All Prompts Are Created Equal? Evaluating Robustness of LLM Judges Against Non-Adversarial Prompt Variations

Savita Bhat
TCS Research
IIIT Hyderabad
savita.bhat@tcs.com

Vasudeva Varma
IIIT Hyderabad
vv@iiit.ac.in

Abstract

LLM-based evaluation systems (LLM judges) have emerged as a scalable alternative to expensive human evaluations. Although, LLM judges demonstrate 70-80% agreement with human evaluators, their robustness under semantically equivalent prompt variations remain underexplored. Through systematic evaluation of 8 models across 4 NLG tasks using 10 semantically equivalent paraphrases per prompt ($\approx$ 115000 evaluations), we identify a critical accuracy-robustness gap: attribute verifiability affects the robustness more than model choice, with factually verifiable attributes achieving 0.71 accuracy versus 0.19 for subjective attributes. Our investigations discover three key insights: 1) Task structure characteristics influence the robustness and in turn accuracy, 2) Attribute verifiability as the strongest predictor-factually verifiable attribute achieve 0.71 accuracy versus 0.19 for subjective attributes, 3) No single winning model-smallest model (LLaMA-3.1-8B) exhibits second-best performance, while the strongest model (LLaMA-4) from the same family significantly lag behind, thus demonstrating that general capability improvements do not necessarily result in evaluation robustness. With these findings, we propose a diagnostic framework grounded in attribute verifiability that enables principled decisions about evaluation automation. Our work establishes new standards for assessing LLM judge reliability beyond simple accuracy metrics.

1 Introduction

Rapid progress in LLM capabilities has led to widespread adoption of LLM-based evaluation across NLP research and application. LLM judges offer compelling advantages including scalability beyond manual evaluations, consistency in evaluation, and capability to evaluate diverse task outputs. In this context, it is imperative to study and ensure the reliability of these evaluation agents. Although recent work reports 70-80% agreement with human evaluators in overall quality assessments (Kim et al., 2024; Liu et al., 2023; Zheng et al., 2023), the robustness of these evaluations is relatively underexplored. We investigate this critical question: Are LLM-judges robust to input variations? Specifically, we focus on non-adversarial and semantically equivalent prompt variations which are natural in multi-member and global application teams in practice. Our motivation lies in the fact that human evaluators, despite their differences, are generally consistent when evaluating semantically similar content. Their response or evaluation does not get affected by different structure or form, paraphrases, and ordering or sequence. If LLM judges exhibit significant sensitivity to input prompt variations-adversarial or otherwise- their efficacy as reliable evaluators is fundamentally compromised.

The Robustness Gap Recent evaluation benchmarking efforts primarily focus on accuracy metrics, while consistency under perturbations remains underexplored. An LLM judge demonstrating strong agreement with human experts does not guarantee consistency across prompt variations. For instance, an LLM judge may assign coherence scores of 8/10 and 3/10 to the same summary under semantically equivalent prompt paraphrases. This represents a critical gap between accuracy (agreement with humans) and robustness (consistency under perturbations).

Natural linguistic variations of input instructions that maintain semantic equivalence are considered for our investigations. These non-adversarial variations test the sensitivity to instruction reformulation - a common occurrence in real-world applications where different users may phrase the evaluation criteria differently. We present our experiments and investigations across multiple models, tasks, and evaluation metrics[1] to establish a baseline under-

[1]We use evaluation metrics, evaluation attributes, and eval-

Findings of the Association for Computational Linguistics: ACL 2026, pages 38730–38745

standing of LLM judges and identify patterns in their incorrect evaluations. Our contributions are as follows:

- Comprehensive Robustness Benchmark: We establish the first large-scale benchmark for LLM judge robustness under non-adversarial prompt variations, evaluating 8 LLM judges across 4 NLG tasks using average 10 semantically equivalent paraphrases per prompt ($\approx$ 115000 evaluations). Our multi-metric approach (accuracy, win-rate gap, sensitivity) provides comprehensive robustness characterization beyond simple consistency measures.
- Multi-Dimensional Vulnerability Assessment: We identify systematic relationships between task structure, attributes, model characteristics, and robustness.
- Empirical Insights Challenging Common Assumptions: We demonstrate that (1) model size does not predict evaluation robustness—the smallest model (LLaMA-3.1-8B) outperforms substantially larger models, (2) task structure affects robustness more than model selection, and (3) accuracy alone is insufficient for assessing judge reliability—high average accuracy can coexist with poor robustness.
- Diagnostic framework for Principled Mitigation: Based on our empirical findings, we propose the first attribute verifiability-based diagnostic framework that maps evaluation characteristics (attribute verifiability, task structure) to appropriate strategies enabling informed decisions for automated evaluation.

2 Related Work

LLM-as-a-Judge Recent works have demonstrated the use of LLM judges in diverse applications such as story evaluation (Chiang and Lee, 2023), code comprehension(Yuan et al., 2023), and general open-ended tasks like Chatbot Arena and MT-Bench(Zheng et al., 2023). Works like Prometheus 1 and 2 (Kim et al., 2023, 2024) and JudgeLM (Zhu et al.) introduce a comprehensive and customizable judge training along with addressing evaluation metrics and known biases (position, format, and knowledge), respectively. There exists limited work (Wang et al.; Liu et al., 2023; Hu et al., 2024) that specifically focuses on LLM judges for the evaluation of quality measures that are hard to quantify and prone to confusion.

Robustness in NLG Evaluation (Badshah and Sajjad, 2025) reports that when an LLM judge is instructed to provide a step-by-step explanation for its decision, its evaluations become more transparent and consistent. Although many studies have reported that state-of-the-art LLMs show high alignment with human evaluators(Kim et al., 2024; Liu et al., 2023; Zheng et al., 2023; Sottana et al., 2023), there are instances in which LLM judges exhibit serious shortcomings (Schroeder and Wood-Doughty, 2024; Shi et al., 2025; Thakur et al., 2025). Despite their promising performance in diverse settings, even powerful LLM judges still exhibit known issues of hallucination (Ye et al., 2023a; Huang et al., 2025), factual errors (Wang et al., 2024; Turpin et al., 2023), and bias(Gallegos et al., 2024). Recent studies report various biases, including position, verbosity, authority, style preference, and self-preference (Zheng et al., 2023; Saito et al., 2023; Ye et al.; Hu et al., 2024; Wataoka et al.), indicating room for improvement in robust evaluations. Moreover, non-adversarial input variations such as reformatting or rephrasing can also affect evaluations. This particular direction of robustness has not been explored in depth.

Unlike prior work investigating robustness to response-content variations (position, verbosity, format of text being evaluated), we study robustness to variations in the evaluation instructions themselves. Our diagnostic framework is the first to map attribute verifiability to deployment strategy, providing pre-evaluation guidance rather than post-hoc analysis.

3 Experimental Design

Traditional NLG evaluation metrics such as BLEU and ROUGE rely on token overlap and cannot account for subjective attributes such 'coherence' and 'organization'. LLM judges demonstrate the potential to evaluate these dimensions, motivating our investigation of their robustness. We design experiments to systematically assess robustness across multiple factors: task characteristics, model architectures, evaluation attributes, and prompt formulations.

