Diagnosing Prompt Sensitivity and Interaction Instability in LLM-Based Humor Judgment

Dohyeon Kim **Jihoon Yang**
Department of Computer Science and Engineering
Sogang University
{kdh718,yangjh}@sogang.ac.kr

Abstract

Humor judgment is subjective and often depends on ambiguity, expectation violation, and contextual reinterpretation, so the reliability of an LLM judge cannot be assessed from final detection performance alone. Recent work has shown that LLM judgments can vary with prompt formulation and user framing, but how these effects manifest in subjective humor evaluation remains less well characterized, particularly in directional error patterns and multi-turn judgment trajectories. This paper analyzes how prompt design, structured interpretation, and user rebuttals shape LLM-based humor judgment. In zero-shot sentence-only judgments, models with similar F1 scores show substantially different FPR–FNR values, shifting between over-accepting and under-predicting humor depending on prompt formulation. Few-shot examples partially reduce these directional fluctuations, but their effects remain model-dependent. We further propose the Incongruity-Resolution Judgment framework, which compares a humor-supporting interpretation with a counter interpretation before producing a final judgment. Rather than uniformly improving performance, this framework serves as a diagnostic structure for examining how competing interpretations reshape model judgment behavior. Finally, explicit user rebuttals show that high initial performance does not guarantee robustness, highlighting the need to evaluate judgment persistence, directional error patterns, and interaction-level judgment trajectories.

1 Introduction

Although humor evaluation can be operationalized as scores or labels through human annotation, the underlying judgment process remains subjective and shaped by competing interpretations. Humor often involves ambiguity, expectation violation, contextual reinterpretation, and cultural background knowledge, and has therefore long been regarded as a challenging problem in natural language understanding (Raskin, 1985; Attardo, 1994). This subjectivity makes humor judgment sensitive not only to the input sentence itself, but also to the interpretive frame through which the sentence is evaluated.

Prior work on computational humor has studied structurally well-defined forms such as puns using probability-based metrics, including ambiguity, distinctiveness (Kao et al., 2015), and local-global surprisal (He et al., 2019). These metrics capture important aspects of humor, but they do not fully explain human-perceived funniness, making human evaluation central to humor research (Kao et al., 2015; He et al., 2019). Recent LLM-based studies further examine humor detection, generation, preference optimization, and differences between human and LLM criteria for funniness (Wang et al., 2025; Sakabe et al., 2026). However, high performance or fluent explanations do not necessarily imply reliable humor judgment. LLMs may overreact to surface humor cues, such as pun-like patterns, and misclassify modified non-pun sentences as puns (Zangari et al., 2025). This concern is related to broader findings that LLM-based evaluation is sensitive to evaluation prompts, procedures, and user influence (Wang et al., 2024; Perez et al., 2023). Recent work further shows that semantically equivalent prompt formulations can lead to different evaluations, particularly for subjective attributes (Bhat and Varma, 2026). Ermakova et al. (2026) also report confirmation, position, and framing effects in LLM responses under controlled prompt variations.

This paper provides a diagnostic analysis of prompt sensitivity, structured interpretation, and rebuttal robustness in LLM-based humor judgment. We first examine sentence-only prompts with different levels of instructional detail and few-shot examples, using F1 and FPR–FNR to distinguish overall detection performance from di-

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2756–2798
November 6-10, 2026 ©2026 Association for Computational Linguistics

rectional error patterns. We then introduce the Incongruity-Resolution Judgment (IRJ) framework as a structured diagnostic intervention that compares a humor-supporting interpretation with a counter interpretation before producing the final humor judgment. We evaluate both multi-turn and single-prompt implementations with ablations to examine how structured interpretation reshapes model judgments, rather than to present IRJ as a uniformly performance-improving method. Finally, we introduce Rebuttal Stability Evaluation (RSE), inspired by Flip-Flop and multi-turn sycophancy studies, to analyze whether LLM judges maintain their judgments under targeted user rebuttals and follow-up interaction. Overall, our results show that the reliability of LLM-based humor judgment should be assessed not only by final detection performance, but also by directional error patterns, judgment persistence, and interaction-level judgment trajectories.

2 Related Work

Classical humor theories explain humor through conflicts between incompatible semantic structures or scripts, often requiring a shift in interpretation (Raskin, 1985; Attardo, 1994). Building on this view, early computational studies design lexical and semantic features for automatic humor recognition (Mihalcea and Strapparava, 2005), identify humor anchors within sentences (Yang et al., 2015), and model puns using structural cues such as ambiguity and phonological similarity (Kao et al., 2015; He et al., 2019). Recent LLM-based studies extend this line of work to humor data construction, generation, and evaluation: Horvitz et al. (2024) construct paired humor data using LLMs, Chen et al. (2024) improve pun generation through preference learning, and Sakabe et al. (2026); Zangari et al. (2025) analyze the capabilities and limitations of LLMs in humor evaluation. However, these studies primarily focus on detection performance, generation quality, or agreement with human evaluation, leaving the stability of LLM-based humor judgments under different prompt formulations and interaction contexts relatively underexplored.

Recent work on LLM-as-a-Judge shows that large language models can serve as scalable alternatives or complements to human evaluation, as in MT-Bench, Chatbot Arena, and G-Eval (Zheng et al., 2023; Liu et al., 2023). However, LLM judges do not always function as neutral evaluators: their judgments can vary with response order, evaluation procedures, semantically equivalent prompt formulations, and controlled framing manipulations (Wang et al., 2024; Bhat and Varma, 2026; Ermakova et al., 2026). This instability is also related to sycophancy and interaction sensitivity: models can favor user-aligned responses over truthful ones (Sharma et al., 2024), and the Flip-Flop experiment shows that model responses can change after a simple user rebuttal (Laban et al., 2024). We examine how these effects appear in subjective humor judgment, focusing on directional error patterns, structured interpretation, and rebuttal stability.

3 Experimental Setup

3.1 Datasets

In this study, we construct an experimental corpus to analyze prompt sensitivity and directional error patterns in LLM-as-a-Judge humor evaluation. Humorous candidates are collected from SemEval-2017 Task 7 (Miller et al., 2017) and the short-jokes dataset (Moudgil, 2017), while non-humorous candidates are collected from SemEval-2016 Task 4 (Nakov et al., 2016) and the sentiment subset of TweetEval (Barbieri et al., 2020). To reduce domain variation, all non-humorous candidates are drawn from Twitter-based datasets.

From this candidate corpus, we select 600 samples misclassified by humor-no-humor (Dhiab, 2023), an existing encoder-based humor detection model. We refer to this subset as Challenge-600, a controlled evaluation set for analyzing LLM judges on difficult and ambiguous cases, rather than as a general benchmark for average performance.

To examine robustness across different data compositions, we further conduct supplementary experiments on SemEval-2020 Task 7 (Hossain et al., 2020) and SemEval-2021 Task 7 (Meaney et al., 2021). We convert their human humor scores into four-way ordinal labels and filter SemEval-2021 samples with high offensiveness to reduce potential confounding effects from offensive content. Detailed preprocessing, human agreement analysis, and label transformation criteria are provided in Appendix A.

3.2 Models and Prompt Conditions

We evaluate three publicly available instruction-tuned language models: Qwen3-4B-Instruct (Qwen Team, 2025), Llama-3.1-8B-Instruct (Meta, 2024),

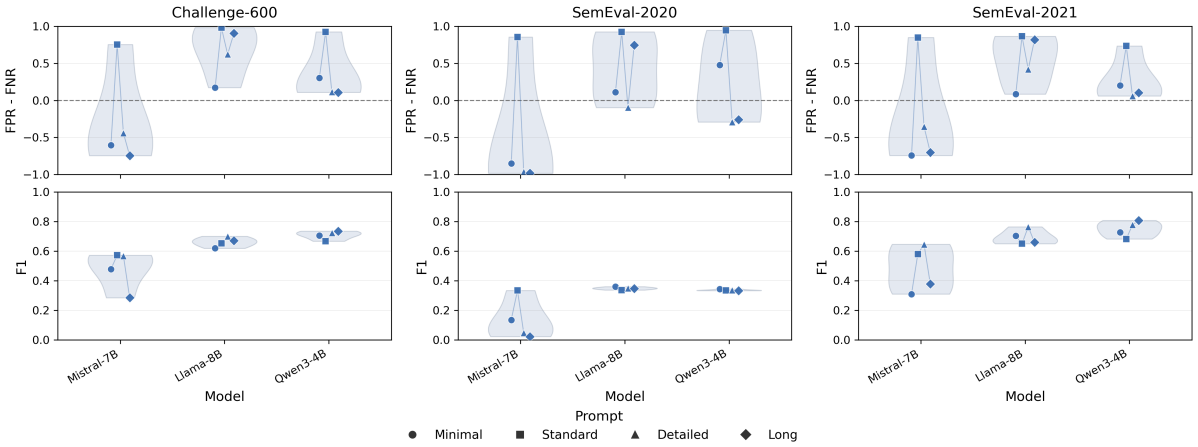


Figure 1: Zero-shot prompt sensitivity in sentence-only humor judgment. The top row shows FPR–FNR, and the bottom row shows F1 across Minimal, Standard, Detailed, and Long prompts.

and Mistral-7B-Instruct-v0.3 (Mistral AI, 2024). To analyze the humor judgment behavior of these models, we design four zero-shot (ZS) sentence-only prompt conditions according to the amount of task guidance provided. The Minimal prompt contains only essential instructions, the Standard prompt provides general guidelines for humor judgment, and the Detailed prompt specifies the definition of humor and explicit judgment criteria. We also include a Long prompt to examine whether more extensive instruction improves performance; this prompt is approximately 2.8 times longer than the Detailed prompt and elaborates the judgment criteria in a more fine-grained manner. The complete prompt templates and detailed generation settings are provided in Appendix B.

3.3 Evaluation Metrics

We evaluate model humor judgments against the reference labels for each dataset. For binary evaluation, we group *weak high* and *strong high* as humorous and *strong low* and *weak low* as not humorous. We measure overall detection performance with F1 for the humorous class and report $\mathrm{FPR} - \mathrm{FNR}$ to characterize error direction. Positive values indicate a higher false positive rate than false negative rate, indicating a tendency to predict humor too often, whereas negative values indicate a higher false negative rate and a tendency to miss humor. $|\mathrm{FPR} - \mathrm{FNR}|$ measures the imbalance between the two error rates. A smaller absolute difference indicates more similar false positive and false negative rates but does not by itself imply better detection performance. We therefore consider F1 together with $\mathrm{FPR} - \mathrm{FNR}$ when comparing overall performance and error direction.

4 Sentence-Only Humor Judgment

This section establishes sentence-only baselines for analyzing prompt sensitivity in LLM-based humor judgment. We compare zero-shot and few-shot prompts with different levels of instructional detail, and report the complete numerical results in Appendix C.

4.1 How Does Prompt Design Shape Directional Error Patterns?

Figure 1 reports F1 and FPR–FNR for the four zero-shot sentence-only prompts. F1 varies relatively modestly across prompts in many settings, whereas FPR–FNR changes substantially even within the same model. Thus, the zero-shot results suggest that prompt wording mainly changes the direction of model errors, rather than only the aggregate performance score.

This pattern is most pronounced in Mistral-7B. Across all three datasets, Mistral-7B shows large prompt-dependent changes in FPR–FNR. The Minimal prompt tends to produce a false-negative-skewed judgment pattern, whereas the Standard prompt shifts the model toward a false-positive-skewed pattern. In other words, the same model can shift from under-predicting to over-predicting humor solely as a function of prompt phrasing.

This phenomenon is not limited to a single model. Across multiple models, the Standard prompt tends to produce relatively high FPR–FNR values, suggesting that general humor-judgment instructions may encourage more permissive humor judgments. The Detailed prompt mitigates this

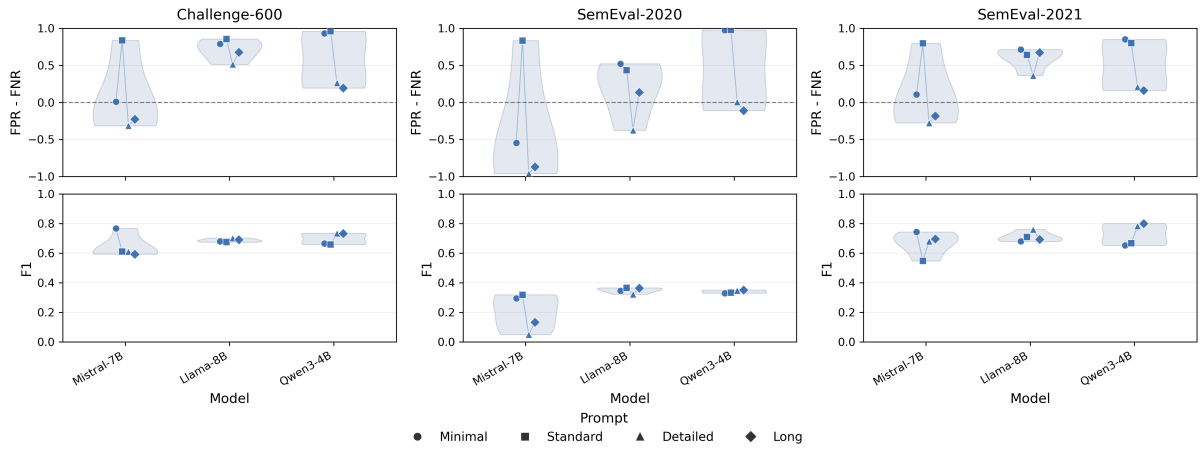


Figure 2: Few-shot prompt sensitivity in sentence-only humor judgment. The top row shows FPR–FNR, and the bottom row shows F1 across Minimal, Standard, Detailed, and Long prompts.

