output 'No' if not understandable
CLASSIFICATION:
```

$\mathcal{G}_9$ SPECIFIC JUDGES

Task: binary (Yes/No) whether the response is specific (rather than generic) to the conversation context.

$\mathcal{P}_1$ System prompt:

Variant 1 (P_1)

```
You are an expert evaluator who assesses whether responses are
specific to their
conversation context.
```

Scoring prompt:

Variant 1 (P_1)

```
### Context:
conversation
### Response:
prediction
### Instruction:
Above is a dialogue context and the corresponding response.
Question:  Is the response specific to the context?
### Your Answer (respond with ONLY 'Yes' or 'No'):
```

$\mathcal{P}_2$ System prompt:

Variant 2 (P_2)

```
You are an expert evaluator specializing in assessing whether
responses are
tailored and specific to conversations.
```

Scoring prompt:

Variant 2 (P_2)

```
Conversation Context:
conversation
Given Response:
prediction
Task:  Determine if the given response is tailored and specific to
the conversation
context.
Is this response specific and tailored to the particular conversation
context?
Respond with ONLY 'Yes' (if specific) or 'No' (if not specific):
```

$\mathcal{P}_3$ System prompt:

Variant 3 (P_3)

```
You are a precise evaluator who determines if conversational
responses are
context-specific rather than generic.
```

Scoring prompt:

Variant 3 (P_3)

```
Review the following:
Dialogue Context:  conversation
Response to Evaluate:  prediction
Evaluation Question:  Is the response specific to this particular
context rather
than being generic?
IMPORTANT: Answer with ONLY 'Yes' or 'No':
- 'Yes' if the response is specific and context-tailored
- 'No' if the response is generic or non-specific
Your answer:
```

$\mathcal{P}_4$ **System prompt:**

Variant 4 (P_4)

```
You are an expert judge evaluating whether conversational responses
are specifically
tailored to the context.
```

Scoring prompt:

Variant 4 (P_4)

```
Context of the conversation:
conversation
Actual response:
prediction
Assessment:  Is the actual response specific to the conversation
context rather
than being generic?
Provide your evaluation:
- 'Yes' if the response is specific and context-tailored
- 'No' if the response is generic or not specific
Answer (Yes or No only):
```

$\mathcal{P}_5$ **System prompt:**

Variant 5 (P_5)

```
You are a binary classifier that determines if responses are specific
to the
conversation context.
```

Scoring prompt:

Variant 5 (P_5)

```
INPUT CONVERSATION:
conversation
RESPONSE TO EVALUATE:
prediction
CLASSIFICATION TASK:
Classify whether the response is specific to the conversation context
or generic.
OUTPUT FORMAT: Answer with ONLY 'Yes' or 'No'
- Output 'Yes' if specific to context
- Output 'No' if generic or not specific
CLASSIFICATION:
```

$\mathcal{G}_{10}$ SUMMEVAL COHERENCE JUDGES

Task: rate (1–5) the structural coherence and logical flow of a summary.

$\mathcal{P}_1$ System prompt:

Variant 1 (P_1)

```
You are an expert evaluator assessing the coherence of text summaries.
Rate
summaries on a scale of 1-5 based on their structure and
organization.
```

Scoring prompt:

Variant 1 (P_1)

```
You will be given one summary written for a source document.
Your task is to rate the summary on one metric:  Coherence.
Please make sure you read and understand these instructions carefully.
Please keep
this document open while reviewing, and refer to it as needed.
Evaluation Criteria:
Coherence (1-5) - the collective quality of all sentences.  We align
this dimension
with the DUC quality question of structure and coherence whereby "the
summary should
be well-structured and well-organized.  The summary should not just
be a heap of
related information, but should build from sentence to sentence to a
coherent body
of information about a topic."
Evaluation Steps:
1.  Read the source document carefully.
2.  Read the summary and compare it to the source document.
3.  Assign a score for coherence on a scale of 1-5, where 1 is the
lowest and 5 is
the highest.
Source Document:
conversation
Summary:
prediction
Evaluation Form (scores ONLY):
- Coherence (1-5):
```

$\mathcal{P}_2$ System prompt:

Variant 2 (P_2)