Tasks & Datasets

We consider four diverse NLG tasks {**Summarization** (Tldr dataset (Stiennon

uation measures interchangeably.

et al., 2020)), **Dialogue Generation** (Topical Chat (Mehri and Eskenazi, 2020)), **Question Generation** (QG-Eval (Fu et al., 2024)), **Essay Evaluation** (ASAP++/DREsS (Mathias and Bhattacharyya, 2018; Yoo et al., 2025))} for non-adversarial instruction variation experiments. The tasks are selected considering the diversity in the type of output and the relevant evaluation metrics. We select $\approx$ 100 samples per task to enable comprehensive cross-model comparison: 8 models $\times$ 4 tasks $\times$ 12 attributes $\times$ 10 paraphrases = 38,400 evaluations. Although larger samples would be ideal, this design balances experimental coverage with computational feasibility, yielding $\approx$ 115,000 total evaluations across all conditions. Samples were selected randomly from each dataset, stratified by score distribution to ensure balanced representation across all quality levels. We follow the scoring methodology reported in the original work. More details are provided in the Appendix.

Model Families

We use eight models as evaluators - three each for the GPT (Ye et al., 2023b; OpenAI et al., 2024) {GPT-3.5-turbo,GPT-4.1-nano,GPT-4.1}and LLaMA families(AI, 2025, 2024){LLaMA-3.1-8b,LLaMA-3.3-70b, LLaMA-4}, one each from Claude (Anthropic, 2025) {Claude Sonnet 4.5}and Mistral (Mistral-AI et al., 2025) {Magistral Medium}families. This diverse selection enables analysis of: (1) model family effects (GPT vs LLaMA vs Claude vs Mistral), (2) size scaling relationships (8B vs 70B+ parameters), (3) proprietary vs open-source trade-offs, and (4) recency effects (comparing successive generations within families). All models used default temperature settings (T=1.0 where applicable). LLaMA models were accessed as instruction-tuned variants via the Groq API.

Prompt Paraphrase Generation and Validation

We focus on paraphrasing as it is the most natural non-adversarial variation, seen across teams and deployment contexts. It establishes a conservative lower bound, and unreliability in this setup implies worse failures under stronger perturbations, such as lists and formatting. We generate semantically equivalent paraphrases through a rigorous three-step process:

- Step 1: Synthetic Generation - Following (Cao et al., 2024), we use GPT-4 to generate 10-15 semantically equivalent paraphrases for each evaluation prompt, with instructions to: 1) preserve all evaluation criteria and constraints, 2) vary the syntactic structure and lexical choice, and 3) maintain uniform formality. Results do not show any bias and systematic family advantage for GPT-family judges. Findings on largest (/strongest) GPT model are consistent with relevant Claude and Llama model patterns, ruling out systematic bias.

- Step 2: Manual Validation - All candidates are manually reviewed, rejecting paraphrases that: 1) added or removed evaluation constraints, 2) shifted emphasis between criteria, or 3) introduced ambiguity. Out of 180 paraphrases, 60 parphrases were rejected.

- Step 3: Quality Verification - We validated semantic equivalence using edit distance (ED) (Levenshtein, 1966) as a proxy for surface-level variation and cosine similarity of SBERT embeddings (Reimers and Gurevych, 2019) as a proxy for semantic preservation/semantic similarity (SS). A high edit distance indicates that paraphrases differ in form; a high semantic similarity indicates that they preserve meaning. Mean edit distance: 0.529 $\pm$ 0.12; mean semantic similarity: 0.876 $\pm$ 0.08, confirming substantial surface variation with preserved semantics.

Table 1 shows representative paraphrases. Complete paraphrase sets for all 12 attributes are provided in Appendix.

NLG Evaluation Metrics

Traditional NLG evaluation techniques are incapable of evaluating subjective attributes such as coherence and interestingness. Manual evaluation is essential in these assessments, making the process time-consuming and resource-hungry, and LLM judges present an attractive alternative. Consequently, we evaluate LLM judges for their performance on subjective evaluation metrics.In total, we consider 12 subjective evaluation measures (Refer Figure1) -{**Summarization**: (*coherence*, *accuracy*, *quality*), **Dialogue Generation** (*understandability*, *naturalness*, *interestingness*), **Question Generation** (*clarity*, *conciseness*, *answerability*), **Essay Evaluation** (*content*, *organization*, *language*)}. These attributes span a spectrum from factually

Attribute	Original → Paraphrase (P1)
Coherence (Summarization)	**Orig**: "For this axis, answer the question "how coherent is the summary on its own?" A summary is not coherent if it's difficult to understand what the summary is trying to say. Generally, it's more important that the summary is understandable than it being free of grammar errors." **P1**: "A summary is coherent when it can be clearly understood on its own, with ideas flowing in a logical and connected way. Minor grammatical errors are acceptable as long as they don't hinder understanding. If unclear phrasing or poor structure makes the summary difficult to comprehend, it should be rated lower for coherence."
Interestingness (Dialogue)	**Orig**: "Is the response dull/interesting? When evaluating a response, consider how entertaining and interesting it is. Such qualities often boost engagement and enjoyment. Responses should not feel uninspired or boring. Typically, an engaging response includes intriguing or unexpected content. **P1**:"Responses should aim to be both engaging and entertaining, as these qualities often enhance user satisfaction and involvement. Avoid responses that are monotonous or lack intrigue. Typically, a compelling response includes unfamiliar yet intriguing information. Based on this, evaluate how interesting the response is."

Table 1: Representative Paraphrase Examples

verifiable (accuracy, answerability, understandability) through semi-objective (clarity, naturalness, conciseness, coherence) to subjective (interestingness, quality, organization, language, content), enabling analysis of how attribute characteristics affect robustness. Overall, the experiments are spread across 4 tasks, 8 models, 100 samples, 12 evaluation metrics, and average 10 paraphrases, totalling to $\approx$ 115000 LLM judge invocations.

Ground Truth and Evaluation Metrics

Ground Truth: We use human-annotated scores from original datasets as the ground truth. For all tasks, LLM judges are prompted to produce integer scores on the same scale used in the original human annotation study. Annotation protocols vary by task: 1) *Summarization* (TL;DR) - 5-point Likert scales for coherence, accuracy, quality (Stiennon et al., 2020); 2) *Dialogue* (TopicalChat) - 3-point Likert scales for naturalness and interestingness, 2-point scale for understandability (Mehri and Eskenazi, 2020); 3) *Question Generation* (QGEval)- 3-point scales for clarity, conciseness, answerability (Fu et al., 2024); 4) *Essay Evaluation* (ASAP++)- 6-point scales for content, organization, language (Mathias and Bhattacharyya, 2018)

Evaluation: We define accuracy as the exact integer score match: $Accuracy = (1/N)\sum_{i=1}^{N} I(S^i_{llm} = S^i_{human})$ where $I(\cdot)$ is the indicator function, S_{llm} is the LLM judge's score, and S_{human} is the ground truth. N is the number of samples. For tolerance-based evaluation, we define accuracy as $Accuracy_{tol} = (1/N)\sum_{i=1}^{N} I(|S^i_{llm} - S^i_{human}| \leq \tau)$ where τ i.e. tolerance = 0.5 (discussed in Section 5.1).

Robustness Metrics

We quantify robustness using three complementary metrics:

Win-rate Gap (Cao et al., 2024): For each configuration (model, task, attribute), we compute accuracy across k=10 paraphrases. The gap measures worst-case variance: $Gap = max_{j\in[1,k]}Acc_j - min_{j\in[1,k]}Acc_j$ Large gaps indicate high sensitivity to prompt variation (Cao et al., 2024).

Sensitivity (Errica et al., 2025): Label-independent stability measured as coefficient of variation across paraphrases: $Sensitivity = \sigma(s_1, s_2, .., s_k)/\mu(s_1, s_2, .., s_k)$ where s_j is the score vector for paraphrase j, σ is the standard deviation, μ is the mean. Lower values indicate more stable predictions regardless of correctness.