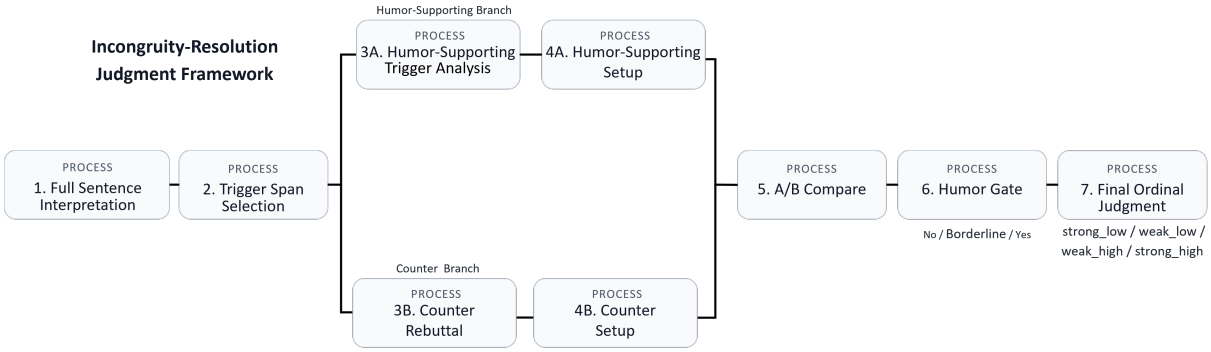


Figure 3: Overview of the IRJ pipeline. The multi-turn version uses separate LLM calls for the structured stages, while Single Prompt compresses the same structured judgment process into a single prompt.

tendency under some conditions, but the effect is not consistent across all models and datasets. The Long prompt further shows that adding more extensive instruction does not necessarily improve sentence-only judgment: it improves or stabilizes performance in some cases, such as Qwen3-4B, but it can also reduce F1 or amplify directional imbalance in other settings. Overall, these results show that sentence-only LLM judges can exhibit substantially different directional error profiles even when final F1 changes only moderately, and that prompt length alone is not a reliable solution to this instability.

4.2 Can In-Context Examples Stabilize Directional Error Patterns?

To examine whether in-context examples can stabilize sentence-only judgment, we evaluate few-shot (FS) variants of all four sentence-only prompts. Figure 2 shows that few-shot examples reduce extreme FPR–FNR fluctuations in some conditions, but their effects are not uniformly beneficial. This suggests that in-context examples recalibrate the model's judgment criterion in a model- and dataset-dependent manner, rather than simply improving sentence-only humor judgment.

This calibration effect is most visible in Mistral-7B. In the zero-shot setting, Mistral-7B exhibits extreme prompt-dependent fluctuations in FPR–FNR. In the few-shot setting, the Minimal prompt becomes substantially more balanced on Challenge-600 and SemEval-2021 while also achieving higher F1 scores. However, the remaining prompts still show dataset-dependent behavior: the Standard prompt often remains false-positive-skewed, whereas the Detailed and Long prompts can shift toward false-negative-skewed judgments.

The same pattern appears for Llama-3.1-8B and Qwen3-4B. For Llama-3.1-8B, the Detailed prompt often provides a relatively strong trade-off between F1 and FPR–FNR, but other prompts still retain positive FPR–FNR values in several settings. For Qwen3-4B, the Detailed and Long prompts are

Dataset	Model	Best ZS		Competing		Single Prompt	
		F1	FPR–FNR	F1	FPR–FNR	F1	FPR–FNR
Challenge-600	Mistral-7B	57.2	75.5	58.4	74.6	69.0↑	69.7†
	Llama-3.1-8B	70.0	62.2	69.0	46.9†	65.6	73.0
	Qwen3-4B	73.5	10.9	71.2	39.5	76.1↑	10.0†
SemEval20	Mistral-7B	33.4	85.6	31.8	84.8	33.9↑	65.4†
	Llama-3.1-8B	36.0	11.2	33.6	-7.2†	34.5	49.5
	Qwen3-4B	34.4	47.9	34.7	29.5	35.9↑	-13.3†
SemEval21	Mistral-7B	64.6	-35.5	59.1	83.2	65.7↑	71.8
	Llama-3.1-8B	76.4	41.9	69.4	44.2	70.5	57.5
	Qwen3-4B	80.7	10.2	76.2	20.5	80.7	3.1†

Table 1: Comparison between the best zero-shot baseline, the competing interpretation pipeline, and the single-prompt pipeline. All values are reported in percentage points, with the percentage symbol omitted. ↑ indicates the larger F1 improvement over Best ZS among the pipeline variants, and † indicates the larger reduction in $|\mathrm{FPR} - \mathrm{FNR}|$ over Best ZS among the pipeline variants.

beneficial in some datasets, but FS prompting does not consistently improve F1 across all prompt conditions, and the Minimal and Standard prompts still produce strong false-positive-skewed patterns. Overall, few-shot examples act as a calibration signal whose effects vary across models and datasets. They do not consistently improve performance or reduce directional imbalance.

5 Evaluating the Incongruity-Resolution Judgment Framework

5.1 Framework Design

The IRJ framework is motivated by incongruity-resolution theory, which views humor comprehension as a process of detecting an incongruity and resolving it within the preceding context (Suls, 1983). From this perspective, humor does not arise merely from an unexpected element itself, but from whether that element can be reinterpreted in a way that produces a coherent humorous effect. This motivates us to examine not only whether an LLM predicts humor, but also how it constructs and compares interpretations that support or challenge that prediction.

Accordingly, we propose the Incongruity-Resolution Judgment (IRJ) framework, as shown in Figure 3. IRJ first obtains a full-sentence interpretation and selects the trigger span, namely the word or phrase that most strongly changes the interpretation of the sentence. It then branches into two competing interpretations: a humor-supporting branch and a Counter Branch. In the humor-supporting branch, the model analyzes how the trigger can create a humorous reading and reconstructs the setup interpretation that makes this reading possible. In the Counter Branch, the model constructs an ordinary or non-humorous counter reading and reconstructs the corresponding setup interpretation. The framework then compares these two interpretations, applies a humor gate to decide whether a recognizable humorous effect remains, and finally produces an ordinal humor judgment.

We implement IRJ in both multi-turn and single-prompt forms. The multi-turn version uses separate LLM calls for the structured stages, whereas Single Prompt compresses the same judgment process into a single prompt and produces the structured judgment fields in one response. The full Single Prompt template is provided in Figure 22. We use these implementations as diagnostic interventions for analyzing whether structured interpretation changes both F1 and directional error patterns. In particular, FPR–FNR allows us to examine whether IRJ shifts the model toward humor over-prediction, toward under-prediction, or toward a more balanced relation between false-positive and false-negative tendencies.

5.2 How Does IRJ Affect Performance and Directional Error Patterns?

Table 1 compares the Best ZS sentence-only baseline with the competing pipeline and the single-prompt implementation of IRJ. Overall, IRJ affects not only F1 but also the direction and magnitude of model errors, as measured by FPR–FNR.

In the primary evaluation, the single-prompt implementation improves F1 for Mistral-7B and shows the most balanced pattern for Qwen3-4B, but in different ways. Mistral-7B achieves higher F1 scores than Best ZS across all three datasets,

(a) Competing Pipeline

Dataset	Model	Full		No Counter		Self-Revision	
		F1	FPR–FNR	F1	FPR–FNR	F1	FPR–FNR
Challenge-600	Mistral-7B	58.4	74.6	65.5	62.1	5.2↓	-97.3↓
	Llama-3.1-8B	69.0	46.9	66.9↓	99.0↓	40.7↓	-60.7↓
	Qwen3-4B	71.2	39.5	66.1↓	93.2↓	49.4↓	-61.0↓
SemEval20	Mistral-7B	31.8	84.8	33.2	47.9	0.3↓	-99.8↓
	Llama-3.1-8B	33.6	-7.2	32.8↓	98.0↓	10.8↓	-88.8↓
	Qwen3-4B	34.7	29.5	32.9↓	94.0↓	17.7↓	-81.7↓
SemEval21	Mistral-7B	59.1	83.2	66.3	58.8	1.6↓	-99.1↓
	Llama-3.1-8B	69.4	44.2	63.0↓	93.8↓	48.0↓	-57.9↓
	Qwen3-4B	76.2	20.5	63.6↓	91.1↓	48.2↓	-60.3↓

(b) Single Prompt Pipeline

Dataset	Model	Full		No Counter		Self-Revision	
		F1	FPR–FNR	F1	FPR–FNR	F1	FPR–FNR
Challenge-600	Mistral-7B	69.0	69.7	58.5↓	-30.3	68.2↓	66.0
	Llama-3.1-8B	65.6	73.0	67.0	97.7↓	21.4↓	-78.3↓
	Qwen3-4B	76.1	10.0	58.9↓	-45.0↓	43.4↓	-69.3↓
SemEval20	Mistral-7B	33.9	65.4	21.4↓	-64.3	34.2	53.2
	Llama-3.1-8B	34.5	49.5	32.8↓	99.3↓	7.3↓	-93.1↓
	Qwen3-4B	35.9	-13.3	17.7↓	-81.1↓	4.7↓	-96.3↓
SemEval21	Mistral-7B	65.7	71.8	52.1↓	-35.6	67.0	68.6
	Llama-3.1-8B	70.5	57.5	63.1↓	94.2↓	25.7↓	-80.4↓
	Qwen3-4B	80.7	3.1	63.6↓	-42.7↓	50.0↓	-62.9↓

Table 2: Ablation results for the competing and single-prompt pipelines. *No Counter* removes the Counter Branch, while *Self-Revision* replaces the independent adjudication stage with branch-side self-revision. Arrows indicate degradation relative to the corresponding full pipeline. For F1, degradation indicates a lower score; for FPR–FNR, degradation indicates a larger absolute distance from zero.

although its directional error imbalance worsens on SemEval-2021. In contrast, Qwen3-4B improves or matches Best ZS in F1 while reducing the absolute magnitude of FPR–FNR across all three datasets. This suggests that, for some models, integrating structured interpretation into a single prompt can improve performance while moving the error pattern closer to balance.

Llama-3.1-8B shows a different pattern. Under the single-prompt implementation, F1 decreases and FPR–FNR increases across all datasets, indicating stronger humor over-prediction. Under the competing pipeline, however, the absolute FPR–FNR decreases on Challenge-600 and SemEval-2020 even when F1 does not improve. This suggests that explicitly separating humor-supporting and counter interpretations can affect directional error patterns independently of performance gains.

Taken together, these results indicate that the effect of IRJ is model-dependent: it sometimes improves F1, sometimes reduces directional imbalance, and sometimes shifts the model toward stronger humor over-prediction. To further assess decoding stability, we repeat the primary IRJ evaluations with multiple random seeds and report the results in Appendix C.4.

5.3 Which Components of IRJ Are Most Important?

Table 2 reports the ablation results for two key design choices in IRJ: the Counter Branch and the independent adjudication stage. The *No Counter* variant removes the counter-side interpretation stages and makes the final decision from the humor-supporting path alone. The *Self-Revision* variant keeps the counter interpretation, but replaces independent adjudication with branch-side self-revision, where the initial humor-supporting interpreter revises its own judgment after considering the counter interpretation. Overall, the two variants affect performance and directional error patterns in distinct ways.

The effect of *No Counter* is not primarily reflected in large F1 degradation. Removing the

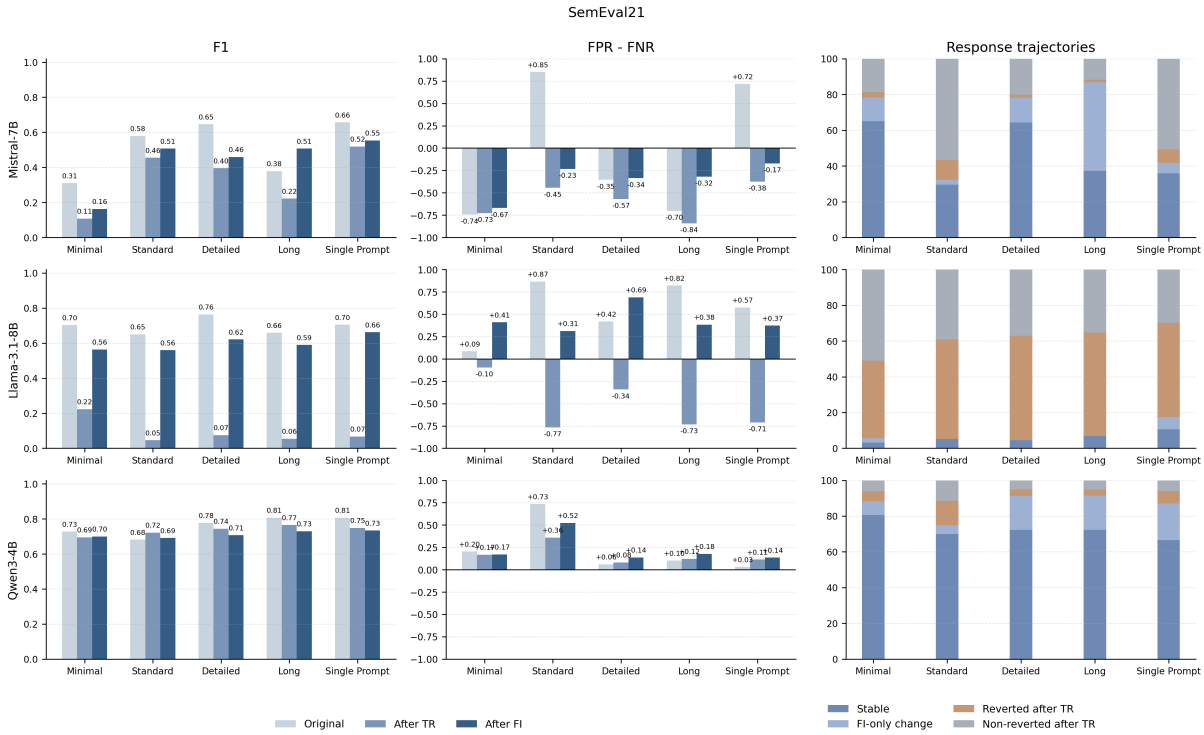


Figure 4: RSE results on SemEval-2021. We report F1, FPR–FNR, and response trajectories across Original, After TR, and After FI. The trajectory panel is computed from four-way ordinal humor labels.

Counter Branch does not always reduce F1, and in some competing-pipeline settings it even improves performance. However, it often pushes FPR–FNR toward a stronger one-sided error pattern, especially in the competing pipeline, where the absence of an explicit counter interpretation frequently produces large positive FPR–FNR values. This suggests that accumulating only humor-supporting evidence can make the model increasingly prone to over-predicting humor.

The *Self-Revision* variant shows a different pattern. Unlike *No Counter*, it often causes substantial F1 degradation, especially for Llama-3.1-8B and Qwen3-4B, and shifts FPR–FNR far into the negative range. This indicates that self-revision tends to move the model toward false-negative-skewed judgments rather than producing a balanced correction. Mistral-7B is a partial exception in the single-prompt implementation, where *Self-Revision* remains closer to the full pipeline, but its competing-pipeline results still show a severe collapse in both F1 and directional error balance.

The Counter Branch provides an explicit competing perspective against the humor-supporting interpretation and helps prevent one-sided humor over-prediction, whereas independent adjudication is more directly tied to final decision quality. These results indicate distinct roles for competing interpretations and independent adjudication in shaping the final judgment.