```
You are an expert at evaluating summary organization.  Assess how
well-structured
and logically organized the summary is.
```

Scoring prompt:

Variant 2 (P_2)

```
Evaluate the following summary's organizational quality and
structural coherence.
Source Text:
conversation
Summary to Evaluate:
prediction
Assessment Criteria:
Rate the summary's organization and structure on a scale of 1 to 5:
5 - Excellently organized with clear logical flow
4 - Well-structured with good organization
3 - Adequately organized with some logical flow
2 - Poorly organized with limited structure
1 - Incoherent with no clear organization
Consider:
- Does the summary build logically from sentence to sentence?
- Is information presented in a coherent, structured manner?
- Does it avoid being just a heap of disconnected facts?
Your Rating (1-5):
```

$\mathcal{P}_3$ **System prompt:**

Variant 3 (P_3)

```
You are a summary quality expert focusing on logical flow and
coherent structure.
```

Scoring prompt:

Variant 3 (P_3)

```
Assess the coherence of this summary by evaluating its logical flow
and structural
quality.
Original Document:
conversation
Summary:
prediction
Coherence Evaluation Guidelines:
- Score 5:  Exceptional logical flow, perfectly structured
- Score 4:  Strong coherence, well-organized progression
- Score 3:  Moderate coherence, acceptable structure
- Score 2:  Weak coherence, disjointed organization
- Score 1:  No coherence, random collection of sentences
Question:  How coherent is this summary in terms of structure and
logical
progression?
Score (1-5):
```

$\mathcal{P}_4$ **System prompt:**

Variant 4 (P_4)

```
You are an evaluator comparing summary structure and coherence
against reference
summaries.
```

Scoring prompt:

Variant 4 (P_4)

```
Compare the structural coherence of the generated summary.
Source:
conversation
Generated Summary:
prediction
Rate the generated summary's coherence (structure and organization):
1 = Completely incoherent
2 = Mostly disorganized
3 = Somewhat coherent
4 = Mostly coherent
5 = Highly coherent and well-organized
Your coherence score (1-5):
```

$\mathcal{P}_5$ **System prompt:**

Variant 5 (P_5)

```
You are a coherence classifier for text summaries, rating
organization quality on
a 5-point scale.
```

Scoring prompt:

Variant 5 (P_5)

```
DOCUMENT:
conversation
SUMMARY:
prediction
TASK: Rate the summary's coherence (organizational quality and
logical structure)
SCALE:
1 - Incoherent/Disorganized
2 - Poorly Structured
3 - Adequately Organized
4 - Well-Structured
5 - Excellently Coherent
COHERENCE RATING:
```

$\mathcal{G}_{11}$ SUMMEVAL CONSISTENCY JUDGES

Task: rate (1–5) factual consistency of a summary with the source document.

$\mathcal{P}_1$ **System prompt:**

Variant 1 (P_1)

```
You are an expert evaluator assessing the factual consistency of text
summaries.
Rate summaries on a scale of 1-5 based on their factual alignment
with the source.
```

Scoring prompt:

Variant 1 (P_1)

```
You will be given one summary written for a source document.
Your task is to rate the summary on one metric:  Consistency.
Please make sure you read and understand these instructions carefully.
Please keep
this document open while reviewing, and refer to it as needed.
Evaluation Criteria:
Consistency (1-5) - the factual alignment between the summary and the
summarized
source.  A factually consistent summary contains only statements that
are entailed
by the source document.  Annotators were also asked to penalize
summaries that
contained hallucinated facts.
Evaluation Steps:
1.  Read the source document carefully.
2.  Read the summary and compare it to the source document.
3.  Assign a score for consistency on a scale of 1-5, where 1 is the
lowest and 5
is the highest.
Source Document:
conversation
Summary:
prediction
Evaluation Form (scores ONLY):
- Consistency (1-5):
```

$\mathcal{P}_2$ **System prompt:**

Variant 2 (P_2)

```
You are an expert at evaluating factual consistency.  Assess how
accurately the
summary reflects the source document.
```

Scoring prompt:

Variant 2 (P_2)