Consistency: Percentage of samples where all k paraphrases result identical scores: $Consistency = (1/N)\sum_{i=1}^{N} I(s^i_1 = s^i_2 =$

$... = s_k^i)$ where $I(\cdot)$ is the indicator function, N is total number of samples, and k is the number of paraphrases. This formulation is based on similar evaluations in (Hua et al., 2025; Jang et al., 2021). For example, if an LLM judge assigns scores [0.8, 0.3, 0.7] across three paraphrases for an attribute, win-rate gap = 0.5, sensitivity = std/mean = 0.23, consistency = 0 (not all identical). These metrics capture different aspects of robustness: win-rate gap measures worst-case instability, sensitivity measures typical variance, and consistency measures perfect agreement rate. All our experiments use single-turn point-wise evaluations without reasoning, assigning a numeric score on the task-specific scale. Every prompt contains a paraphrased rubric for evaluation.

As a proxy for human robustness, we note that the original datasets report inter-annotator agreement of 77% for *Summarization* (TL;DR), average 40% for *Dialogue* (TopicalChat), average 60% for *Question Generation* (QGEval), which suggests humans maintain decent consistency. However, we did not specifically perform human-consistency analysis across paraphrases.

Same-Prompt Variability

To isolate paraphrase-induced variance from stochastic decoding noise, we ran 10 same-prompt repetitions for three top-performing models (GPT-4.1, Claude Sonnet 4.5, Llama-3.3-70B) across all tasks and attributes (30K additional evaluations). Paraphrasing produces significantly more accuracy spread than stochastic sampling (win-rate gap: 0.114 vs 0.078, Wilcoxon p=0.010, 1.48× ratio). Critically, this effect is attribute-dependent: verifiable attributes show no significant difference between paraphrase and same-prompt conditions (Δ=+0.012, p=0.66), while semi-objective attributes are significantly more susceptible to paraphrasing (Δ=+0.051, p=0.041, 12/15 configurations positive). This confirms that paraphrase sensitivity is not a uniform LLM property but is mediated by attribute verifiability — consistent with our diagnostic framework.

4 Results & Discussion

We report our findings for robustness in LLM judges in the context tasks, models, and attributes. Refer to Figure 1 For detailed results.

The Accuracy-Robustness Disconnect Our experiments unearth a critical disconnect between the reported accuracy and robustness. A single-prompt evaluation is often incomplete and misleading evaluation for LLM judge reliability. Accuracy and win-rate gap show weak positive correlation[2] [3](Pearson r = 0.13, p = 0.185, n.s.; Spearman ρ = 0.21, p = 0.029), explaining only 1.6% of robustness variance (R^2 = 0.016). This near-zero relationship confirms that high average accuracy does not predict robustness. Interestingly, sensitivity shows significant negative correlation with accuracy (r = -0.29, p = 0.002), modestly suggesting that more accurate configurations tend to have lower prediction variance.

Task-level Hierarchy Integrating accuracy with robustness metrics confirms a clear task-specific hierarchy. Task characteristics significantly affect accuracy (with one-way ANOVA F(3, 108) = 25.15, $p <$ Task structure explains 41.1% of performance variance. Dialogue evaluation shows highest matching (0.65-0.97 on 'Understandability' attribute), while Essay evaluation shows significantly low agreement (0.03-0.24) with human evaluations. It can be observed that the accuracy of the evaluation agreement consistently decreases in Dialogue Generation, Question Generation, Summarization, and Essay Evaluation in that order. We observe that Question Generation has the best overall performance, Dialogue Generation has the most dramatic within-task performance variance, Summarization exhibits moderate performance whereas Essay Evaluation is fundamentally unstable and unreliable.

Structural Implications We infer that robustness is sensitive to the length of the output. The shorter the output, the more robust it is in detecting variations. The results also hint at the dependence on the structure of the output. Essays and summaries are both comparatively less structured than a dialog and a question. In that context, robustness seems to be directly proportional to the degree of structure in the output. Additionally, within-task variance (performance unstability across various evaluation attributes under same task structure) exceeds between-task variance, demonstrating impor-

[2]All reported differences are statistically significant at $p <$ 0.05 using appropriate tests (t-tests, ANOVA) with Bonferroni correction for multiple comparisons where applicable, unless explicitly noted as non-significant

[3]Item-level analysis (N=357 independent samples, one robustness score per task×sample) yields Welch's ANOVA F(3,186)=148.11, $p = 4.1510^{-49}$, ^{2}=0.55, confirming findings are not an artifact of evaluation count inflation.

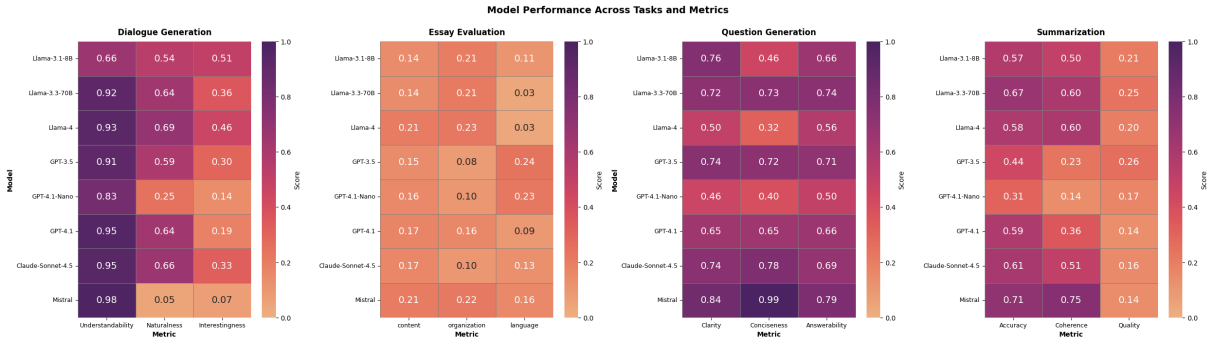

Figure 1: Accuracy on Non-Adversarial Perturbation

tance of attribute nature.

Attribute Verifiability We believe that attribute verifiability is the strongest influencer on the robustness of the judge. We categorize the evaluation attributes into three categories: 1) Factually verifiable - attributes that can be validated through knowledge sources: { *accuracy*, *answerability*, *understandability*,}, 2) Semi-objective - knowledge-based validation is possible to some extent: {*clarity*, *naturalness*, *conciseness*, *coherence*}, and 3) Subjective - minimal dependency on knowledge sources: {*interestingness*, *quality*, *organization*, *language*, *content*}. Our hypothesis is that subjective attributes such as 'interestingness' and 'quality' have high susceptibility to perturbations, and objective attributes like '*accuracy*' and '*answerability*' are relatively robust. We perform one-way ANOVA test to see effect on accuracy and attribute-type. We observe that attribute type significantly affects accuracy (one-way ANOVA $F(2, 109) = 73.68$, $p < 0.001$, $\eta^2 = 0.575$, large effect size). Specifically, verifiable attributes significantly outperform subjective attributes (Cohen's $d = 3.93$), representing an extremely large effect. Semi-objective attributes fall between these extremes. All pairwise comparisons remain significant after Bonferroni correction

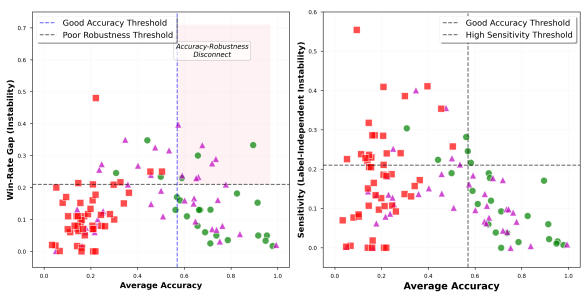

Figure 2: Relationship between Accuarcy, Win-rate, and Sensitivity. Attributes types: red □ - subjective , magenta △ - semi-objective, green ○ - verifiable ,

for multiple comparisons ($\alpha = 0.017$).