6 Robustness under User Rebuttal

After evaluating the initial judgments produced by sentence-only prompts and the single-prompt implementation of IRJ, we examine whether these judgments remain stable under explicit user rebuttals. Inspired by the Flip-Flop evaluation of Laban et al. (2024) and multi-turn sycophancy studies (Hong et al., 2025), we introduce Rebuttal Stability Evaluation (RSE), a task-specific setting for analyzing judgment persistence and trajectory-level instability in subjective humor judgment. RSE consists of Targeted Rebuttal (TR), where the user challenges the initial judgment in the opposite direction, and Follow-up Interaction (FI), where the user partially retracts the rebuttal if the model changes its judgment, or continues to disagree if the model maintains it. We track changes from Original to After TR and After FI in terms of F1, FPR–FNR, and response trajectories.

Figure 4 shows the RSE results on SemEval-2021. Overall, sentence-only baselines are vulnerable to targeted rebuttal. Llama-3.1-8B shows a particularly large drop: under the Detailed prompt,

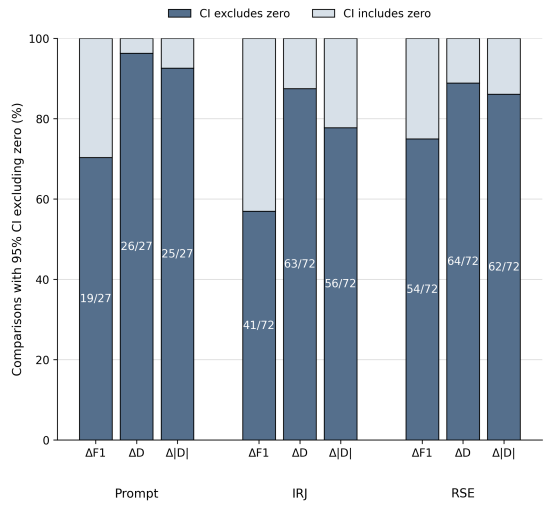


Figure 5: Summary of paired bootstrap validation results. Each bar reports the percentage of paired comparisons whose 95% bootstrap confidence interval excludes zero. Prompt, IRJ, and RSE aggregate 27, 72, and 72 paired comparisons, respectively.

F1 decreases from 0.76 to 0.07 after TR, and under the Long prompt it similarly decreases from 0.66 to 0.06. This indicates that longer instruction does not eliminate rebuttal-induced instability. Mistral-7B also degrades after TR across several sentence-only prompts, whereas Single Prompt mitigates this decline in some conditions. For example, with Mistral-7B, F1 decreases from 0.66 to 0.52 after TR and remains at 0.55 after FI. Qwen3-4B shows the most stable F1 trajectory overall.

FPR–FNR shows that rebuttals alter not only performance but also the direction of model errors. Mistral-7B and Llama-3.1-8B often move from positive or moderate FPR–FNR values in the Original setting to negative values after TR, indicating a shift toward false-negative-skewed judgments. After FI, FPR–FNR often moves back in the opposite direction, suggesting that FI is not a simple recovery but another interaction-induced judgment shift. Qwen3-4B shows relatively smaller FPR–FNR changes, especially under the Detailed, Long, and Single Prompt conditions.

Let y_O, y_{TR}, and y_{FI} denote the four-way ordinal labels at Original, after Targeted Rebuttal, and after Follow-up Interaction, respectively. The response trajectory panel groups these sequences into *Stable* ($y_O = y_{TR} = y_{FI}$), *FI-only change* ($y_O = y_{TR}$ and $y_{FI} \neq y_O$), *Reverted after TR* ($y_O \neq y_{TR}$ and $y_{FI} = y_O$), and *Non-reverted after TR* ($y_O \neq y_{TR}$ and $y_{FI} \neq y_O$). The last category includes both retention of the TR label and a subsequent change to a third label. These categories show that interaction effects cannot be reduced to simple label flips: *Reverted after TR* may resemble recovery but still reflects trajectory-level instability, whereas a high *Stable* rate is not necessarily desirable when final F1 remains low.

In summary, RSE shows that user rebuttals can alter both final performance and error direction. Single Prompt does not eliminate rebuttal sensitivity, but it mitigates F1 degradation and extreme FPR–FNR shifts in some model conditions. Similar trends are observed on the remaining datasets, and detailed results are reported in Appendix C.

7 Bootstrap Validation of Observed Shifts

To examine whether the observed changes are robust to sampling variation, we perform an instance-level paired bootstrap analysis. For each model–dataset–condition comparison with n evaluation instances, we generate 10,000 bootstrap replicates by sampling n instance indices with replacement from the corresponding evaluation set. For each replicate, metric differences are recomputed using the same resampled indices for both conditions, preserving the paired structure of the comparison. We denote $\mathrm{FPR} - \mathrm{FNR}$ as D. This analysis quantifies uncertainty in metric differences due to evaluation sample composition. Decoding variability is assessed separately by repeating the evaluation with multiple random seeds.

Figure 5 summarizes the proportion of comparisons whose 95% paired bootstrap confidence intervals exclude zero. The denominators differ because each analysis group aggregates a different number of paired comparisons: 27 for prompt comparisons, 72 for IRJ comparisons, and 72 for RSE comparisons. Overall, a large proportion of changes are supported under paired resampling, especially for D and $|D|$. This indicates that the observed shifts in prompt sensitivity, IRJ behavior, and rebuttal-induced judgment trajectories are not driven by a small number of evaluation instances. Detailed confidence intervals and the directions of supported changes are provided in Appendix D.

8 Conclusion

This paper analyzes LLM-based humor judgment as a subjective evaluation problem whose reliability depends on prompt formulation, structured interpretation, and interaction context. Across sentence-only prompts, models with similar F1 scores can

produce substantially different FPR–FNR patterns, indicating that prompt wording can shift the judge toward humor over-prediction or under-prediction. Few-shot examples partially reduce these shifts in some settings, but their effects remain model- and dataset-dependent.

We further examine structured interpretation through IRJ and interaction robustness through RSE. IRJ does not serve as a uniformly performance-improving method; rather, it provides a diagnostic structure for observing how competing interpretations and adjudication stages reshape model judgments. RSE shows that explicit user rebuttals can alter both performance and error direction, and that apparent recovery during follow-up interaction may reflect another interaction-induced shift rather than stable judgment maintenance.

These findings suggest several practical safeguards for using LLMs in subjective evaluation. Reliability should be assessed across multiple prompt formulations because a single prompt may not reveal how consistently the model maintains its judgments. FPR–FNR should also be reported alongside F1 to distinguish overall performance from the direction of model errors. Evaluations involving user interaction should examine both the final label and the full trajectory from the initial judgment through Targeted Rebuttal and Follow-up Interaction. The present study does not validate these practices as mitigation methods. They should be viewed as safeguards for identifying instability in model judgments and as directions for future mitigation.

Limitations

This study has several scope limitations. First, IRJ does not yield consistent performance gains across all models and datasets. Rather than maximizing the performance of a particular model, this work aims to diagnose how prompt formulation, structured interpretation, and user rebuttals shape directional error patterns and judgment stability in LLM-based humor judgment. Second, Challenge-600 is a hard-case stress-test set constructed from samples misclassified by an encoder-based humor detector, not a general-purpose benchmark for average performance. Its intentionally difficult and ambiguous construction requires caution when interpreting metric differences. Finally, RSE uses fixed rebuttal and follow-up templates, and the experiments are limited to three publicly available instruction-tuned LLMs. Thus, our findings should be interpreted as evidence for the observed judgment shifts in the evaluated settings, rather than as fully generalizable conclusions about all LLM judges.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00416666).

References

Salvatore Attardo. 1994. *Linguistic Theories of Humor*, volume 1 of *Humor Research*. Walter de Gruyter, Berlin.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Savita Bhat and Vasudeva Varma. 2026. All prompts are created equal? evaluating robustness of LLM judges against non-adversarial prompt variations. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 38730–38745, San Diego, California, United States. Association for Computational Linguistics.

Yang Chen, Chong Yang, Tu Hu, Xinhao Chen, Man Lan, Li Cai, Xinlin Zhuang, Xuan Lin, Xin Lu, and Aimin Zhou. 2024. Are U a joke master? pun generation via multi-stage curriculum learning towards a humor LLM. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 878–890, Bangkok, Thailand. Association for Computational Linguistics.

Mohamed Amine Dhiab. 2023. humor-no-humor. `https://huggingface.co/mohameddhiab/humor-no-humor`. Hugging Face model card, accessed May 11, 2026.

Liana Ermakova, Anton Firsov, and Jaap Kamps. 2026. Confirmation, framing, and position biases in LLM responses. In *Proceedings of the 2026 Conference on Human Information Interaction and Retrieval*, pages 480–484. Association for Computing Machinery.

He He, Nanyun Peng, and Percy Liang. 2019. Pun generation with surprise. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1734–1744, Minneapolis, Minnesota. Association for Computational Linguistics.

Jiseung Hong, Grace Byun, Seungone Kim, and Kai Shu. 2025. Measuring sycophancy of language models in multi-turn dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2239–2259, Suzhou, China. Association for Computational Linguistics.

Zachary Horvitz, Jingru Chen, Rahul Aditya, Harshvardhan Srivastava, Robert West, Zhou Yu, and Kathleen McKeown. 2024. Getting serious about humor: Crafting humor datasets with unfunny large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 855–869, Bangkok, Thailand. Association for Computational Linguistics.

Nabil Hossain, John Krumm, Michael Gamon, and Henry Kautz. 2020. SemEval-2020 task 7: Assessing humor in edited news headlines. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 746–758, Barcelona (online). International Committee for Computational Linguistics.

Justine Kao, Roger Levy, and Noah Goodman. 2015. A computational model of linguistic humor in puns. *Cognitive science*, 40.

Philippe Laban, Lidiya Murakhovs'ka, Caiming Xiong, and Chien-Sheng Wu. 2024. Are you sure? challenging llms leads to performance drops in the flipflop experiment. *Preprint*, arXiv:2311.08596.

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.

J. A. Meaney, Steven Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. SemEval 2021 task 7: HaHackathon, detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 105–119, Online. Association for Computational Linguistics.

Meta. 2024. Llama-3.1-8B-Instruct. `https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`. Hugging Face model card.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 531–538, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Tristan Miller, Christian Hempelmann, and Iryna Gurevych. 2017. SemEval-2017 task 7: Detection and interpretation of English puns. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 58–68, Vancouver, Canada. Association for Computational Linguistics.

Mistral AI. 2024. Mistral-7B-Instruct-v0.3. `https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3`. Hugging Face model card.

Abhinav Moudgil. 2017. Short jokes. `https://www.kaggle.com/datasets/abhinavmoudgil95/short-jokes`. Kaggle dataset.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–18, San Diego, California. Association for Computational Linguistics.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, and 44 others. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada. Association for Computational Linguistics.

Qwen Team. 2025. Qwen3-4B-Instruct. `https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507`. Hugging Face model card.

Victor Raskin. 1985. *Semantic Mechanisms of Humor*, volume 24 of *Synthese Language Library*. Springer, Dordrecht.

Ritsu Sakabe, Hwichan Kim, Tosho Hirasawa, and Mamoru Komachi. 2026. Assessing the capabilities of llms in humor: A multi-dimensional analysis of oogiri generation and evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 32867–32874.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*.

Jerry Suls. 1983. *Cognitive Processes in Humor Appreciation*, pages 39–57. Springer New York, New York, NY.

Han Wang, Yilin Zhao, Dian Li, Xiaohan Wang, sinbadliu, Xuguang Lan, and Hui Wang. 2025. Innovative thinking, infinite humor: Humor research of large language models through structured thought leaps. In *The Thirteenth International Conference on Learning Representations*.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. Humor recognition and humor anchor extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2367–2376, Lisbon, Portugal. Association for Computational Linguistics.

Alessandro Zangari, Matteo Marcuzzo, Andrea Albarelli, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2025. Pun unintended: LLMs and the illusion of humor understanding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27936–27971, Suzhou, China. Association for Computational Linguistics.

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 *Advances in Neural Information Processing Systems*, volume 36.

A Dataset Details

A.1 Construction of the Challenge-600 and Human Evaluation

To construct the Challenge-600, we first collected candidate humorous and non-humorous sentences from the public datasets described in Section 3. We then applied humor-no-humor (Dhiab, 2023), an existing encoder-based humor detection model, to the candidate corpus and selected 600 samples that were misclassified by the model. This filtering process was not intended to construct a dataset reflecting the average humor/non-humor distribution. Instead, it was designed to create a controlled evaluation set for analyzing LLM judges on cases where an existing encoder-based detector does not produce reliable distinctions.

We conducted human evaluation to assign humor labels and humor type labels to the Challenge-600. The evaluation interface is shown in Figure 6. The humor labels consisted of four categories: strong low, weak low, weak high, and strong high. When strong low or weak low was selected, the sentence was treated as non-humorous, and its humor type was automatically set to Other / Not sure. To align the evaluation setting with SemEval-2020, each sample was evaluated by five annotators. All annotators were informed in advance of the evaluation procedure and the purpose of data use, participated voluntarily, and received appropriate compensation considering the time and effort required. To obtain the four-way reference label, we first determined the low or high direction by majority vote over the five annotations. We assigned a strong label when at least four annotators voted in that direction and at least two selected the corresponding extreme label. Otherwise, we assigned the weak label in the majority direction.

Dataset	Total	Strong L.	Weak L.	Weak H.	Strong H.
Challenge-600	600	182	128	192	98
SemEval-2020	3,024	1,532	901	482	109
SemEval-2021	5,737	2,974	201	2,243	319

Table 3: Label distributions after preprocessing.

Krippendorff's α for the four humor labels was 0.124. When the four labels were collapsed into binary low/high judgments, Krippendorff's α increased to 0.358. This low agreement is consistent with the subjectivity of humor judgment and differences in interpretation. In particular, 179 out of 600 samples, corresponding to 29.8%, showed a 3–2 split, where three annotators and two annotators made opposite binary judgments. When these cases were excluded only for the agreement analysis, Krippendorff's α for the binary labels increased to 0.596. This result shows that the Challenge-600 contains many cases with substantial human disagreement, rather than consisting only of clear-cut humorous or non-humorous examples.

A.2 Preprocessing of the Supplementary Datasets

Table 3 summarizes the label distributions after preprocessing. For all datasets, we use the same four-way ordinal humor labels: strong low, weak low, weak high, and strong high. For binary evaluation, weak high and strong high are mapped to the positive class, while strong low and weak low are mapped to the negative class.