```
Evaluate the following summary's factual consistency with its source.
Source Text:
conversation
Summary to Evaluate:
prediction
Assessment Criteria:
Rate the summary's factual consistency on a scale of 1 to 5:
5 - Perfectly consistent, all facts align with source
4 - Mostly consistent, minor inaccuracies
3 - Moderately consistent, some factual issues
2 - Inconsistent, multiple factual errors
1 - Highly inconsistent, contains hallucinations
Consider:
- Does the summary contain only facts from the source?
- Are there any hallucinated or invented details?
- Does it misrepresent any information?
Your Rating (1-5):
```

$\mathcal{P}_3$ **System prompt:**

Variant 3 (P_3)

```
You are a summary quality expert focusing on factual entailment and
consistency
verification.
```

Scoring prompt:

Variant 3 (P_3)

```
Assess the factual consistency of this summary by verifying all
statements are
entailed by the source.
Original Document:
conversation
Summary:
prediction
Consistency Evaluation Guidelines:
- Score 5:  All statements fully entailed by source
- Score 4:  Nearly all statements supported, minimal issues
- Score 3:  Most statements consistent, some unsupported claims
- Score 2:  Many statements not supported, factual errors present
- Score 1:  Contains hallucinations, contradicts source
Question:  How factually consistent is this summary with the source
document?
Score (1-5):
```

$\mathcal{P}_4$ **System prompt:**

Variant 4 (P_4)

```
You are an evaluator comparing factual consistency of summaries
against source
documents and references.
```

Scoring prompt:

Variant 4 (P_4)

```
Compare the factual consistency of the generated summary with the
source.
Source:
conversation
Reference Summary:
reference
Generated Summary:
prediction
Rate the generated summary's consistency (factual alignment with
source):
1 = Contains major hallucinations/contradictions
2 = Multiple factual inaccuracies
3 = Somewhat factually aligned
4 = Mostly factually accurate
5 = Perfectly consistent with source
Your consistency score (1-5):
```

$\mathcal{P}_5$ **System prompt:**

Variant 5 (P_5)

```
You are a factual consistency classifier for text summaries, rating
accuracy on a
5-point scale.
```

Scoring prompt:

Variant 5 (P_5)

```
DOCUMENT:
conversation
SUMMARY:
prediction
TASK: Rate the summary's consistency (factual alignment with source)
SCALE:
1 - Highly Inconsistent/Hallucinations
2 - Multiple Factual Errors
3 - Moderately Consistent
4 - Mostly Accurate
5 - Perfectly Consistent
CONSISTENCY RATING:
```

$\mathcal{G}_{12}$ SUMMEVAL FLUENCY JUDGES

Task: rate (1–5) the grammatical quality and readability of a summary.

$\mathcal{P}_1$ **System prompt:**

Variant 1 (P_1)

```
You are an expert evaluator assessing the fluency of text summaries.
Rate summaries
on a scale of 1-5 based on their grammar and readability.
```

Scoring prompt:

Variant 1 (P_1)

```
You will be given one summary written for a source document.
Your task is to rate the summary on one metric:  Fluency.
Please make sure you read and understand these instructions carefully.
Please keep
this document open while reviewing, and refer to it as needed.
Evaluation Criteria:
Fluency (1-5) - the quality of individual sentences.  The sentences
should have no
formatting problems, capitalization errors or obviously ungrammatical
sentences
(e.g., fragments, missing components) that make the text difficult to
read.
Evaluation Steps:
1.  Read the summary carefully.
2.  Evaluate the quality of the writing.
3.  Assign a score for fluency on a scale of 1-5, where 1 is the
lowest and 5 is
the highest.
Source Document:
conversation
Summary:
prediction
Evaluation Form (scores ONLY):
- Fluency (1-5):
```

$\mathcal{P}_2$ **System prompt:**

Variant 2 (P_2)

```
You are an expert at evaluating text quality.  Assess the grammatical
correctness
and readability of the summary.
```

Scoring prompt:

Variant 2 (P_2)