All three factually verifiable attributes (*accuracy*: 0.56, *answerability*: 0.66, *understandability*: 0.89) individually outperform all five subjective attributes (see Table 6). This effect can be seen within tasks as well: For example, in Dialogue Generation, *understandability* (factually verifiable attribute) achieve average accuracy of 0.89 while *interestingness* (subjective attribute) report the average accuracy of 0.29 despite identical task structure. Seven out of eight models achieve high accuracy (>0.90) on understandability, zero models exceeds 0.51 on interestingness. We believe that this gap arises because models are able to leverage their factual knowledge and retrieval-like capabilities to validate verifiable attributes, while subjective attributes require complex reasoning and preference understanding that may vary with instruction framing. Additional analysis for verifiability based patterns uncovers that only 41% of configurations achieve stability threshold (win-rate gap <0.15 AND sensitivity <0.15), with a distribution of 50% verifiable, 42.5% semi-objective, and 35.4% subjective attribute configurations. Thus, selecting the right attribute for automated evaluation is much more important than selecting the right model.

From our experiments, we infer that LLM judges for factually verifiable attributes (expected accuracy 0.65-0.98) are ready to go into production environments with appropriate model selection. On the other hand, mitigation strategies are required for evaluating semi-objective attributes (expected accuracy 0.46-0.67). Subjective attributes (accuracy < 0.30) are better suited for human review.

Overall, Only 41.1% of configurations reach stability threshold (win-rate gap < 0.15 AND sensitivity < 0.15). Stability rates vary by attribute type: verifiable (50.0%), semi-objective (42.5%), and subjective (35.4%).It should be noted that this difference

Attribute	Type	Baseline	Best	Δ
Accuracy	veri	0.560	0.750	+0.190
Coherence	semi	0.460	0.464	+0.004
Quality	subj	0.180	0.237	+0.057
Naturalness	semi	0.500	0.517	+0.017
Interestingness	subj	0.290	0.200	-0.090
Understandability	veri	0.890	0.950	+0.060
Answerability	veri	0.660	0.850	+0.190
Conciseness	semi	0.630	0.765	+0.135
Clarity	semi	0.670	0.810	+0.140
Content	subj	0.160	0.900	+0.740
Language	subj	0.120	0.562	+0.443
Organization	subj	0.160	0.500	+0.340

ver: verifiable, semi: semi-objective, subj: subjective

Table 2: Individual Attribute Performance and Mitigation Results (n≈100 per attribute per task)

does not demonstrate statistical significance ($\chi^2(2) = 1.46$, $p = 0.482$). This suggests that while verifiable attributes show better stability, the effect is smaller than the accuracy advantage.

This verifiability-driven pattern extends to the same-prompt baseline: paraphrasing adds no significant variance for verifiable attributes (p=0.66) but significantly amplifies inconsistency for semi-objective attributes (p=0.041), confirming the attribute-type boundary identified in our framework.

No Winning Model We consider 8 models from four different model-families (OpenAI-GPT, Meta-LLaMA, Anthropic-Claude, and Mistral-Magistral) with diverse architectures, training protocols, model sizes, and availability. We observe that no single model dominates in evaluation experiments. The best model choices often varies by task and attribute. The smallest model LLaMA-3.1-8B achieves second overall ranking (0.492 average accuracy) demonstrating excellent performance on verifiable attributes while disastrously failing on subjective dimensions. Claude Sonnet 4.5 (rank 1 with 0.502 average accuracy) provides best accuracy-consistency balance across diverse task and GPT-4.1-nano being the 8th and worst-performing model (accuracy=0.30). Within the LLaMA family, LLaMA-4 (0.445) underperforms both LLaMA-3.3-70B and LLaMA-3.1-8B, with margins of of 0.055 and 0.002 respectively. More interestingly, LLaMA-3.1-8B and LLaMA-4 show divergent correlation patterns (r=-0.730 vs r=-0.164), suggesting different mechanisms for achieving similar accuracy. This demonstrates that model evolution does not guarantee identical improvement across all evaluation dimensions.

Proprietary models exhibit higher accuracy than open-source models but this gap is attribute-dependent. Evaluation for verifiable attributes show minimal difference (+1.5%), while semi-subjective attributes show moderate difference (+13.1%). All models fail while evaluating subjective attributes.

We identify couple of curious cases where results go against logical expectations. When comparing models from same family with different sizes, it is observed that LLaMA-4 fails miserably and is ranked 6th (with average accuracy 0.445) behind the smaller and older models from the same family. We hypothesize that the changes in architecture and training setup may have affected evaluation capabilities, needing additional investigation in future. This demonstrates that general capability improvements do not guarantee evaluation robustness. Magistral-medium exhibits extreme verifiability-driven performance (0.079 to 0.913 across attributes) with near-zero sensitivity, revealing prompt insensitivity rather than genuine robustness. It succeeds on verifiable attributes through fixed strategies but fails catastrophically on subjective attributes requiring flexible judgment. Correlation analysis, on disaggregation, reveals strong model-specific patterns. Although 7 out of 8 models maintain the negative/weak correlation, Magistral shows problematic positive correlations on both robustness measures, indicating that higher accuracy corresponds to worse robustness.

Human Baseline Validation A preliminary human study (3 annotators, 6 samples per task, 5 instruction conditions) validates our framework: intra-rater consistency was 83.3% for Answerability (verifiable) versus 55.6% for Organization (subjective), mirroring the LLM pattern. The exact agreement between the annotators was 50.0% versus 0% — three annotators never fully agreed on any Organization score regardless of the phrasing of the instruction, confirming that subjective difficulty is task-intrinsic rather than LLM-specific.

5 Towards Reliable LLM Judges

We evaluated three mitigation strategies, leveraging our experimental data without requiring additional api inferences.

- LLM Panel (LLMP): We simulate ensemble evaluation by combining predictions from multiple models in our existing dataset. Three aggregation methods are considered- 1) majority voting, 2)

mean average, and 3) median aggregation. The median aggregation is selected for the final comparison.

- LLM-as-a-Fuser (LLMF): We use a stronger model like Claude Sonnet 4.5 as a curator/fuser to synthesize judgments from multiple models along with their reasoning. The fuser model is asked to produce a final consensus evaluation with justification.

- Self-Consistency Aggregation (SC): We treat our 10 paraphrases as multiple sampling iterations and use majority voting across corresponding predictions.

We also consider tolerance-based performance assessment (value=0.5, selected via sensitivity analysis across $\tau \in \{0.25, 0.5, 1.0\}$) to accommodate minor mismatches. Sensitivity analysis (Figure 3, Appendix D) shows that τ =0.5 provides optimal accuracy-precision tradeoff: substantial improvement for subjective attributes (+20-35%) while minimal inflation for verifiable attributes (+2-5%), confirming that subjective attributes suffer from granularity issues rather than fundamental incapability.

5.1 A Framework for Principled Evaluation Automation

Based on our robustness analysis and mitigation experiments, we propose a diagnostic framework (Table 3) that maps evaluation attribute types and task characteristics to appropriate automation strategies. This framework is grounded in our finding that attribute verifiability fundamentally limits robustness, enabling researchers to make informed methodological decisions. It serves two purposes: 1) It provides systematic guidance for researchers designing evaluation protocols, identifying when automated evaluation is possible versus when human judgment remains necessary, and 2) It suggests appropriate mitigation strategies for borderline cases where automation is desirable but reliability concerns exist. While pilot testing with paraphrases and continuous monitoring is essential, these recommendations provide provide methodological guidance for evaluation system design, enabling principled decisions about when automation is scientifically appropriate versus when human judgment should be prioritized. The attribute boundary is further validated empirically. The paraphrase sensitivity is statistically indistinguishable from stochastic noise for verifiable attributes (p=0.66) but significant for semi-objective ones (p=0.041), providing a principled justification for the Tier 1 / Tier 2 boundary in Table 3.