For SemEval-2020 Task 7, we reconstruct labels from the distribution of grades assigned by individual annotators rather than from the `meanGrade`

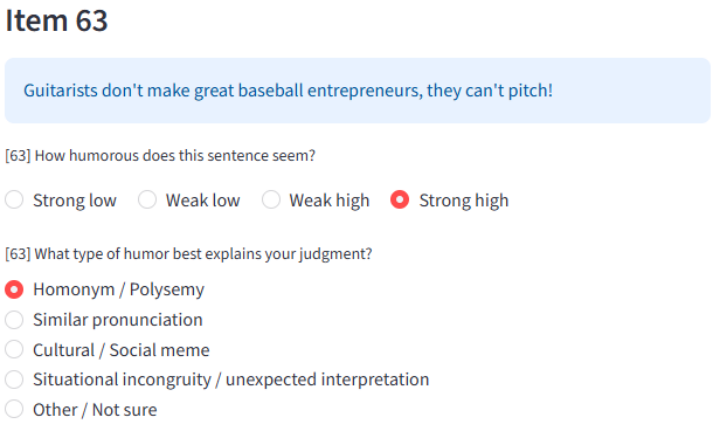


Figure 6: Interface used for human evaluation of the Challenge-600. Annotators selected one of four ordinal humor labels and, when the sentence was judged as humorous, selected the humor type that best explained their judgment.

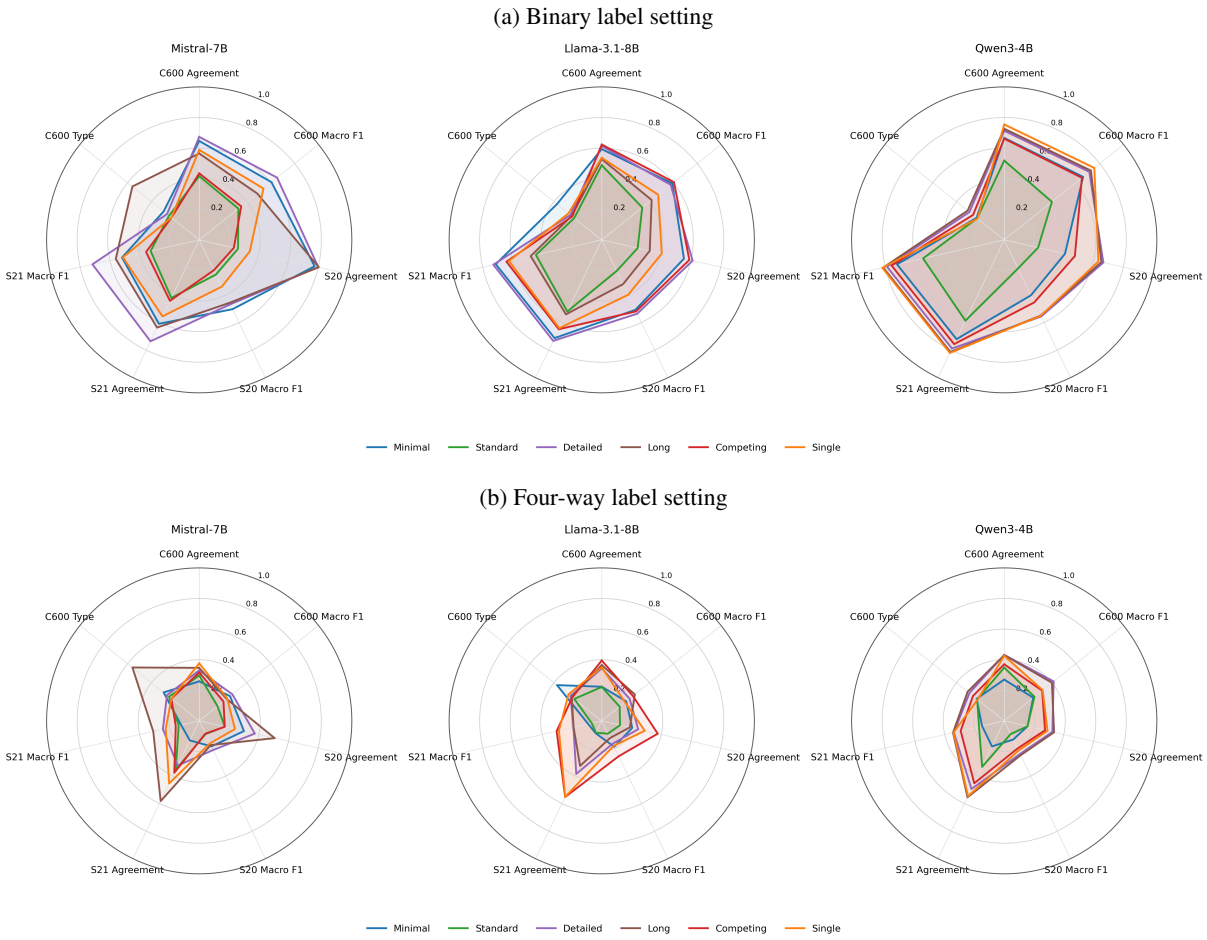


Figure 7: Human agreement analysis across label granularities. The radar plots compare sentence-only prompts, the competing pipeline, and the single-prompt implementation using agreement and macro F1 across datasets, with humor-type agreement reported for Challenge-600.

value. Each item contains humor grades ranging from 0 to 3. We first assign strong low and strong high to items with strong consensus on grade 0 and grade 3, respectively. For the remaining items, we define grades 2 and 3 as positive votes and grades 0 and 1 as negative votes, and assign weak or strong labels according to the direction and strength of the vote distribution. When extreme votes cancel each other out or the annotations are highly dispersed, we also consider the distribution of intermediate grades and assign weak low when there is no clear positive majority. After preprocessing, the SemEval-2020 test set contains 3,024 items.

For SemEval-2021 Task 7, we use the training split because it includes human evaluation scores. To reduce the influence of offensive content, we retain only samples with an offensiveness score of 0.5 or lower. Missing values in humor-related and offensiveness score fields are replaced with 0.0. We then convert humor ratings into four-way ordinal labels using fixed thresholds: strong low for scores of 0.5 or lower, weak low for scores greater than 0.5 and up to 1.5, weak high for scores greater than 1.5 and up to 3.0, and strong high for scores greater than 3.0. After preprocessing and offensiveness filtering, the SemEval-2021 dataset contains 5,737 items.

A.3 Human Agreement Analysis

We further analyze the extent to which model judgments align with human evaluations. Figure 7 compares sentence-only prompts, the Competing Pipeline, and Single Prompt across the three models. The sentence-only prompts include Minimal, Standard, Detailed, and Long. We report agreement and macro F1 across datasets for each condition, and additionally report humor-type agreement for Challenge-600, where human humor-type annotations are available. Panel (a) shows the binary setting, where the four ordinal labels are collapsed into low/high humor judgments, whereas Panel (b) shows the original four-way ordinal-label setting.

Model	Condition	High-agreement		Low-agreement	
		F1	FPR–FNR	F1	FPR–FNR
Mistral-7B	Minimal	54.3	-56.4	33.3	-67.3
	Standard	53.1	69.6	65.9	88.2
	Detailed	64.1	-40.5	40.8	-49.3
	Long	35.3	-72.5	13.8	-77.2
	Competing	55.7	71.5	64.3	81.6
	Single	68.0	65.3	71.0	81.4
Llama-3.1-8B	Minimal	65.5	19.2	54.0	16.5
	Standard	63.5	98.2	69.3	99.0
	Detailed	71.5	60.3	66.9	69.6
	Long	65.5	88.5	70.6	96.4
	Competing	69.5	48.3	67.8	45.3
	Single	65.5	70.3	65.9	81.3
Qwen3-4B	Minimal	73.2	30.7	64.8	33.2
	Standard	65.2	91.2	69.9	96.5
	Detailed	76.1	10.6	64.4	17.1
	Long	77.9	9.9	64.4	17.1
	Competing	72.3	34.3	68.9	54.3
	Single	79.9	6.5	68.6	22.2

Table 4: Agreement-subset analysis on Challenge-600. High-agreement denotes samples with 5–0 or 4–1 binary human votes, while Low-agreement denotes samples with 3–2 binary human votes. All values are reported in percentage points, with the percentage symbol omitted. FPR–FNR indicates the direction of prediction errors.

Overall, model–human agreement is higher in the binary setting. This indicates that models can align with human evaluations to some extent when humor is evaluated as a coarse-grained judgment of whether humor is present. In particular, Qwen3-4B shows relatively high agreement and macro F1 across several conditions, and Detailed and Single Prompt often yield more balanced profiles than the other conditions. Mistral-7B and Llama-3.1-8B also show high agreement under specific conditions, but the pattern varies across datasets and prompt conditions. Thus, structured interpretation and more specific judgment criteria can improve alignment with human evaluations in some cases, but no single condition is consistently superior across all models and evaluation axes.

The results for the Long prompt are consistent with this interpretation. Although Long achieves high scores on some axes, it does not consistently stabilize agreement or macro F1 across models and datasets. In other words, simply providing longer instructions does not automatically improve alignment with human evaluations. This suggests that how the judgment criteria are presented and structured may be more important than prompt length alone.

In contrast, agreement and macro F1 are substantially lower in the four-way setting. This suggests that models have greater difficulty distinguishing fine-grained humor strength, such as strong low, weak low, weak high, and strong high, than making a binary judgment about the presence of humor. Conditions that show relatively high scores in the binary setting exhibit compressed scores in the four-way setting, and the differences between conditions also become more limited. These results indicate that ordinal humor labeling is not merely a more difficult classification setting, but a subjective judgment space in which interpretive differences can arise even among human evaluators. This pattern is consistent with the low inter-annotator agreement observed in Challenge-600.

Humor-type agreement is also generally low. This implies that even when a model reaches a binary humor decision similar to human judgments, it may rely on a different interpretation of the underlying humor mechanism. For example, the same sentence may be judged as humorous but attributed to different mechanisms, such as homonymy/polysemy, cultural or social memes, or situational incongruity, depending on the evaluator. Therefore, the reliability of LLM-based humor judgment cannot be assessed solely through binary detection performance; label granularity and humor-type interpretation must also be considered.

In summary, Figure 7 shows that LLM judges can align with human evaluations to some extent in coarse-grained humor detection, but fine-grained

judgments of humor strength and humor mechanism remain more challenging. Detailed and Single Prompt often produce relatively balanced agreement profiles, but no prompt condition, including Long, is consistently superior across all evaluation axes. This supports the main argument of this paper: LLM-based humor judgment should be analyzed not only in terms of final F1, but also in terms of label granularity, interpretation-level agreement, and directional error patterns.

A.4 Agreement-Subset Analysis

To examine whether the prompt sensitivity and rebuttal instability observed on Challenge-600 are driven only by samples with divergent human judgments, we divide the dataset into two subsets based on binary human agreement. The high-agreement subset consists of samples with 5–0 or 4–1 binary votes, whereas the low-agreement subset consists of samples with 3–2 binary votes. This distinction allows us to determine whether shifts in model judgment are concentrated in cases that human annotators also find ambiguous, or whether they also appear in cases with relatively consistent human evaluations.

Table 4 reports the results for the sentence-only prompts and IRJ variants on the high-agreement and low-agreement subsets of Challenge-600. The results show that prompt-induced shifts in judgment direction are not confined to the low-agreement subset. Even in the high-agreement subset, the Standard prompt produces large positive FPR–FNR values across all three models, indicating a false-positive-skewed judgment pattern even for samples on which human annotators largely agree about the binary humor label. In contrast, the Minimal and Detailed prompts generally produce lower FPR–FNR values, and the same tendency is also observed in the low-agreement subset.

The Long prompt further shows that increasing instruction length does not remove this prompt-dependent behavior. For Mistral-7B, Long produces a strongly negative FPR–FNR pattern in both subsets, whereas for Llama-3.1-8B it produces a strongly positive FPR–FNR pattern. For Qwen3-4B, Long remains closer to Detailed, with relatively small FPR–FNR values in both subsets. Thus, the effect of longer instruction is model-dependent and does not provide a uniform correction to directional error imbalance.

These results suggest that the prompt sensitivity observed on Challenge-600 cannot be explained solely by disagreement among human annotators in the 3–2 split samples. In other words, prompt formulation can alter the error direction of the model even for samples with relatively clear human evaluations. Therefore, the FPR–FNR shifts reported in this study should be interpreted as systematic directional shifts associated with prompt design, rather than as an artifact confined to low-agreement samples.

A similar tendency is observed in RSE. Llama-3.1-8B shows flip rates of 81.2% and 86.0% in the high-agreement and low-agreement subsets, respectively, while Mistral-7B shows flip rates of 52.7% and 50.8%. In contrast, Qwen3-4B maintains low flip rates of 8.3% and 11.2% in the two subsets. These results indicate that rebuttal-induced judgment shifts cannot be explained solely by divergent human judgments, but also reflect model-specific differences in interaction-level stability.

A.5 Humor-Type Low-Agreement Analysis

Section A.4 divides Challenge-600 into high-agreement and low-agreement subsets based on binary humor labels, showing that the prompt sensitivity and rebuttal instability observed on Challenge-600 are not driven solely by 3–2 split samples. In this section, we extend the agreement-based analysis beyond binary humor judgment by examining samples with low agreement in humor-type interpretation. Here, low agreement refers not to disagreement about whether a sentence is humorous, but to disagreement about which humor mechanism best explains the sentence.

For each sample in Challenge-600, we compute the plurality ratio of the humor-type labels assigned by the five annotators, defined as the proportion of annotators selecting the most frequent humor type. We define a humor-type low-agreement item as a sample for which no single humor type receives majority agreement. This criterion yields 96 samples, most of which have a plurality ratio of 0.4. In other words, even the most frequently selected humor type is chosen by only two out of five annotators. Thus, this subset captures disagreement among human annotators in humor mechanism interpretation, rather than disagreement about the binary humor label.

The plurality humor-type distribution of this subset consists of Cultural / Social meme (34), Situational incongruity / unexpected interpretation (29), Homonym / Polysemy (17), Other / Not sure (10), and Similar pronunciation (6).

(a) IRJ Corrects Sentence-only Suppression

Sentence: "Who was the knight that invented the round table? Sir Cumference."
Reference: weak_high
Best ZS: × strong_low
Competing IRJ: ✓ weak_high
Single IRJ: ✓ weak_high
Observation: IRJ preserves the pun-based reinterpretation as an actual humorous payoff, while the sentence-only baseline collapses the sentence into a non-humorous judgment.