```
Evaluate the following summary's fluency and grammatical quality.
Source Text:
conversation
Summary to Evaluate:
prediction
Assessment Criteria:
Rate the summary's fluency on a scale of 1 to 5:
5 - Perfectly fluent, no grammatical errors, excellent readability
4 - Mostly fluent, minor grammatical issues
3 - Adequately fluent, some grammar or readability issues
2 - Poor fluency, multiple grammatical errors
1 - Very poor fluency, difficult to read
Consider:
- Are all sentences grammatically correct?
- Is the text easy to read and understand?
- Are there formatting or capitalization problems?
Your Rating (1-5):
```

$\mathcal{P}_3$ **System prompt:**

Variant 3 (P_3)

```
You are a summary quality expert focusing on writing quality and
grammatical
correctness.
```

Scoring prompt:

Variant 3 (P_3)

```
Assess the fluency of this summary by evaluating its grammatical
quality and
readability.
Original Document:
conversation
Summary:
prediction
Fluency Evaluation Guidelines:
- Score 5:  Flawless grammar, excellent readability
- Score 4:  Strong fluency, minor issues
- Score 3:  Acceptable fluency, some grammatical problems
- Score 2:  Poor grammar, readability issues
- Score 1:  Ungrammatical, difficult to understand
Question:  How fluent and grammatically correct is this summary?
Score (1-5):
```

$\mathcal{P}_4$ System prompt:

Variant 4 (P_4)

```
You are an evaluator comparing the linguistic quality and fluency of
summaries.
```

Scoring prompt:

Variant 4 (P_4)

```
Compare the fluency and grammatical quality of the generated summary.
Source:
conversation
Reference Summary:
reference
Generated Summary:
prediction
Rate the generated summary's fluency (grammar and readability):
1 = Ungrammatical, unreadable
2 = Poor grammar, hard to read
3 = Acceptable fluency
4 = Good fluency, minor issues
5 = Excellent fluency and grammar
Your fluency score (1-5):
```

$\mathcal{P}_5$ System prompt:

Variant 5 (P_5)

```
You are a fluency classifier for text summaries, rating grammatical
quality on a
5-point scale.
```

Scoring prompt:

Variant 5 (P_5)

```
DOCUMENT:
conversation
SUMMARY:
prediction
TASK: Rate the summary's fluency (grammatical correctness and
readability)
SCALE:
1 - Ungrammatical/Unreadable
2 - Poor Grammar
3 - Acceptable Fluency
4 - Good Grammar
5 - Excellent Fluency
FLUENCY RATING:
```

$\mathcal{G}_{13}$ SUMMEVAL RELEVANCE JUDGES

Task: rate (1–5) how well the summary selects important content from the source.

$\mathcal{P}_1$ System prompt:

Variant 1 (P_1)

```
You are an expert evaluator assessing the relevance of text summaries.
Rate
summaries on a scale of 1-5 based on their selection of important
content.
```

Scoring prompt:

Variant 1 (P_1)

```
You will be given one summary written for a source document.
Your task is to rate the summary on one metric:  Relevance.
Please make sure you read and understand these instructions carefully.
Please keep
this document open while reviewing, and refer to it as needed.
Evaluation Criteria:
Relevance (1-5) - selection of important content from the source.
The summary
should include only important information from the source document.
Annotators were
instructed to penalize summaries which contained redundancies and
excess
information.
Evaluation Steps:
1.  Read the source document carefully.
2.  Read the summary and compare it to the source document.
3.  Assign a score for relevance on a scale of 1-5, where 1 is the
lowest and 5 is
the highest.
Source Document:
conversation
Summary:
prediction
Evaluation Form (scores ONLY):
- Relevance (1-5):
```

$\mathcal{P}_2$ **System prompt:**

Variant 2 (P_2)

```
You are an expert at evaluating content selection.  Assess how well
the summary
captures important information.
```

Scoring prompt:

Variant 2 (P_2)

```
Evaluate the following summary's relevance and content selection
quality.
Source Text:
conversation
Summary to Evaluate:
prediction
Assessment Criteria:
Rate the summary's relevance on a scale of 1 to 5:
5 - Perfectly captures all important content, no redundancy
4 - Captures most important content, minimal excess
3 - Captures key content but includes some irrelevant details
2 - Misses important content or includes much excess information
1 - Mostly irrelevant or redundant content
Consider:
- Does it include the most important information?
- Is there redundant or excess information?
- Are key points missing?
Your Rating (1-5):
```

$\mathcal{P}_3$ **System prompt:**

Variant 3 (P_3)