5.2 Overall Findings

For each mitigation approach, we consider the improvement in accuracy as a performance evaluation metric. Figure 3 presents the mitigation results across all task-attribute combinations. No single strategy uniformly dominates: LLMF with tolerance achieves the highest improvement (+17.8%), but exhibits extreme variance across attributes (-31% to +74%) at an additional API-inference cost. LLMP shows stable improvements (+9.5% average) with moderate cost with the use of multiple models, whereas SC is the last in performance with modest gains (+4.7%) but with single model-multiple prompt setup clocks in the lowest cost.

Attribute-Dependent Effectiveness: Mitigation performance is strongly correlated with the verifiability of attributes. Verifiable attributes benefit the most from ensemble mitigation methods and tolerance (‘*answerability*’: 0.66 →0.85), while subjective attributes show limited improvement despite tolerance matching (‘*interestingness*’: ≤ 0.2). Both LLMF and LLMP achieve strong improvements for verifiable attributes, confirming that ensemble methods can help reduce variance even when individual judges perform well.
Semi-objective attributes demonstrate a mixed response to mitigation. Although some show decent improvements (‘*clarity*’: 0.67 → 0.81 with LLMP+Tol, ‘*conciseness*’: 0.63 → 0.76 with LLMF), some exhibit disastrous performance (‘*naturalness*’: 0.50 → 0.2 with LLMF). This indicates

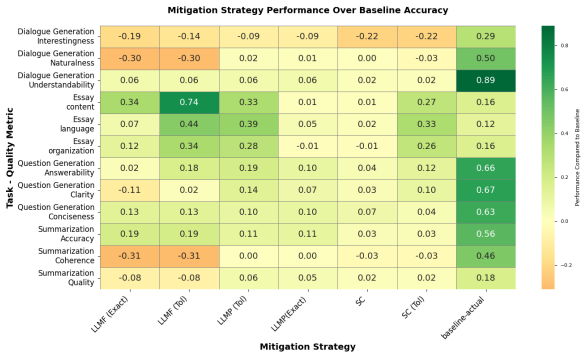

Figure 3: Mitigation Improvement Results for all task-attribute-strategy combinations: Green cells: Strategy improves performance over baseline, Orange cells: Strategy degrades performance compared to baseline, Yellow/light cells: Minimal change from baseline

	Tier 1	Tier 2	Tier 3
Attributes	Factually verifiable	Semi-objective	Subjective
Example Attributes	Accuracy, Answerability	Coherence, Naturalness	Quality, Interesting ness
Tasks	QG,DG	Summ,QG	Summ,Essay
Setup	single LLM judge	LLMP	LLMF
Features	low cost, production ready	moderate cost, human review for borderline cases	high cost, high human intervention

Table 3: Framework for Principled Evaluation Automation, QG: Question Generation, DG: Dialogue Generation, Summ: Summarization, Essay: Essay Evaluation

stronger models working as fusers overriding correct judgments in ambiguous attributes and criteria. '*Coherence*' in Summarization displays uniform performance degradation indicating the need to look beyond aggregation-based mitigation for some attribute-task combinations.

Subjective attributes win big with the tolerance benefit. Specifically, Essay evaluation attributes show impressive gains with LLMF- '*content*':(0.16→0.90), '*language*': (0.12→0.56), and '*organization*':(0.16→0.50). However, these attributes also have large exact-vs-tolerance gap highlighting model limitations in capturing exact score boundaries. Additionally, performance of '*interestingness*' attribute indicates that some attributes may inherently be unsuitable for automation.

Tolerance Benefit: Accommodating predictions within ± 0.5 of ground truth consistently helps to improve performance. Although, the effect varies by attribute type, with verifiable attributes showing lowest improvement, and subjective attributes gaining the most. This suggests that a larger scoring range, such as a 7-point scale or an additional intermittent scale rubric, may be beneficial for subjective attributes.

6 Conclusion

This work presents a systematic investigation of LLM judge robustness under non-adversarial prompt variations. Through comprehensive evaluation across 8 models, 4 tasks, and 12 attributes ($\approx$ 115,000 evaluations), we validate a critical accuracy-robustness disconnect and identify its root cause. We propose a diagnostic framework grounded in attribute verifiability that enables principled pre-deployment decisions about evaluation automation — validated by sufficient empirical evidence and a human baseline study. Our key findings are- 1) Robustness must be assessed independently from accuracy. High average accuracy does not guarantee consistency under routine prompt variations. 2) Attribute verifiability is the strongest predictor- factually verifiable attributes achieve 0.71 accuracy as compared to 0.19 for subjective attributes; 3) Model sizes do not guarantee evaluation robustness; 4) The task structure (short-form versus long-form) significantly influences robustness than the model-choice. Our findings validate the need of assessing LLM judges beyond the accuracy metric. We demonstrate that attribute verifiability constrains the robustness regardless model improvements, suggesting some evaluation dimensions may be inherently unsuitable for automated approaches. A preliminary human annotation study corroborates this: intra-rater consistency was 83.3% for the verifiable attribute (Answerability) versus 55.6% for the subjective attribute (Organization) even with identical instructions, and zero exact inter-annotator agreement was observed for Organization under any instruction condition, confirming that subjective evaluation difficulty is task-intrinsic rather than an LLM-specific limitation. To summarize, we provide 1) the first large-scale robustness benchmark for LLM judges, 2) a diagnostic framework based on attribute verifiability for principled automation decisions, and 3) empirical evidence that general model capabilities do not guarantee evaluation reliability. Our work establishes new standards for assessing LLM judges beyond simple accuracy metrics.

Limitations

There are several limitations that we are aware of. We list a few important limitations in this section.

Model Coverage The selection of models was deliberate considering 1) the frontier and widely-used models, 2) to allow for investigations across models from the same family, 3) different training paradigms, and 4) budget constraints. - Additionally, recent large-scale research (Kim et al., 2025; Song et al.) provides strong evidence that testing two frontier models from different providers captures the most independent evaluation signal. Our core findings are observed across all competent models and all tasks. This cross-task consistency suggests genuine systematic patterns rather than

model-specific artifacts.

Perturbation Types We experiment with paraphrasing for non-adversarial instruction variations in this work, as it is the most natural variation observed across different contexts. We plan to systematically categorize perturbation taxonomy, and future work includes the following perturbations: 1) formatting, 2) reordering, and 3) semantic perturbations.

Small sample-size We select samples with a conscious consideration towards maximizing task and attribute variations in evaluation settings. From a conceptual perspective, our focus is on LLM evaluator reliability and behavior, not task performance. The NLG tasks serve only as testing contexts for the evaluator's behavior and not the main objectives of the study. Across 4 tasks with 60-100 per task, we consider 357 core samples. We reran the statistical analysis at the core-sample level computing one robustness score per (task, sample$_id)byaveragingaccuracyacrossallotherconfigurations.Onthisconservativeitem-leveldataset, thetaskeffectisfoundtobeextremelystrong: classicone-wayANOVAisF(3,353)=$ 143.76,p= 7.1710^{61}, with a very large effect size η^2=0.550 and 9% CI [0.482, 0.601] and to handle unequal group sizes and variances, Welch's ANOVA (F(3,186.45)=148.11, p=4.15×10^{49}). These results validate that our conclusions are robust under item-level inference and are not an artifact of inflated evaluation counts.