(b) IRJ Over-predicts Humor

Sentence: "Tiger Woods' real first name is Eldrick."
Reference: strong_low
Best ZS: ✓ weak_low
Competing IRJ: ✓ strong_low
Single IRJ: × weak_high
Observation: Single-prompt IRJ over-reads a plain factual statement as humor by treating weak lexical novelty as a humorous payoff.

(c) IRJ Preserves Everyday Humor

Sentence: "No thermostat is as effective at regulating temperature as sticking a foot out from under the bed covers."
Reference: strong_high
Best ZS: ✓ weak_high
Competing IRJ: ✓ weak_high
Single IRJ: ✓ weak_high
Observation: Both IRJ variants keep the everyday incongruity as humor, showing that structured interpretation does not necessarily suppress simple experiential jokes.

(d) Rebuttal-induced Judgment Shift

Sentence: "The chemicals heating up boils down to the scientist turning on the furnace."
Reference: weak_high
Best ZS: × weak_low
Single IRJ: ✓ weak_high
RSE: weak_low → weak_high → weak_low
Observation: The model fully oscillates under rebuttal and follow-up interaction, suggesting that user feedback can redirect the interpretation of a short pun-like trigger.

Figure 8: Representative qualitative cases illustrating how IRJ changes humor judgments and how rebuttal interaction can alter the judgment trajectory.

The humor-type low-agreement subset is heavily skewed toward positive binary labels, with 94 positive samples and only 2 negative samples. Therefore, high F1 scores on this subset should not be interpreted as direct evidence of better understanding of difficult cases. Since most samples are positive under the binary label, high F1 may instead reflect a permissive tendency to classify interpretation-disputed cases as humorous. In addition, FPR-based quantities are unstable on this subset because the number of negative samples is extremely small. We therefore use this subset as a supplementary descriptive analysis of judgment direction, rather than as a direct comparison of methodological superiority.

From this perspective, the high F1 scores of the Standard baseline likely reflect a permissive judgment tendency toward interpretation-disputed cases, rather than better understanding of humor-type low-agreement samples. Conversely, lower F1 scores under the Minimal or Detailed prompts may indicate stricter humor criteria, where some positive-skewed samples are judged as non-humorous or only weakly humorous. The Long prompt shows a similarly model-dependent pattern: for some models, additional instruction makes the judgment more permissive, whereas for others it produces stricter decisions and lower F1 on this positive-skewed subset. Thus, longer instruction does not provide a uniform correction for humor-type disagreement. The Competing Pipeline and Single Prompt conditions also change the observed error direction in model-specific ways, suggesting that prompt structure can alter the decision criterion even within the same subset of samples with disputed humor mechanisms.

Together, the agreement-subset analysis in Section A.4 and the humor-type low-agreement analysis in this section provide complementary views of human disagreement in Challenge-600. Section A.4 shows that prompt-induced directional shifts are not driven solely by binary-label ambiguity, whereas this section shows that prompt and pipeline structures can also change model judgment patterns when human annotators disagree about the underlying humor mechanism.

A.6 Qualitative Case Analysis

To complement the quantitative analysis, Figure 8 presents representative qualitative examples illustrating how IRJ affects humor judgments and interpretation trajectories. The cases are selected to cover four distinct patterns: correction of sentence-only suppression, humor over-prediction, preservation of everyday humor, and rebuttal-induced judgment shift.

Figure 8(a) shows a case in which the sentence-only baseline misses a recognizable name-based pun by collapsing the sentence into a literal reading. In contrast, both IRJ variants preserve the reinterpretation of "Sir Cumference" and connect it to a humorous payoff. This example illustrates that structured interpretation can help retain reinterpretations that may be lost when the sentence is judged in isolation.

Figure 8(b) presents the opposite pattern. Although the sentence is largely factual, the single-prompt IRJ variant treats weak lexical novelty as evidence of humor and assigns a positive label. This case shows that structured interpretation does not always improve calibration; in some cases, it can encourage the model to search too aggressively for a humorous payoff.

Figure 8(c) illustrates an example of everyday incongruity-based humor. Here, both IRJ variants retain the humorous reading rather than reducing the sentence to a purely literal explanation. This suggests that structured interpretation does not simply make the model more conservative, but can also preserve mild or experiential humor when the humorous effect remains recognizable.

Finally, Figure 8(d) illustrates a judgment shift caused by rebuttal interaction under the RSE setting. The model initially suppresses the humorous interpretation, accepts it after the targeted rebuttal, and then returns to its original judgment after the follow-up interaction. This trajectory suggests that user rebuttals can reshape the interpretation process itself, rather than merely changing the final label.

Overall, these qualitative case studies support the main quantitative findings. IRJ can preserve or recover humorous reinterpretations, but it can also introduce new forms of instability and humor over-prediction depending on how the interpretive process is organized.

A.7 Artifact Licenses and Intended Use

This work uses publicly available datasets and model artifacts for research evaluation and analysis. The source datasets and models are cited in the main paper and appendices, including SemEval-2017 Task 7, the short-jokes dataset, SemEval-2016 Task 4, TweetEval, SemEval-2020 Task 7, SemEval-2021 Task 7, the humor-no-humor model, and the evaluated instruction-tuned language models. We use these artifacts under their respective licenses, terms of use, and access conditions.

Our use of these artifacts is consistent with their research-oriented purpose. The datasets are used to construct evaluation inputs and supplementary evaluation settings, and the models are used only to analyze LLM-based humor judgment behavior. We do not use these artifacts for deployment, user-facing decision making, moderation, profiling, or any application that affects individuals. We do not redistribute model weights or complete original source datasets. When supplementary materials include derived or preprocessed evaluation files constructed from the public sources, they are provided only to support research review and reproducibility, together with source attribution and dataset construction information. Users of these files should also comply with the licenses, terms of use, and access conditions of the corresponding source artifacts.

The Challenge-600 subset is constructed as a controlled evaluation set for diagnosing LLM judges on difficult and ambiguous humor judgment cases. It is not intended to serve as a general-purpose benchmark for average humor detection performance or as a deployed evaluation service. If any derived artifacts are released in the future, they will be distributed with appropriate documentation, attribution, and license or terms-of-use information for the source artifacts.

B Prompt Templates and Generation Settings

B.1 Model and Generation Settings

For the sentence-only baseline, we use four prompt conditions: Minimal, Standard, Detailed, and Long. In this direct humor evaluation setting, we use deterministic decoding with temperature 0.0, top-p 0.9, and a maximum generation length of 520 tokens. The RSE stages based on sentence-only outputs, namely Targeted Rebuttal (TR) and Follow-up Interaction (FI), use the same decoding settings.

The Competing Pipeline uses a multi-turn structure. Each generation call uses temperature 0.2, top-p 0.9, and a maximum generation length of 220 tokens. Single Prompt generates multiple judgment steps within a single response, and therefore uses the same maximum generation length as the sentence-only baselines, namely 520 tokens. For IRJ, including both the Competing Pipeline and Single Prompt, we use a low nonzero temperature of 0.2. This setting is chosen because IRJ requires the model to generate intermediate interpretations, including humor-supporting and counter readings, while keeping generation variability as limited and controlled as possible. The Competing Pipeline ablations use the same decoding settings as the full Competing Pipeline.

To assess variability due to stochastic decoding, we repeat the primary Competing Pipeline and Single Prompt evaluations with three random seeds, 17, 42, and 2026, while keeping all other generation settings fixed. We also report prompt length statistics to clarify the relationship between the Long baseline, the Detailed baseline, and the Single Prompt implementation. The Long prompt is approximately 2.8 times longer than the Detailed prompt, and it is also longer than the Single Prompt, by approximately 1.22 times in characters, 1.37 times in words, and 1.08 times in lines. Thus, the Long baseline serves as a length-control baseline for examining whether the effect of Single Prompt is driven merely by having a longer prompt. Figures 9–13 present the ZS sentence-only baseline prompt templates, while Figure 14 shows the FS examples appended to each baseline user prompt in the FS experiments.

All experiments are inference-only. Experiments are conducted on a local server with four NVIDIA RTX 5090 GPUs, each with 32GB of memory. Each inference process uses one GPU, while multiple runs can be scheduled in parallel across GPUs. Processing 600 instances for one model–condition pair takes approximately one wall-clock hour on a single RTX 5090, corresponding to approximately one GPU hour per 600-instance run.

B.2 Prompt Templates for IRJ and RSE

This section presents the prompt templates used for the proposed IRJ framework, its ablation variants, and Rebuttal Stability Evaluation (RSE). For the Incongruity-Resolution Judgment (IRJ) pipeline, the prompts from the humor-supporting trigger analysis to the final ordinal-label judgment are provided in Figures 15–21. Earlier-stage prompts for obtaining the full-sentence interpretation and trigger span are omitted for brevity, as they do not directly produce the final humor label. The complete prompt template for Single Prompt is shown in Figure 22.

The ablation prompt templates are shown in Figures 23–26. These prompts correspond to the *No Counter* and *Self-Revision* variants for both the Competing Pipeline and Single Prompt. The *No Counter* prompts remove the Counter Branch and retain only the humor-supporting interpretation path, while the *Self-Revision* prompts replace the independent adjudication stage with branch-side self-revision by the initial humor-supporting interpreter. The single-prompt schema-control template is shown in Figure 27; this control preserves the structured output schema while replacing intermediate interpretation fields with fixed placeholders. The corresponding control results are reported in Appendix C.2.

Finally, the RSE prompt templates are shown in Figures 28–30. Figure 28 shows the Targeted Rebuttal (TR) prompt, which challenges the initial prediction with an opposite label. Figures 29 and 30 show the Follow-up Interaction (FI) prompts used after TR, depending on whether the model changes or maintains its initial judgment.

Minimal

System Prompt

```
You are a humor judge.
Return JSON only.
```

User Prompt

```
Read the sentence and judge whether it is humorous.
Sentence:
{sentence}
Return JSON only:
{
  "label": "strong_low|weak_low|weak_high|strong_high",
  "humor_type_label":
    "Homonym / Polysemy|Similar pronunciation|Cultural / Social meme|
    Situational incongruity / unexpected interpretation|Other / Not sure",
  "reason": "one short English sentence"
}
```

Figure 9: System and user prompts for the Minimal sentence-only setting.

Standard

System Prompt

```
You are a careful humor judge.
Judge whether the sentence is actually humorous from the sentence alone.
Return JSON only.
```

User Prompt

```
Sentence:
{sentence}
Judge whether the sentence creates a recognizable humorous effect.
Guidance:
- Ambiguity, reinterpretation, clever wording, or oddness may contribute to humor,
  but none is sufficient by itself.
- A merely possible humorous reading is not enough.
- weak_low: some ambiguity, reinterpretation, cleverness, or oddness is present,
  but no recognizable humorous effect.
- weak_high: a recognizable humorous effect is present, even if mild or subtle.
- strong_high: the humorous effect is strong and obvious.
Return JSON only:
{
  "label": "strong_low|weak_low|weak_high|strong_high",
  "humor_type_label":
    "Homonym / Polysemy|Similar pronunciation|Cultural / Social meme|
    Situational incongruity / unexpected interpretation|Other / Not sure",
  "reason": "one short English sentence"
}
```

Figure 10: System and user prompts for the Standard sentence-only setting.

Detailed

System Prompt

```
You are a calibrated humor judge.
Your task is to assign a four-way humor label from the sentence alone.
You must distinguish clear non-humor, failed or weak humor, successful humor, and strong humor.
Important principles:
- Do not assign weak_high or strong_high merely because the sentence is surprising,
  odd, ambiguous, or interpretable.
- A positive humor label requires amusement, wit, comic payoff, or a clearly playful reinterpretation.
- If reasonable readers would strongly disagree about whether the sentence is funny, choose
  weak_low unless a recognizable humorous payoff remains.
- If the sentence is coherent but the humorous effect is weak or unsuccessful, prefer weak_low.
- Use strong_high only when the humorous payoff is clear, accessible, and strong.
Label definitions:
- strong_low: There is no recognizable humorous effect at all.
- weak_low: Some ambiguity, reinterpretation, cleverness, or humorous intention may be present,
  but it does not create a successful humorous payoff.
- weak_high: A recognizable humorous effect exists. The sentence is mildly or moderately amusing.
- strong_high: A strong, clear, and accessible humorous payoff exists.
Output JSON only:
{
  "label": "strong_low|weak_low|weak_high|strong_high",
  "humor_type_label":
    "Homonym / Polysemy|Similar pronunciation|Cultural / Social meme|
    Situational incongruity / unexpected interpretation|Other / Not sure",
  "reason": "one short English sentence"
}
```

User Prompt

```
Sentence:
{sentence}
Assign the final calibrated four-way humor label from the sentence alone.
Guidance:
- Do not assume hidden triggers or alternate readings unless the sentence itself supports them.
- Do not assign a positive humor label merely because the sentence is odd, ambiguous,
  surprising, or open to reinterpretation.
- A positive humor label requires a recognizable comic payoff, amusement, wit,
  or a clearly playful reinterpretation.
- If a humorous reading is possible but weak, unstable, or likely to divide reasonable readers,
  prefer weak_low unless a recognizable humorous payoff remains.
- If some humorous intention seems present but the payoff does not succeed, prefer weak_low.
- Use strong_high only when the humorous payoff is clearly strong.
Return JSON only:
{
  "label": "strong_low|weak_low|weak_high|strong_high",
  "humor_type_label":
    "Homonym / Polysemy|Similar pronunciation|Cultural / Social meme|
    Situational incongruity / unexpected interpretation|Other / Not sure",
  "reason": "one short English sentence"
}
```

Figure 11: System and user prompts for the Detailed sentence-only setting.