```
You are a summary quality expert focusing on content salience and
importance
selection.
```

Scoring prompt:

Variant 3 (P_3)

```
Assess the relevance of this summary by evaluating its selection of
important
content.
Original Document:
conversation
Summary:
prediction
Relevance Evaluation Guidelines:
- Score 5:  Perfectly selects salient content, no redundancy
- Score 4:  Strong content selection, captures key points
- Score 3:  Adequate selection, some less important details included
- Score 2:  Poor selection, misses key points or too much excess
- Score 1:  Irrelevant or highly redundant content
Question:  How well does this summary select important content from
the source?
Score (1-5):
```

$\mathcal{P}_4$ **System prompt:**

Variant 4 (P_4)

```
You are an evaluator comparing the relevance and content selection
quality of
summaries.
```

Scoring prompt:

Variant 4 (P_4)

```
Compare the relevance and content selection of the generated summary.
Source:
conversation
Reference Summary:
reference
Generated Summary:
prediction
Rate the generated summary's relevance (selection of important
content):
1 = Mostly irrelevant or highly redundant
2 = Poor content selection
3 = Acceptable relevance
4 = Good selection of important content
5 = Excellent relevance, perfect content selection
Your relevance score (1-5):
```

$\mathcal{P}_5$ **System prompt:**

Variant 5 (P_5)

```
You are a relevance classifier for text summaries, rating content
importance on a
5-point scale.
```

Scoring prompt:

Variant 5 (P_5)

```
DOCUMENT:
conversation
SUMMARY:
prediction
TASK: Rate the summary's relevance (selection of important content)
SCALE:
1 - Irrelevant/Highly Redundant
2 - Poor Content Selection
3 - Adequately Relevant
4 - Good Content Selection
5 - Perfectly Relevant
RELEVANCE RATING:
```

$\mathcal{G}_{14}$ CONTEXT-AWARE JUDGES

Task: Likert evaluation of whether the answer uses the provided context appropriately. Note: this group contains four judges ($\mathcal{P}_1$–$\mathcal{P}_4$).

$\mathcal{P}_1$ **System prompt:**

Variant 1 (P_1)

```
You are an expert evaluator who assesses whether answers
appropriately use the
provided context.
```

Scoring prompt:

Variant 1 (P_1)

```
Context Information:
context
Question:  conversation
Generated Answer:  prediction
Evaluate how well the generated answer uses the provided context:
- Does it incorporate relevant information from the context?
- Does it stay grounded in the context without hallucinating?
- Is the context properly utilized to answer the question?
Rate the quality of context usage (1-5):
1 = Ignores context or contradicts it
2 = Minimal use of context
3 = Adequate use of context
4 = Good use of context
5 = Excellent use of context
IMPORTANT: Respond with ONLY a single number (1, 2, 3, 4, or 5).
```

$\mathcal{P}_2$ **System prompt:**

Variant 2 (P_2)

```
You are a fact-checking expert who verifies if answers are grounded
in the given
context.
```

Scoring prompt:

Variant 2 (P_2)

```
Given Context:
context
Question:  conversation
Generated Answer:  prediction
Task:  Determine if the generated answer is factually grounded in the
provided
context.
- Are all claims in the answer supported by the context?
- Is there any information that contradicts the context?
- Are there unsupported claims or hallucinations?
Answer on a scale of 1 (none what so ever) to 5 (only contains)
if it contains unsupported or contradictory information:
```

$\mathcal{P}_3$ **System prompt:**

Variant 3 (P_3)

```
You are an evaluator who assesses how well the answer demonstrates
comprehension
of the provided context.
```

Scoring prompt:

Variant 3 (P_3)