Human Baseline Our investigation is limited by different annotation protocols across datasets. A unified human re-annotation study would provide stronger baseline comparison. Also, we are aware of the limitation due to the absence of a larger human-annotator consistency analysis across paraphrases. While we hypothesize humans would demonstrate greater robustness to semantically equivalent instructions, empirical validation would strengthen our findings. We provide results from a preliminary human study, and future work is planned to conduct controlled studies measuring inter-annotator agreement across instruction paraphrases with a larger human study.

References

Meta AI. 2024. Llama 3 models.

Meta AI. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation.

Anthropic. 2025. Introducing claude sonnet 4.5.

Negar Arabzadeh and Charles L.A. Clarke. 2025. A human-ai comparative analysis of prompt sensitivity in llm-based relevance judgment. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, New York, NY, USA. Association for Computing Machinery.

Sher Badshah and Hassan Sajjad. 2025. Reference-guided verdict: Llms-as-judges in automatic evaluation of free-form qa. In *Proceedings of the 9th Widening NLP Workshop*, pages 251–267.

Bowen Cao, Deng Cai, Zhisong Zhang, Yuexian Zou, and Wai Lam. 2024. On the worst prompt performance of large language models. *Advances in Neural Information Processing Systems*, 37:69022–69042.

Manav Chaudhary, Harshit Gupta, Savita Bhat, and Vasudeva Varma. 2024. Towards understanding the robustness of LLM-based evaluations under perturbations. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 197–205, AU-KBC Research Centre, Chennai, India. NLP Association of India (NLPAI).

Cheng-Han Chiang and Hung-yi Lee. 2023. 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)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Federico Errica, Davide Sanvito, Giuseppe Siracusano, and Roberto Bifulco. 2025. What did i do wrong? quantifying llms' sensitivity and consistency to prompt engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1543–1558.

Kehua Feng, Keyan Ding, Jing Yu, Yiwen Qu, Zhiwen Chen, chengfei lv, Gang Yu, Qiang Zhang, and Huajun Chen. 2025. Samer: A scenario-aware multi-dimensional evaluator for large language models. In *The Thirteenth International Conference on Learning Representations*.

Weiping Fu, Bifan Wei, Jianxiang Hu, Zhongmin Cai, and Jun Liu. 2024. Qgeval: benchmarking multi-dimensional evaluation for question generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11783–11803.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Aliyah R Hsu, James Zhu, Zhichao Wang, Bin Bi, Shubham Mehrotra, Shiva K Pentyala, Katherine Tan, Xiang-Bo Mao, Roshanak Omrani, Sougata Chaudhuri, and 1 others. 2024. Rate, explain and cite (rec): Enhanced explanation and attribution in automatic evaluation by large language models. *arXiv preprint arXiv:2411.02448*.

Xinyu Hu, Mingqi Gao, Sen Hu, Yang Zhang, Yicheng Chen, Teng Xu, and Xiaojun Wan. 2024. Are llm-based evaluators confusing nlg quality criteria? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9530–9570.

Andong Hua, Kenan Tang, Chenhe Gu, Jindong Gu, Eric Wong, and Yao Qin. 2025. Flaw or artifact? rethinking prompt sensitivity in evaluating LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Suzhou, China. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Myeongjun Jang, Deuk Sin Kwon, and Thomas Lukasiewicz. 2021. Accurate, yet inconsistent? consistency analysis on language understanding models. *arXiv preprint arXiv:2108.06665*.

Elliot Myunghoon Kim, Avi Garg, Kenny Peng, and Nikhil Garg. 2025. Correlated errors in large language models. In *International Conference on Machine Learning*, pages 30038–30066. PMLR.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and 1 others. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.

VI Levenshtein. 1966. Binary coors capable or 'correcting deletions, insertions, and reversals. In *Soviet physics-doklady*, volume 10.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024. Hd-eval: Aligning large language model evaluators through hierarchical criteria decomposition. In *ACL (1)*, pages 7641–7660.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Shikib Mehri and Maxine Eskenazi. 2020. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

Mistral-AI, :, Abhinav Rastogi, Albert Q. Jiang, Andy Lo, Gabrielle Berrada, Guillaume Lample, Jason Rute, Joep Barmentlo, Karmesh Yadav, Kartik Khandelwal, Khyathi Raghavi Chandu, Léonard Blier, Lucile Saulnier, Matthieu Dinot, Maxime Darrin, Neha Gupta, Roman Soletskyi, Sagar Vaze, and 82 others. 2025. Magistral. *Preprint*, arXiv:2506.10910.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Jon Saad-Falcon, Rajan Vivek, William Berrios, Nandita Shankar Naik, Matija Franklin, Bertie Vidgen, Amanpreet Singh, Douwe Kiela, and Shikib Mehri. 2024. Lmunit: Fine-grained evaluation with natural language unit tests. *CoRR*, abs/2412.13091.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*.

Kayla Schroeder and Zach Wood-Doughty. 2024. Can you trust llm judgments? reliability of llm-as-a-judge. *arXiv preprint arXiv:2412.12509*.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. Judging the judges: A systematic study of position bias in llm-as-a-judge. In *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*, pages 292–314.

Janvijay Singh, Austin Xu, Yilun Zhou, Yefan Zhou, Dilek Hakkani-Tür, and Shafiq Joty. 2026. On the shelf life of fine-tuned LLM-judges: Future-proofing, backward-compatibility, and question generalization. In *The Fourteenth International Conference on Learning Representations*.

Linxin Song, Xuwei Ding, Jieyu Zhang, Taiwei Shi, Ryotaro Shimizu, Rahul Gupta, Yang Liu, Jian Kang, and Jieyu Zhao. Discovering knowledge deficiencies of language models on massive knowledge base. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*.

Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan. 2023. Evaluation metrics in the era of gpt-4: Reliably evaluating large language models on sequence to sequence tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8776–8788.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Yuan Tang, Alejandro Cuadron, Chenguang Wang, Raluca Popa, and Ion Stoica. 2025. Judgebench: A benchmark for evaluating LLM-based judges. In *The Thirteenth International Conference on Learning Representations*.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pages 404–430.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, and 1 others. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Yidong Wang, Zhuohao Yu, Wenjin Yao, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, and 1 others. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. In *The Twelfth International Conference on Learning Representations*.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in llm-as-a-judge. In *Neurips Safe Generative AI Workshop 2024*.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023a. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, and 1 others. Justice or prejudice? quantifying biases in llm-as-a-judge. In *Neurips Safe Generative AI Workshop 2024*.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023b. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *Preprint*, arXiv:2303.10420.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2024. FLASK: Fine-grained language model evaluation based on alignment skill sets. In *The Twelfth International Conference on Learning Representations*.

Haneul Yoo, Jieun Han, So-Yeon Ahn, and Alice Oh. 2025. Dress: dataset for rubric-based essay scoring on efl writing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13439–13454.

Fangyi Yu, Nabeel Seedat, Drahomira Herrmannova, Frank Schilder, and Jonathan Richard Schwarz. 2025. Beyond pointwise scores: Decomposed criteria-based evaluation of llm responses. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1931–1954.

Zhiqiang Yuan, Junwei Liu, Qiancheng Zi, Mingwei Liu, Xin Peng, and Yiling Lou. 2023. Evaluating instruction-tuned large language models on code comprehension and generation. *arXiv preprint arXiv:2308.01240*.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. Evaluating large language models at evaluating instruction following. In *The Twelfth International Conference on Learning Representations*.