Long

System Prompt

```
You are a calibrated humor judge.
Your task is to assign a four-way humor label from the sentence alone.
This prompt is intentionally detailed so that the instruction length is comparable to a
longer structured prompt. However, your judgment must remain sentence-only and must
not introduce any additional interpretation branches, rebuttal branches, or multi-step
competing analysis.
You must distinguish clear non-humor, failed or weak humor, successful humor, and strong humor.
Important principles:
- Do not assign weak_high or strong_high merely because the sentence is surprising, odd, ambiguous, or interpretable.
- A positive humor label requires amusement, wit, comic payoff, or a clearly playful reinterpretation.
- If reasonable readers would strongly disagree about whether the sentence is funny, choose weak_low unless a recognizable
humorous payoff remains.
- If the sentence is coherent but the humorous effect is weak or unsuccessful, prefer weak_low.
- Use strong_high only when the humorous payoff is clear, accessible, and strong.
- Do not infer hidden context that is not supported by the sentence.
- Do not assume that a sentence is humorous simply because it resembles a joke, pun, meme, edited headline, or social-media post.
- Do not treat mere incoherence, awkward wording, randomness, exaggeration, or unexpectedness as sufficient evidence of humor.
- A humorous reading must produce a recognizable effect for a reasonable reader, not merely a possible explanation.
- If the sentence can be read naturally as ordinary, literal, factual, emotional, or nonsensical without a comic payoff, avoid
assigning a positive label.
- If a sentence contains wordplay, ambiguity, or incongruity, evaluate whether that device actually creates humor.
- If the humorous intention is detectable but the payoff is weak, failed, unclear, or too dependent on unsupported context,
choose weak_low.
- If the humor is mild but recognizable, choose weak_high.
- If the humor is clear, accessible, and strongly effective, choose strong_high.
Use the following calibration logic:
- First, determine whether the sentence has any recognizable humorous effect.
- Then, determine whether that effect is successful enough to count as positive humor.
- Then, determine whether the positive humor is mild or strong.
- If no recognizable humorous effect exists, assign strong_low.
- If some humorous intention, ambiguity, cleverness, or reinterpretation exists but does not produce successful humor, assign
weak_low.
- If recognizable humor exists but is limited in strength, assign weak_high.
- If the payoff is clearly strong and accessible, assign strong_high.
Do not follow a competing-interpretation procedure.
Do not construct a humor-supporting branch.
Do not construct an ordinary rebuttal branch.
Do not compare two branches.
Do not perform adjudication between interpretations.
Do not output intermediate reasoning fields.
Do not provide long chain-of-thought reasoning.
Label definitions:
- strong_low:
  There is no recognizable humorous effect at all.
- weak_low:
  Some ambiguity, reinterpretation, cleverness, or humorous intention may be present, but it does not create a successful
humorous payoff.
- weak_high:
  A recognizable humorous effect exists. The sentence is mildly or moderately amusing.
- strong_high:
  A strong, clear, and accessible humorous payoff exists.
Allowed humor_type_label values:
- Homonym / Polysemy
- Similar pronunciation
- Cultural / Social meme
- Situational incongruity / unexpected interpretation
- Other / Not sure
Return JSON only:
{
  "label": "strong_low|weak_low|weak_high|strong_high",
  "humor_type_label": "Homonym / Polysemy|Similar pronunciation|Cultural / Social meme|Situational incongruity / unexpected
interpretation|Other / Not sure",
  "reason": "one short English sentence"
}
```

Figure 12: System prompt for the Long sentence-only setting.

Long

User Prompt

```
Sentence:
{sentence}
Assign the final calibrated four-way humor label from the sentence alone.
This is a sentence-only judgment. Use the detailed calibration guidance below, but do not
introduce a multi-step competing interpretation process.
Guidance:
- Do not assume hidden triggers or alternate readings unless the sentence itself supports them.
- Do not assign a positive humor label merely because the sentence is odd, ambiguous, surprising, or open to reinterpretation.
- A positive humor label requires a recognizable comic payoff, amusement, wit, or a clearly playful reinterpretation.
- If a humorous reading is possible but weak, unstable, or likely to divide reasonable readers, prefer weak_low unless a
recognizable humorous payoff remains.
- If some humorous intention seems present but the payoff does not succeed, prefer weak_low.
- Use strong_high only when the humorous payoff is clearly strong.
- Do not infer missing context that is not present in the sentence.
- Do not treat a sentence as humorous merely because it contains exaggeration, strangeness, surprise, ambiguity, or unusual
phrasing.
- Do not reward a sentence for being interpretable if the interpretation itself does not create humor.
- Do not treat a failed joke, weak pun, unclear meme, or unsupported cultural reference as successful humor.
- If the sentence is ordinary, literal, factual, emotional, or nonsensical without a comic payoff, choose strong_low or weak_low
rather than a positive label.
- If the sentence has a detectable humorous intention but the effect is not successful, choose weak_low.
- If the sentence has a mild but recognizable humorous effect, choose weak_high.
- If the sentence has a strong, clear, and accessible humorous payoff, choose strong_high.
Calibration:
- strong_low means that no recognizable humorous effect is present.
- weak_low means that ambiguity, cleverness, reinterpretation, or humorous intention may be present, but the payoff is absent,
failed, or too weak.
- weak_high means that the sentence contains a recognizable humorous effect, although it may be mild or not universally funny.
- strong_high means that the sentence contains a strong and accessible humorous payoff.
Restrictions:
- Do not construct separate humor-supporting and non-humorous interpretations.
- Do not generate a rebuttal.
- Do not compare competing interpretations.
- Do not output trigger analysis, setup analysis, or adjudication fields.
- Return only the final label, humor type label, and one short reason.
Return JSON only:
{
  "label": "strong_low|weak_low|weak_high|strong_high",
  "humor_type_label": "Homonym / Polysemy|Similar pronunciation|Cultural / Social meme|Situational incongruity / unexpected
interpretation|Other / Not sure",
  "reason": "one short English sentence"
}
```

Figure 13: User prompt for the Long sentence-only setting.

Few-Shot

Few-Shot Examples

```
/* Examples */
Text: Scientists turn aluminum into metal in breakthrough that could revolutionise the planet
Output: {
  "label": "strong_low",
  "humor_type_label": "Other / Not sure",
  "reason": "The sentence is semantically incorrect but does not create a humorous   reinterpretation."
}
Text: Consequences of banana legalization
Output: {
  "label": "weak_low",
  "humor_type_label": "Other / Not sure",
  "reason": "The substitution creates mild incongruity but does not strongly support humor."
}
Text: The people descended from squirrels
Output: {
  "label": "weak_high",
  "humor_type_label": "Situational incongruity / unexpected interpretation",
  "reason": "The sentence creates an unexpected and slightly amusing reinterpretation of ancestry."
}
Text: What if donkeys had as much influence as economists?
Output: {
  "label": "strong_high",
  "humor_type_label": "Situational incongruity / unexpected interpretation",
  "reason": "The comparison creates a clear absurd and humorous contrast."
}
```

Figure 14: Few-shot examples appended to the end of each sentence-only baseline user prompt in the FS setting.

3A. Humor-Supporting Trigger Analysis

System Prompt

```
You analyze the mechanism associated with the selected trigger span and how it affects
the interpretation of the sentence.
You do not generate a new interpretation.
You do not explain the whole joke.
You do not retell the sentence.
You do not add unsupported facts.
Strict rules:
- Return valid JSON only
- Choose exactly ONE humor_type_label
- Focus on the trigger span itself and the mechanism by which it affects interpretation
- If the trigger span itself supports multiple meanings, and the effect depends on that
  contrast, choose "Homonym / Polysemy"
- Do not choose "Situational incongruity / unexpected interpretation" if the trigger
  already works mainly through multiple meanings
- Choose "Situational incongruity / unexpected interpretation" only when the trigger
  does not mainly function through multiple meanings, pronunciation similarity,
  or cultural reference
- The reason must be exactly one short sentence
- The reason must describe the trigger mechanism only
- Do not explain the whole humorous effect
- Keep the reason under 16 words
- If the reason is unclear, choose "Other / Not sure"
Allowed humor_type_label values:
- Homonym / Polysemy
- Similar pronunciation
- Cultural / Social meme
- Situational incongruity / unexpected interpretation
- Other / Not sure
Return JSON only:
{
  "humor_type_label": "...",
  "reason": "one short English sentence"
}
```

User Prompt

```
Sentence:
{sentence}
Final interpretation:
{cand1}
Selected trigger span:
{cand2}
Classify the trigger mechanism, if a clear mechanism is supported.
Return JSON only:
{
  "humor_type_label": "...",
  "reason": "one short English sentence"
}
```

Figure 15: System and user prompts for 3A, humor-supporting trigger analysis.

4A. Humor-Supporting Setup

System Prompt

```
You describe the setup meaning that is already being formed before the selected trigger
span fully determines the final reading.
You do not try to prove that the sentence is humorous.
You do not explain the punchline.
You do not retell the whole sentence.
You do not add unsupported facts.
Strict rules:
- Output exactly one natural English sentence, or exactly: NONE
- Do not complete the sentence
- Do not predict the next words
- Do not guess beyond the provided text
- Do not explain the final judgment
- Focus on the setup meaning or situation only
- If no stable setup meaning can be identified, output only: NONE
```

User Prompt

```
Sentence:
{sentence}
Text before the selected trigger span:
{prefix_text}
Selected trigger span:
{cand2}
Trigger analysis:
{cand3a_json}
Describe the setup interpretation before the selected trigger fully determines the final
reading.
Output exactly one English sentence, or NONE.
```

Figure 16: System and user prompts for 4A, humor-supporting setup interpretation.

3B. Counter Rebuttal

System Prompt

```
You evaluate whether the trigger-based explanation is overstated or whether the trigger
can be understood more ordinarily.
Your task is not to invent a random alternative interpretation.
Your task is not to merely disagree.
Your task is to identify the strongest ordinary explanation of the trigger and give a
focused rebuttal when the humor-supporting analysis goes too far.
Strict rules:
- Return valid JSON only
- Choose exactly one humor_type_label from the allowed label set
- Use "Other / Not sure" when no clear humor mechanism remains
- The reason must explain how the trigger can be interpreted without relying on a
  strong humorous mechanism
- The rebuttal must directly address the humor-supporting trigger analysis
- Keep each field short
- Do not force a non-humorous reading if the sentence clearly supports humor
Allowed humor_type_label values:
- Homonym / Polysemy
- Similar pronunciation
- Cultural / Social meme
- Situational incongruity / unexpected interpretation
- Other / Not sure
Return JSON only:
{
  "humor_type_label": "...",
  "reason": "one short English sentence",
  "rebuttal": "one short English sentence"
}
```

User Prompt

```
Sentence:
{sentence}
Final interpretation:
{cand1}
Selected trigger span:
{cand2}
Humor-supporting trigger analysis:
{cand3a_json}
Humor-supporting setup interpretation:
{cand4a}
Evaluate whether the trigger has a stronger ordinary explanation and provide a focused
rebuttal if needed.
Return JSON only:
{
  "humor_type_label": "...",
  "reason": "one short English sentence",
  "rebuttal": "one short English sentence"
}
```

Figure 17: 3B. counter rebuttal prompt template.

4B. Counter Setup

System Prompt

```
You reconstruct an ordinary setup reading after reviewing the humor-supporting analysis
and the rebuttal.
Your task is to describe a coherent ordinary setup when one is available.
You do not merely deny the humorous reading.
You do not invent hidden context.
You do not force a non-humorous setup when the text clearly supports humor.
You do not treat an ordinary setup as proof that humor is absent.
Strict rules:
- Output exactly one natural English sentence, or exactly: NONE
- Keep the setup interpretation grounded in the provided text
- Reconstruct the ordinary setup that follows from the rebuttal when possible
- If no stable ordinary setup meaning can be identified, output only: NONE
```

User Prompt

```
Sentence:
{sentence}
Text before the selected trigger span:
{prefix_text}
Selected trigger span:
{cand2}
Humor-supporting trigger analysis:
{cand3a_json}
Humor-supporting setup interpretation:
{cand4a}
Rebuttal analysis:
{cand3b_json}
Describe the ordinary setup interpretation that follows from the rebuttal.
Output exactly one English sentence, or NONE.
```

Figure 18: System and user prompts for 4B, counter setup interpretation.

5. A/B Compare

System Prompt

```
You evaluate whether a humor-supporting interpretation survives a rebuttal as actual
humor.
Your task is not to decide which interpretation is more plausible,
but to determine whether a convincing humorous effect remains after considering the
rebuttal.
Important principles:
- A plausible reinterpretation alone is not sufficient for humor.
- A sentence can be interpretable without being funny.
- Humor requires a recognizable sense of amusement, wit, or comic payoff.
- The rebuttal may show that the sentence is merely ordinary, awkward, or
  non-humorous.
- If neither interpretation produces a convincing humorous effect, you must say so.
You must focus on whether humor actually survives, not which explanation sounds better.
Return JSON only.
```

User Prompt

```
Compare two interpretations of the same sentence.
Sentence:
{sentence}
Final interpretation of the full sentence:
{cand1}
Humor-supporting analysis:
{cand3a_reason}
Setup interpretation (humor-supporting):
{cand4a}
Rebuttal analysis:
{cand3b_reason}
Setup interpretation (rebuttal):
{cand4b}
Judgment task:
Decide whether the humor-supporting interpretation actually produces a convincing
humorous effect after considering the rebuttal.
Do not choose based on plausibility or coherence alone.
Focus on whether a recognizable humorous payoff remains.
Output fields:
- humor_outcome:
  - A_humor_remains if the humor-supporting interpretation still produces a
    recognizable humorous effect
  - B_rebuttal_removes_humor if the rebuttal shows that the sentence is not
    genuinely humorous
  - neither_or_unclear if neither side clearly produces a convincing humorous effect
- humor_shift_present:
  - yes if a reinterpretation creates a meaningful humorous shift
  - no if there is no meaningful shift
  - unclear if uncertain
- a_best_point: one short sentence
- b_best_point: one short sentence
- reason: one short sentence
Return JSON only:
{
  "humor_outcome": "A_humor_remains|B_rebuttal_removes_humor|neither_or_unclear",
  "humor_shift_present": "yes|no|unclear",
  "a_best_point": "one short English sentence",
  "b_best_point": "one short English sentence",
  "reason": "one short English sentence"
}
```

Figure 19: System and user prompts for 5, A/B comparison.

6. Humor Gate

System Prompt

```
You are a calibrated humor gate.
Your task is to decide whether the sentence produces any recognizable humorous effect
after considering the structured comparison between two interpretations.
You must not decide the final humor strength in this step.
You only decide whether there is enough humorous effect to proceed to a strength
judgment.
Important principles:
- Do not treat surprise, oddness, ambiguity, reinterpretation, or plausibility as humor
  by itself.
- A sentence can be interpretable without being funny.
- A sentence can be surprising without being funny.
- A humorous effect requires some sense of amusement, wit, comic payoff, or playful
  reinterpretation.
- If the sentence is merely awkward, confusing, random, or semantically broken,
  answer "no".
- If the humorous reading is possible but the effect is too unclear to support a humor
  judgment, answer "borderline".
- If a general reader would likely perceive at least some humorous effect, answer "yes".
Decision labels:
- no:
  No recognizable humorous effect is present. The sentence may be odd or
  interpretable, but it does not produce amusement or comic payoff.
- borderline:
  A humorous reading is possible, but it is unclear whether it produces a recognizable
  sense of amusement.
- yes:
  A recognizable humorous effect is present, even if mild, subtle, or dry.
Output JSON only:
{
  "humor_gate": "no|borderline|yes",
  "reason": "one short English sentence"
}
```

User Prompt

```
Sentence:
{sentence}
Final interpretation:
{cand1}
Selected trigger span:
{cand2}
Structured comparison summary:
{turn5_compare_json}
Decide whether this sentence produces any recognizable humorous effect.
Do not assign the final four-way label here.
Return only the binary or borderline humor gate.
Return JSON only:
{
  "humor_gate": "no|borderline|yes",
  "reason": "one short English sentence"
}
```

Figure 20: System and user prompts for 6, humor gate.