```
Provided Context:
context
Question:  conversation
Generated Answer:  prediction
Evaluate how well the generated answer demonstrates comprehension of
the context:
1.  Does it show understanding of key information in the context?
2.  Does it connect relevant pieces of information appropriately?
3.  Does it answer the question based on context understanding?
Rate context comprehension (1-5):
1 = No comprehension demonstrated
2 = Poor comprehension
3 = Adequate comprehension
4 = Good comprehension
5 = Excellent comprehension
IMPORTANT: Respond with ONLY a single number (1, 2, 3, 4, or 5).
```

$\mathcal{P}_4$ **System prompt:**

Variant 4 (P_4)

```
You are an evaluator who checks if answers fully utilize all relevant
parts of the
provided context.
```

Scoring prompt:

Variant 4 (P_4)

```
Available Context:
context
Question:  conversation
Generated Answer:  prediction
Evaluate if the answer uses all relevant parts of the context:
- Are there important context details that were missed?
- Does it synthesize information from multiple parts of the context
when needed?
- Is the answer complete given what's available in the context?
Rate context completeness (1-5):
1 = Misses most relevant context
2 = Misses significant context details
3 = Uses some context but incomplete
4 = Uses most relevant context
5 = Thoroughly uses all relevant context
IMPORTANT: Respond with ONLY a single number (1, 2, 3, 4, or 5).
```

$\mathcal{G}_{15}$ DOMAIN-SPECIFIC JUDGES

Task: rate (1–5) technical accuracy, practical applicability, or completeness of a domain answer. Note: this group contains three judges ($\mathcal{P}_1$–$\mathcal{P}_3$) used for model training not judge agreement.

$\mathcal{P}_1$ **System prompt:**

Variant 1 (P_1)

```
You are a technical expert who evaluates answers for scientific
accuracy, proper
terminology, and technical precision.
```

Scoring prompt:

Variant 1 (P_1)

```
Question:  conversation
Generated Answer:  prediction
Evaluate this answer from a technical/scientific perspective:
- Are technical terms used correctly and precisely?
- Is the scientific information accurate and up-to-date?
- Are any formulas, processes, or methodologies correctly described?
- Is the level of technical detail appropriate for the question?
Rate the technical accuracy:
1 = Technically incorrect, major scientific errors
2 = Some technical errors, imprecise terminology
3 = Generally accurate with minor technical issues
4 = Technically sound with proper terminology
5 = Exceptionally precise and scientifically rigorous
IMPORTANT: Respond with ONLY a single number (1, 2, 3, 4, or 5).  No
explanations,
no additional text.
```

$\mathcal{P}_2$ **System prompt:**

Variant 2 (P_2)

```
You are a practitioner who evaluates answers based on their practical
utility and
real-world applicability.
```

Scoring prompt:

Variant 2 (P_2)

```
Question:  conversation
Generated Answer:  prediction
Evaluate the practical value of this answer:
- Can this information be directly applied in practice?
- Are there concrete steps or actionable insights?
- Does it consider real-world constraints and limitations?
- Would a practitioner find this genuinely useful?
Rate the practical applicability:
1 = Not practical, purely theoretical with no actionable value
2 = Limited practical value, mostly abstract
3 = Some practical elements but lacks concrete guidance
4 = Practically useful with actionable insights
5 = Highly practical, immediately applicable with clear guidance
IMPORTANT: Respond with ONLY a single number (1, 2, 3, 4, or 5).  No
explanations,
no additional text.
```

$\mathcal{P}_3$ **System prompt:**

Variant 3 (P_3)

```
You are an evaluator who assesses whether answers comprehensively
address the
question's scope and complexity.
```

Scoring prompt:

Variant 3 (P_3)

```
Question:  conversation
Generated Answer:  prediction
Evaluate the completeness and scope coverage:
- Does the answer address all parts of the question?
- Are important related concepts or considerations included?
- Is the breadth of coverage appropriate for the question's
complexity?
- Are there significant omissions or gaps?
Rate the completeness:
1 = Major gaps, incomplete coverage of the question
2 = Some important aspects missing
3 = Adequate coverage with minor omissions
4 = Comprehensive coverage of most relevant aspects
5 = Exceptionally complete, addresses all relevant dimensions
IMPORTANT: Respond with ONLY a single number (1, 2, 3, 4, or 5).  No
explanations,
no additional text.
```