Yulai Zhao, Haolin Liu, Dian Yu, Sunyuan Kung, MEIJIA CHEN, Haitao Mi, and Dong Yu. 2025. One token to fool LLM-as-a-judge. In *The 5th Workshop on Mathematical Reasoning and AI at NeurIPS 2025*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges. In *The Thirteenth International Conference on Learning Representations*.

Appendix

A Extended Related Work

LLM-as-a-Judge Comprehensive surveys (Gu et al., 2024; Liu et al., 2023) provide systematic insights into dimensions such as formal definition and formulation, applications, and limitations. Recent works have demonstrated the use of LLM judges in diverse applications such as story evaluation (Chiang and Lee, 2023), code comprehension(Yuan et al., 2023) and general open-ended tasks like Chatbot Arena and MT-Bench(Zheng et al., 2023). Prometheus 1 and 2 (Kim et al., 2023, 2024) , fine-tuned evaluators based on LLaMA-2 with customizable evaluation criteria, demonstrate performance matching with state-of-the-art proprietary models. JudgeLM (Zhu et al.) introduces a comprehensive judge training along with addressing known biases (position,format, and knowledge). There exists limited work (Wang et al.; Liu et al., 2023; Hu et al., 2024) that specifically focuses on LLM judges for the evaluation of quality measures that are hard to quantify and prone to confusion. Experimental results reveal that LLMs possess inherent oversensitivity and confusion about quality attributes. This motivates our work on the reliability of evaluations for NLG quality measures.

Robustness in NLG Evaluation A central concern in using LLM judges is calibration of their robustness and reliability: are they consistent while evaluating the content or are they vulnerable to irrelevant factors and minor perturbations? (Badshah and Sajjad, 2025) reports that when an LLM judge is instructed to provide a step-by-step explanation for its decision, its evaluations become more transparent and consistent. Although many studies have reported that state-of-the-art LLMs show high alignment with human evaluators(Kim et al., 2024; Liu et al., 2023; Zheng et al., 2023; Sottana et al., 2023), there are instances in which LLM judges exhibit serious shortcomings (Schroeder and Wood-Doughty, 2024; Shi et al., 2025; Thakur et al., 2025). Despite their promising performance in diverse settings, even powerful LLM judges still exhibit known issues of hallucination (Ye et al., 2023a; Huang et al., 2025), factual errors (Wang et al., 2024; Turpin et al., 2023), and bias(Gallegos et al., 2024). Specifically, recent studies have reported challenges such as position bias(Zheng et al., 2023), verbosity bias(Saito et al., 2023; Zheng et al., 2023), confusing evaluation criteria, focusing more on style compared to factual information(Hu et al., 2024) and self-preference(Wataoka et al.). (Ye et al.) systematically identifies in total 12 types of disinct biases. These studies indicate a significant room for improvement in achieving robust, bias-free evaluations. Moreover, non-adversarial input variations such as reformatting or rephrasing can also affect evaluations. This particular direction of robustness has not been explored in depth.

Evaluation Explainability Although LLM judges can provide a numerical score or a preference, details about *how* and *why* this evaluation is achieved remain opaque. Human evaluators typically consider multiple quality dimensions such as coherence, fluency, and relevance when assessing text.If an LLM judge disagrees with the human evaluator, it is critical to understand from where the disagreement comes from. However, LLM judges do not naturally provide this information, and this is a crucial gap to make the evaluation explainable. When an LLM judge provides a monolithic judgment, users have little insight into whether the model perhaps misjudged the factuality of the text, misinterpreted the prompt, or over-emphasized minor grammatical issues. Recent studies like G-Eval have studied multidimensional evaluation in their framework. (Hsu et al., 2024) presents fine-tuned evaluators, specifically designed to evaluate across quality dimensions including faithfulness, coherence, and completeness. These models not only provide ratings for these metrics but also offer detailed explanation and verifiable citation, thereby enhancing trust in the content. G-Eval(Liu et al., 2023) proposes a foundational framework that uses GPT-4-turbo with

chain-of-thought reasoning to decompose given evaluation criteria into structured steps. (Feng et al., 2025) introduces SaMer, a scenario-aware multidimensional evaluator that is capable of adapting evaluation dimensions based on query context. Unlike fixed-dimension approaches, SaMer automatically identifies and prioritizes relevant evaluation dimensions tailored to specific scenarios, achieving superior performance across diverse evaluation tasks. FLASK (Fine-Grained Language Model Evaluation)(Ye et al., 2024) also focuses on multidimensional evaluation and decomposes coarse-level scoring into skill set-level evaluations for each instruction. This approach enables more interpretable assessment by considering instance-wise skill composition rather than overall preference-based evaluation. Similarly, LMUnit(Saad-Falcon et al., 2024) also decomposes the response quality evaluation into natural language unit criteria tests which are explicit and testable. This framework improves inter-annotator agreement and enables more effective LLM development workflows. (Liu et al., 2024) presents HD-EVAL, a framework based on hierarchical criteria decomposition used to align LLM evaluators with human preferences. The framework decomposes evaluation tasks into finer-grained criteria using iterative alignment training. DeCE (Decomposed Criteria-Based Evaluation) (Yu et al., 2025) demonstrates the effectiveness of decomposed evaluation in expert domains by separating precision and recall in legal QA evaluation. Recent work (Chaudhary et al., 2024) investigates robustness of LLM-based evaluators under perturbations, reporting limited alignment with humans. However, their focus on input perturbations differs from our attribution-based diagnostic approach for interpretation.

Prompt Robustness Works in (Tan et al., 2025; Zeng et al.; Zhao et al., 2025; Singh et al., 2026; Arabzadeh and Clarke, 2025) study robustness to variations in the content being evaluated. For example, the position of the answer, the verbosity of the response, and the format of the generated text. In many of these works, actual evaluation instruction is fixed; only the content to be evaluated changes. In contrast, our work focuses on robustness to variations in evaluation instructions themselves (paraphrasing), while the content under evaluation remains unchanged. Specifically, (Arabzadeh and Clarke, 2025) investigates the prompt sensitivity in IR domain with a focus on relevance judgement with no mitigation strategies and is limited to 3 models. In contrast, we study prompt robustness across 4 NLG tasks and 8 models with mitigation strategies and attribute verifiable taxonomy. JudgeBench in (Tan et al., 2025) evaluates LLM judges capability on objectively verifiable response pairs, while we assess judge reliability under input variations for NLG subjective quality attribute evaluation. (Singh et al., 2026) investigates the temporal robustness of finetuned judges to distributional shifts in generator models, while we study whether LLM judges produce consistent evaluations under semantically equivalent instruction variations. These are entirely different robustness axes. In (Zhao et al., 2025), they study adversarial inputs (model responses) specifically designed to exploit judge vulnerabilities, while we focus on non-adversarial, semantically equivalent prompt variations for evaluation instructions, targeting different components of the evaluation. LLMBar in (Zeng et al.) targets instruction following and evaluates whether LLM evaluators can distinguish instruction-following outputs from incorrect ones. We test the robustness of the judge under paraphrasing. We agree that paraphrasing the evaluation instruction is a subset of judge-input perturbations studied in prior work. Our contribution focuses mainly on subjective evaluation criteria, which can be described in multiple semantically equivalent ways. Evaluation rubric paraphrase sensitivity for subjective NLG criteria is under-explored relative to response-style perturbations.

B Types of Perturbations

Input instructions are broadly categorized into two classes:

1) **Adversarial Perturbations:** These perturbations are deliberately crafted inputs, designed to mislead the LLM judge. This tests whether LLM judges can maintain independent judgment even when presented with contradictory information.