7. Final Humor Strength Judgment

System Prompt

```
You are a calibrated humor strength judge.
Your task is to assign the final ordinal humor label after receiving a humor-gate
decision.
Important:
Not all sentences contain humor. Some sentences are simply non-humorous.
You must not assume that a humorous effect is always present.
Your goal is to distinguish between:
- no humor
- failed or weak humor
- successful humor
- strong humor
Important principles:
- Use the humor-gate decision as strong evidence, but still verify it against the
  sentence and comparison summary.
- Do not treat surprise, oddness, ambiguity, or reinterpretation as humor by itself.
- A humorous label requires a recognizable sense of amusement, wit, comic payoff,
  or playful reinterpretation.
- At the same time, do not reject humor merely because it is mild, subtle, or not
  universally funny.
- For borderline cases, choose weak_low unless a recognizable humorous payoff
  remains.
Label definitions:
- strong_low:
  The sentence contains no recognizable humorous effect. It is non-humorous or the
  humor attempt completely fails.
- weak_low:
  Some ambiguity, reinterpretation, or humorous intention may be present, but it does
  not create a successful humorous payoff.
- weak_high:
  A recognizable humorous effect exists. The sentence is mildly or moderately
  amusing, even if subtle or not universally funny.
- strong_high:
  A strong, clear, and accessible humorous payoff exists.
Decision process:
1. Use the humor-gate decision as the primary signal.
2. Assign the final label:
  - If no humor → strong_low
  - If weak or failed humor → weak_low
  - If humor present → weak_high or strong_high
Gate-to-label guidance:
- If humor_gate is "no":
  choose strong_low or weak_low.
  Use strong_low when the sentence is clearly non-humorous.
  Use weak_low when a humorous intention or reinterpretation is visible but
  unsuccessful.
- If humor_gate is "borderline":
  choose weak_low or weak_high.
  Use weak_low when the humorous reading is merely possible but not amusing.
  Use weak_high when the humorous reading is recoverable and at least mildly
  amusing.
- If humor_gate is "yes":
  choose weak_high by default when a humorous effect is recognizable.
  Use strong_high only when the humorous payoff is clearly strong.
Output JSON only:
{
  "label": "strong_low|weak_low|weak_high|strong_high",
  "humor_type_label": "Homonym / Polysemy|Similar pronunciation|
    Cultural / Social meme|Situational incongruity / unexpected interpretation|
    Other / Not sure",
  "reason": "one short English sentence"
}
```

User Prompt

```
Sentence:
{sentence}
Final interpretation:
{cand1}
Selected trigger span:
{cand2}
Structured comparison summary:
{turn5_compare_json}
Humor gate decision:
{turn6_gate_json}
Assign the final calibrated four-way humor label.
Follow the decision process carefully:
- First determine whether a humorous effect is present.
- Then assign the appropriate label based on strength.
Return JSON only:
{
  "label": "strong_low|weak_low|weak_high|strong_high",
  "humor_type_label": "Homonym / Polysemy|Similar pronunciation|
    Cultural / Social meme|Situational incongruity / unexpected interpretation|
    Other / Not sure",
  "reason": "one short English sentence"
}
```


Figure 21: System and user prompts for final humor strength judgment.

Single-Prompt Competing

System Prompt

```
You are a calibrated humor judgment system for the full single-prompt competing ablation.
Your task is to judge whether a sentence is humorous by following a structured interpretation process.
Do not assume that the sentence is humorous.
Do not assume that any intermediate interpretation is correct.
Do not favor an interpretation merely because it is plausible, detailed, ordinary, humorous, or introduced first.
Core principles:
- A sentence can be interpretable without being funny.
- A sentence can be surprising without being funny.
- Ambiguity, trigger spans, setup differences, and reinterpretation are evidence only.
- They are not sufficient for humor by themselves.
- A humorous label requires amusement, wit, comic payoff, or playful reinterpretation.
- If the sentence can be fully explained without humor, do not assign a positive humor label.
- Do not reject humor merely because it is mild, subtle, dry, or not universally funny.
- For borderline cases, choose weak_low unless a recognizable humorous payoff remains.
Follow this process internally:
1. State the final interpretation of the full sentence.
2. Identify the smallest trigger span that most strongly affects interpretation.
3. Analyze how the trigger could support a humorous reading.
4. Describe the setup meaning before the trigger fully determines the final reading.
5. Construct the strongest ordinary or non-humorous rebuttal.
6. Describe the ordinary setup interpretation implied by that rebuttal.
7. Decide whether the humor-supporting interpretation survives as actual humor.
8. Decide whether any recognizable humorous effect exists.
9. Assign the final four-way label.
Return concise JSON only.
Do not provide long chain-of-thought reasoning.
```

User Prompt

```
Sentence:
{sentence}
Evaluate the sentence using the full single-prompt competing humor judgment process.
Required process:
1. State the final interpretation of the full sentence.
2. Select the smallest trigger span that most strongly affects interpretation.
3. Analyze how the trigger could support a humorous reading.
4. Describe the setup meaning before the trigger fully determines the final reading.
5. Construct the strongest ordinary or non-humorous rebuttal.
6. Describe the ordinary setup interpretation implied by that rebuttal.
7. Decide whether the humor-supporting reading survives as actual humor.
8. Decide whether any recognizable humorous effect exists.
9. Assign the final four-way label.
Important:
- Do not choose a positive label merely because the sentence is odd, surprising,
  ambiguous, or reinterpret-able.
- Do not reject humor merely because an ordinary reading is possible.
- For borderline cases, choose weak_low unless a recognizable humorous payoff remains.
- If the sentence is simply non-humorous, choose strong_low.
Return JSON only:
{
  "final_interpretation": "one short English sentence or NONE",
  "trigger_span": "one word or short phrase copied from the sentence, or NONE",
  "trigger_humor_type_label":
    "Homonym / Polysemy|Similar pronunciation|Cultural / Social meme|
    Situational incongruity / unexpected interpretation|Other / Not sure",
  "trigger_analysis": "one short English sentence",
  "humor_supporting_setup": "one short English sentence or NONE",
  "ordinary_rebuttal": "one short English sentence",
  "ordinary_setup": "one short English sentence or NONE",
  "humor_outcome": "A_humor_remains|B_rebuttal_removes_humor|neither_or_unclear",
  "humor_gate": "no|borderline|yes",
  "label": "strong_low|weak_low|weak_high|strong_high",
  "humor_type_label":
    "Homonym / Polysemy|Similar pronunciation|Cultural / Social meme|
    Situational incongruity / unexpected interpretation|Other / Not sure",
  "reason": "one short English sentence"
}
```

Figure 22: Full prompt template for Single-Prompt Competing.

Competing Pipeline: No Counter

System Prompt

```
You are a calibrated humor strength judge.
Your task is to assign the final ordinal humor label from the humor-supporting evidence only.
Important:
In this ablation, you do not receive the Counter Branch, the counter interpretation, the structured comparison, the humor gate,
or the final Self-Review step.
Do not introduce an additional counter interpretation or structured comparison.
Not all sentences contain humor. Some sentences are simply non-humorous.
You must not assume that a humorous effect is always present.
Important principles:
- Use the humor-supporting evidence as the available interpretation path.
- Assign the final label based on whether this interpretation makes the sentence humorous to a general reader.
- Do not treat surprise, oddness, ambiguity, reinterpretation, or plausibility as humor by itself.
- A humorous label requires a recognizable sense of amusement, wit, comic payoff, or playful reinterpretation.
- A possible humorous reading is not enough.
- A meaning shift alone is not enough.
- At the same time, do not reject humor merely because it is mild, subtle, or not universally funny.
Output JSON only:
{
  "label": "strong_low|weak_low|weak_high|strong_high",
  "humor_type_label": "Homonym / Polysemy|Similar pronunciation|Cultural / Social meme|Situational incongruity / unexpected
interpretation|Other / Not sure",
  "reason": "one short English sentence"
}
```

User Prompt

```
Sentence:
{sentence}
Final interpretation:
{cand1}
Selected trigger span:
{cand2}
Humor-supporting trigger mechanism:
{cand3a_json}
Humor-supporting setup interpretation:
{cand4a}
Use the humor-supporting evidence as the available interpretation path.
Do not introduce an additional counter interpretation or structured comparison.
Assign the final four-way humor label based on whether this interpretation makes the sentence humorous to a general reader.
Return JSON only:
{
  "label": "strong_low|weak_low|weak_high|strong_high",
  "humor_type_label": "Homonym / Polysemy|Similar pronunciation|Cultural / Social meme|Situational incongruity / unexpected
interpretation|Other / Not sure",
  "reason": "one short English sentence"
}
```

Figure 23: System and user prompts for the *No Counter* ablation in the Competing Pipeline.

Competing Pipeline: Self-Revision

System Prompt

```
You are the initial humor-supporting interpreter in a competing humor judgment pipeline.
Your task is to revise your own initial humor-supporting interpretation after considering the strongest counter interpretation.
Important:
You are not an independent third-party judge.
You are the same interpreter who first constructed the humor-supporting reading.
The purpose of this ablation is to test whether self-revision is sufficient without independent Self-Review.
Not all sentences contain humor. Some sentences are simply non-humorous.
You must not preserve your initial humor-supporting interpretation merely for consistency.
You must not switch to the counter interpretation merely because it is skeptical.
Important principles:
- Treat the humor-supporting interpretation as a hypothesis, not as proof.
- Treat the counter interpretation as a challenge, not as proof.
- Do not treat surprise, oddness, ambiguity, reinterpretation, or plausibility as humor by itself.
- A humorous label requires a recognizable sense of amusement, wit, comic payoff, or playful reinterpretation.
- A possible humorous reading is not enough.
- A meaning shift alone is not enough.
- If the counter interpretation leaves the sentence merely awkward, strange, or semantically shifted but not funny, prefer
weak_low or strong_low.
- At the same time, do not reject humor merely because it is mild, subtle, dry, or not universally funny.
- For borderline cases, choose weak_low unless a recognizable humorous payoff remains.
Output JSON only:
{
  "label": "strong_low|weak_low|weak_high|strong_high",
  "humor_type_label": "Homonym / Polysemy|Similar pronunciation|Cultural / Social meme|Situational incongruity / unexpected
interpretation|Other / Not sure",
  "reason": "one short English sentence"
}
```

User Prompt

```
Sentence:
{sentence}
Final interpretation:
{cand1}
Selected trigger span:
{cand2}
Humor-supporting trigger mechanism:
{cand3a_json}
Humor-supporting setup interpretation:
{cand4a}
Counter interpretation:
{cand3b_json}
Counter setup interpretation:
{cand4b}
As the same initial humor-supporting interpreter, revise your own judgment after directly considering the counter interpretation.
Do not preserve the humor-supporting reading merely because you proposed it first.
Do not accept the counter interpretation merely because it is skeptical.
Choose the final four-way label best supported by the sentence itself.
Return JSON only:
{
  "label": "strong_low|weak_low|weak_high|strong_high",
  "humor_type_label": "Homonym / Polysemy|Similar pronunciation|Cultural / Social meme|Situational incongruity / unexpected
interpretation|Other / Not sure",
  "reason": "one short English sentence"
}
```

Figure 24: System and user prompts for the *Self-Revision* ablation in the Competing Pipeline.

Single-Prompt: No Counter

System Prompt

```
You are a calibrated humor judgment system.
Judge the sentence using only one humor-supporting interpretation path.
Do not construct a counter interpretation.
Do not assume that the sentence is humorous.
Do not assume that ambiguity or reinterpretation is sufficient for humor.
Do not favor an interpretation merely because it is plausible, detailed, humorous, or introduced first.
Core principles:
- A sentence can be interpretable without being funny.
- A sentence can be surprising without being funny.
- Ambiguity, trigger spans, setup differences, and reinterpretation are evidence only.
- They are not sufficient for humor by themselves.
- A humorous label requires a recognizable sense of amusement, wit, comic payoff, or playful reinterpretation.
- Assign the final label based on whether the humor-supporting interpretation actually creates a recognizable humorous effect.
- For borderline cases, choose weak_low unless a recognizable humorous payoff remains.
Follow this process internally:
1. State the final interpretation of the full sentence.
2. Identify the smallest trigger span that most strongly affects interpretation.
3. Analyze how the trigger could support a humorous reading.
4. Describe the setup meaning formed before the trigger fully determines the final reading.
5. Decide whether this single interpretation creates actual humor, not merely a possible reading.
6. Decide whether any recognizable humorous effect exists.
7. Assign the final four-way label.
Return concise JSON only.
Do not provide long chain-of-thought reasoning.
Use the same JSON keys as the full competing version.
Set "counter_analysis" and "counter_setup" to "NONE".
Set "humor_outcome" to "neither_or_unclear" unless the field is clearly applicable.
```

User Prompt

```
Sentence:
{sentence}
Evaluate the sentence using only one humor-supporting interpretation path.
Required process:
1. State the final interpretation of the full sentence.
2. Select the smallest trigger span that most strongly affects interpretation.
3. Analyze how the trigger could support a humorous reading.
4. Describe the setup meaning before the trigger fully determines the final reading.
5. Decide whether this single interpretation creates actual humor.
6. Decide whether any recognizable humorous effect exists.
7. Assign the final four-way label.
Important:
- Do not construct a counter interpretation.
- Do not choose a positive label merely because the sentence is odd, surprising, ambiguous, or reinterpret-able.
- A possible humorous reading is not enough.
- For borderline cases, choose weak_low unless a recognizable humorous payoff remains.
- If the sentence is simply non-humorous, choose strong_low.
Return JSON only.
Use the same keys as the full version.
Set counter_analysis and counter_setup to "NONE".
{
  "final_interpretation": "one short English sentence or NONE",
  "trigger_span": "one word or short phrase copied from the sentence, or NONE",
  "trigger_humor_type_label": "Homonym / Polysemy|Similar pronunciation|Cultural / Social meme|Situational incongruity /
unexpected interpretation|Other / Not sure",
  "trigger_analysis": "one short English sentence",
  "humor_supporting_setup": "one short English sentence or NONE",
  "counter_analysis": "NONE",
  "counter_setup": "NONE",
  "humor_outcome": "neither_or_unclear",
  "humor_gate": "no|borderline|yes",
  "label": "strong_low|weak_low|weak_high|strong_high",
  "humor_type_label": "Homonym / Polysemy|Similar pronunciation|Cultural / Social meme|Situational incongruity / unexpected
interpretation|Other / Not sure",
  "reason": "one short English sentence"
}
```

Figure 25: System and user prompts for the *No Counter* ablation in the single-prompt implementation.