2) **Non-Adversarial Variations:** Natural linguistic variations of input instructions maintaining semantic equivalence are considered under this category. This tests the sensitivity to instruction reformulation- a common occurrence in real-world applications where different users may phrase the evaluation criteria differently.

For example, to evaluate 'Interestingness' of a generated dialogue response, we consider following

paraphrases for the original prompt "Is the response dull/interesting? When evaluating a response, consider how entertaining and interesting it is. Such qualities often boost engagement and enjoyment. Responses should not feel uninspired or boring. Typically, an engaging response includes intriguing or unexpected content.".

1. "To be effective, a response should offer entertainment and spark interest. Responses with these traits usually foster higher user engagement and enjoyment. Refrain from giving plain or uninspired answers. Typically, the most interesting responses include unexpected or little-known details."

2. "Effective responses should be intriguing and fun. These traits often lead to better user engagement and satisfaction. Avoid responses that seem tedious or repetitive. An interesting response usually contains new and unexpected facts."

3. "The best responses are those that are lively and entertaining, which tend to keep users more satisfied and involved. Avoid dull or repetitive replies. Responses that surprise or share lesser-known insights are generally more compelling."

C Data and Code Availability

To facilitate reproducibility, we will release soon: 1) Complete prompt sets: All 120 paraphrased evaluation prompts, 2) Evaluation data: LLM judge responses for all 115,000 evaluations, 3) Analysis scripts: Code for all metrics and statistical tests, and 4) Dataset details: Specific sample IDs and preprocessing steps.

D Mitigation Strategies

Our primary contribution is the diagnostic framework. The mitigation strategies demonstrate that the accuracy-robustness gap is partially addressable. Because our goal is to quantify and reduce prompt-induced instability rather than optimize task-specific accuracy, evaluating mitigation on the same controlled set isolates and validates robustness effects. External out-of-distribution validation is a meaningful extension, and we will add experiments on cross-task transfer or one extra dataset focused on mitigation. Table 3 in the paper provides relative cost indicators. The costing will mainly depend on the budget and API calls.

E Human Study

We provide results from a preliminary human study using paraphrases for NLG evaluations. We consider two setups: 1) Question Generation for 'answerability' attribute, and 2) Essay Evaluation for 'organization' attribute. Each block has an evaluation prompt phrase and 6 samples to evaluate. Common instructions provided in the google form were as follows: Please evaluate each item independently based only on the instruction shown in each block. Do not compare across blocks. Some blocks may appear similar; that is expected. Please answer all items carefully.

Model	Features
GPT-3.5-turbo-0125	baseline-efficiency
GPT-4.1-nano-2025-04-14	smaller-efficient
GPT-4.1-2025-04-14	latest
Claude Sonnet 4.5	strong reasoning
LLaMA-3.1-8b-instant	smaller, efficient
LLaMA-3.3-70b-versatile	larger, capable
LLaMA-4	latest generation
Mistral Magistral Medium	latest,alternative architecture

Table 4: Models considered for robustness experiments. Top four are proprietary models and bottom four are open-source.

Task	Accuracy	Sensitivity	Win-rate Gap
Summ	0.356	0.152	0.146
Dialogue	0.480	0.146	0.172
Question	0.657	0.121	0.156
Essay	0.154	0.162	0.095

Table 5: Task Performance Statistics

Attribute Type	Baseline	Best Strategy	Best Acc. Δ
Factually Verifiable	0.71	PoLL/Fuser	0.89 (+0.18)
Semi-Objective	0.56	PoLL(Tol)	0.73 (+0.17)
Subjective	0.19	Fuser(Tol)	0.54 (+0.35)

Table 6: Aggregate mitigation performance by attribute category. Best strategy varies by attribute verifiability. Even with mitigation, subjective attributes remain below acceptable thresholds.

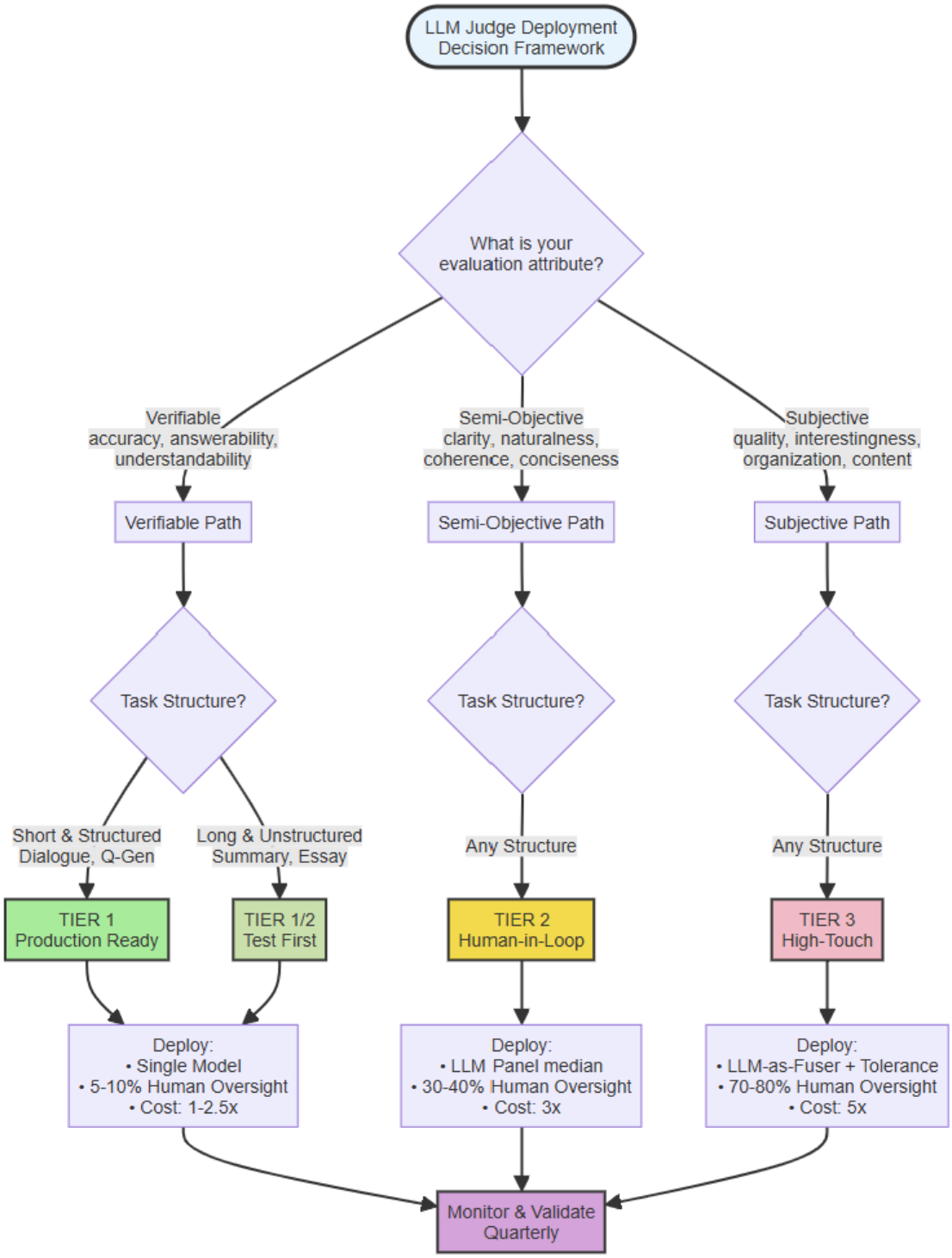

Figure 4: Deployment Ready Reckoner