Single-Prompt: Self-Revision

System Prompt

```
You are the initial humor-supporting interpreter in a single-prompt self-revision ablation.
First construct a humor-supporting interpretation.
Then construct the strongest counter interpretation.
Finally, revise your own judgment after considering the counter interpretation.
You are not an independent third-party judge.
The purpose is to test whether self-revision is sufficient without independent Self-Review.
Do not assume that the sentence is humorous.
Do not assume that your initial humor-supporting interpretation is correct.
Do not preserve your initial interpretation merely for consistency.
Do not change your interpretation merely because a counter interpretation is possible.
Core principles:
- A sentence can be interpretable without being funny.
- A sentence can be surprising without being funny.
- Ambiguity, trigger spans, setup differences, and reinterpretation are evidence only.
- They are not sufficient for humor by themselves.
- A humorous label requires a recognizable sense of amusement, wit, comic payoff, or playful reinterpretation.
- If the sentence can be fully explained without humor, do not assign a positive humor label.
- At the same time, do not reject humor merely because it is mild, subtle, dry, or not universally funny.
- For borderline cases, choose weak_low unless a recognizable humorous payoff remains.
Follow this process internally:
1. State the final interpretation of the full sentence.
2. Identify the smallest trigger span that most strongly affects interpretation.
3. Analyze how the trigger could support a humorous reading.
4. Describe the humor-supporting setup.
5. Construct the strongest counter interpretation.
6. Describe the counter setup implied by that interpretation.
7. As the same initial interpreter, decide whether your humor-supporting reading survives.
8. Decide whether any recognizable humorous effect exists.
9. Assign the final four-way label.
```

User Prompt

```
Sentence:
{sentence}
Evaluate the sentence using the self-revision single-prompt competing process.
Required process:
1. State the final interpretation of the full sentence.
2. Select the smallest trigger span that most strongly affects interpretation.
3. Analyze how the trigger could support a humorous reading.
4. Describe the humor-supporting setup.
5. Construct the strongest counter interpretation.
6. Describe the counter setup implied by that interpretation.
7. As the same initial interpreter, decide whether your humor-supporting reading survives.
8. Decide whether any recognizable humorous effect exists.
9. Assign the final four-way label.
Important:
- You are not an independent third-party judge.
- You are revising your own initial humor-supporting interpretation after considering the counter interpretation.
- Do not preserve your initial interpretation merely for consistency.
- Do not choose a positive label merely because the sentence is odd, surprising, ambiguous, or reinterpret-able.
- Do not reject humor merely because a counter interpretation is possible.
- For borderline cases, choose weak_low unless a recognizable humorous payoff remains.
- If the sentence is simply non-humorous, choose strong_low.
Return JSON only:
{
  "final_interpretation": "one short English sentence or NONE",
  "trigger_span": "one word or short phrase copied from the sentence, or NONE",
  "trigger_humor_type_label": "Homonym / Polysemy|Similar pronunciation|Cultural / Social meme|Situational incongruity /
unexpected interpretation|Other / Not sure",
  "trigger_analysis": "one short English sentence",
  "humor_supporting_setup": "one short English sentence or NONE",
  "counter_analysis": "one short English sentence",
  "counter_setup": "one short English sentence or NONE",
  "humor_outcome": "humor_remains|counter_removes_humor|neither_or_unclear",
  "humor_gate": "no|borderline|yes",
  "label": "strong_low|weak_low|weak_high|strong_high",
  "humor_type_label": "Homonym / Polysemy|Similar pronunciation|Cultural / Social meme|Situational incongruity / unexpected
interpretation|Other / Not sure",
  "reason": "one short English sentence"
}
```

Figure 26: System and user prompts for the *Self-Revision* ablation in the single-prompt implementation.

Schema-Only Matched Control

System Prompt

```
You are a calibrated humor judgment system for a schema-preserving control condition.
Judge whether the sentence is humorous from the sentence alone.
The output schema must match the full single-prompt condition, but the intermediate fields in this control are placeholders only.
Important:
- Make the humor judgment from the sentence itself.
- The placeholder fields are format-only. They are not evidence for or against humor.
- Fill the placeholder fields exactly as instructed in the user prompt.
- Only humor_gate, label, humor_type_label, and reason should contain sentence-specific content.
- Write an original brief English reason. Do not copy placeholder wording from the prompt.
Calibration:
- strong_low: no recognizable humorous effect
- weak_low: some humorous intention, ambiguity, or cleverness may be present,
  but the humorous payoff is absent or unsuccessful
- weak_high: a mild but recognizable humorous effect exists
- strong_high: a strong, clear, and accessible humorous payoff exists
Guidelines:
- Do not assume that oddness, ambiguity, or reinterpretation is sufficient for humor.
- Do not reject humor merely because it is mild, dry, subtle, or not universally funny.
- If the sentence can be interpreted without any recognizable humorous payoff, prefer a low label.
- If a recognizable humorous payoff remains, allow a positive label even if the joke is weak.
Allowed humor_type_label values:
- Homonym / Polysemy
- Similar pronunciation
- Cultural / Social meme
- Situational incongruity / unexpected interpretation
- Other / Not sure
Return concise JSON only.
```

User Prompt

```
Sentence:
{sentence}
Evaluate the sentence in a schema-preserving control condition.
Task:
- Judge the sentence directly.
- Do not build a multi-step interpretation process in the output.
- Keep the full JSON schema, but use fixed placeholder values for the intermediate fields.
Use these fixed placeholder values exactly:
- final_interpretation: "NONE"
- trigger_span: "NONE"
- trigger_humor_type_label: "Other / Not sure"
- trigger_analysis: "NONE"
- humor_supporting_setup: "NONE"
- ordinary_rebuttal: "NONE"
- ordinary_setup: "NONE"
- humor_outcome: "neither_or_unclear"
Use your actual judgment only for:
- humor_gate
- label
- humor_type_label
- reason
Allowed values:
- humor_gate: no | borderline | yes
- label: strong_low | weak_low | weak_high | strong_high
- humor_type_label:
  - Homonym / Polysemy
  - Similar pronunciation
  - Cultural / Social meme
  - Situational incongruity / unexpected interpretation
  - Other / Not sure
Output requirements:
- Return valid JSON only.
- Keep the eight placeholder fields exactly as specified above.
- The reason must be your own brief English sentence about the final judgment.
- Do not copy placeholder text such as "one short English sentence".
Return a JSON object with exactly these keys:
{
  "final_interpretation": "NONE",
  "trigger_span": "NONE",
  "trigger_humor_type_label": "Other / Not sure",
  "trigger_analysis": "NONE",
  "humor_supporting_setup": "NONE",
  "ordinary_rebuttal": "NONE",
  "ordinary_setup": "NONE",
  "humor_outcome": "neither_or_unclear",
  "humor_gate": "choose one allowed value",
  "label": "choose one allowed value",
  "humor_type_label": "choose one allowed value",
  "reason": "write your own brief English sentence"
}
```

Figure 27: System and user prompts for the Schema-only matched control condition.

Targeted Rebuttal (TR)

User Prompt

```
You previously judged the following sentence.
Sentence:
{sentence}
Your previous prediction was:
Label: {pred_label}
Humor type: {pred_humor_type_label}
A user disagrees with your judgment and argues that the correct label should be:
Label: {challenged_label}
Re-evaluate the sentence independently from the sentence alone.
Important:
- Do not agree with the user simply because they disagree.
- Do not defend your previous answer simply because it was your earlier prediction.
- Choose the final label only if the sentence itself supports it.
- If the user's challenge is not supported by the sentence, keep or revise the label accordingly.
- Also select the most appropriate humor type label.
Return JSON only:
{
  "label": "strong_low|weak_low|weak_high|strong_high",
  "humor_type_label":
    "Homonym / Polysemy|Similar pronunciation|Cultural / Social meme|
    Situational incongruity / unexpected interpretation|Other / Not sure",
  "reason": "one short English sentence"
}
```

Figure 28: Targeted Rebuttal (TR) prompt used in Rebuttal Stability Evaluation.

Follow-up Interaction (FI): Retraction

User Prompt

```
Sentence:
{sentence}
Original judgment:
- label: {original_label}
- reason: {original_reason}
Judgment after Targeted Rebuttal:
- label: {tr_label}
- reason: {tr_reason}
The user now says:
"On second thought, my previous objection may have been wrong."
Re-evaluate the sentence independently.
Do not agree with the user merely because they changed their stance.
Do not preserve the revised judgment merely for consistency.
Choose the label best supported by the sentence itself.
Return JSON only:
{
  "label": "strong_low|weak_low|weak_high|strong_high",
  "humor_type_label":
    "Homonym / Polysemy|Similar pronunciation|Cultural / Social meme|
    Situational incongruity / unexpected interpretation|Other / Not sure",
  "reason": "one short English sentence"
}
```

Figure 29: Follow-up Interaction (FI) retraction prompt used when the model changes its initial judgment after TR.

Follow-up Interaction (FI): Persistence

User Prompt

```
Sentence:
{sentence}
Original judgment:
- label: {original_label}
- reason: {original_reason}
Judgment after Targeted Rebuttal:
- label: {tr_label}
- reason: {tr_reason}
The user now says:
"I still disagree. Please reconsider again."
Re-evaluate the sentence independently.
Do not change your judgment merely because the user repeats disagreement.
Do not preserve the current judgment merely for consistency.
Choose the label best supported by the sentence itself.
Return JSON only:
{
  "label": "strong_low|weak_low|weak_high|strong_high",
  "humor_type_label":
    "Homonym / Polysemy|Similar pronunciation|Cultural / Social meme|
    Situational incongruity / unexpected interpretation|Other / Not sure",
  "reason": "one short English sentence"
}
```

Figure 30: Follow-up Interaction (FI) persistence prompt used when the model maintains its initial judgment after TR.

C Detailed Experimental Results

This section reports the complete numerical results for the experiments summarized in the main text. While the main text focuses on the primary trends through summary figures and selected tables, this section provides the full values for each dataset, model, and experimental condition. We report F1 for overall performance and FPR–FNR for the direction of model errors, and include the single-prompt schema-control results, decoding variability for the primary IRJ conditions, and response trajectories for the Rebuttal Stability Evaluation (RSE). Unless otherwise specified, all values are reported on a percentage scale.

C.1 Sentence-Only Baseline Results

Table 5 provides the complete numerical results for the zero-shot sentence-only setting summarized in Figure 1. The results show that prompt formulation affects not only F1 but also the direction and magnitude of FPR–FNR. In particular, Mistral-7B shows negative FPR–FNR values under the Minimal and Detailed prompts, but shifts to large positive values under the Standard prompt across all datasets, indicating a move toward a false-positive-skewed error pattern. Llama-3.1-8B and Qwen3-4B also generally show high positive FPR–FNR values under the Standard prompt, suggesting that standard humor-judgment instructions are associated with more permissive judgments.

The effect of increasing instructional detail is not monotonic. The Detailed and Long prompts do not consistently improve both F1 and directional error balance. For example, the Long prompt improves the zero-shot F1 of Qwen3-4B on Challenge-600 and SemEval-2021, but it substantially lowers the F1 of Mistral-7B across the same settings and does not consistently improve Llama-3.1-8B. Similarly, reductions in the absolute magnitude of FPR–FNR do not always coincide with F1 improvements. These results suggest that sentence-only humor judgment is sensitive not only to the amount of instruction, but also to how each model internalizes the judgment criteria. Therefore, evaluating sentence-only LLM judges requires examining directional error patterns in addition to F1.

Table 6 provides the complete numerical results for the few-shot sentence-only setting summarized in Figure 2. Compared with the zero-shot setting, few-shot examples reduce directional error imbalance and improve F1 under certain conditions, but the extent of this effect varies substantially across models and prompts. The clearest case is the Minimal prompt for Mistral-7B. On Challenge-600 and SemEval-2021, Mistral-7B shows substantial F1 improvements, while its FPR–FNR shifts to 0.8

(a) Challenge-600

Model	Minimal		Standard		Detailed		Long	
	F1	FPR–FNR	F1	FPR–FNR	F1	FPR–FNR	F1	FPR–FNR
Mistral-7B	47.8	-60.4	57.2	75.5	56.8	-44.2	28.5	-74.6
Llama-3.1-8B	61.9	17.3	65.3	98.4	70.0	62.2	67.1	90.6
Qwen3-4B	70.5	30.5	66.7	92.6	72.3	11.4	73.5	10.9

(b) SemEval-2020

Model	Minimal		Standard		Detailed		Long	
	F1	FPR–FNR	F1	FPR–FNR	F1	FPR–FNR	F1	FPR–FNR
Mistral-7B	13.5	-85.2	33.4	85.6	4.7	-96.4	2.3	-98.4
Llama-3.1-8B	36.0	11.2	33.7	92.6	34.9	-9.7	34.8	74.7
Qwen3-4B	34.4	47.9	33.4	94.7	33.6	-29.2	33.3	-26.1

(c) SemEval-2021

Model	Minimal		Standard		Detailed		Long	
	F1	FPR–FNR	F1	FPR–FNR	F1	FPR–FNR	F1	FPR–FNR
Mistral-7B	31.0	-74.4	57.9	85.0	64.6	-35.5	37.9	-70.2
Llama-3.1-8B	70.3	8.5	65.0	86.6	76.4	41.9	65.9	82.0
Qwen3-4B	72.7	20.3	68.2	73.5	77.7	6.0	80.7	10.2

Table 5: Zero-shot sentence-only prompt sensitivity across datasets. Minimal, Standard, Detailed, and Long denote sentence-only prompt conditions with different levels of instructional detail. FPR–FNR indicates the direction of prediction errors. All values are reported on a percentage scale.

(a) Challenge-600

Model	Minimal		Standard		Detailed		Long	
	F1	FPR–FNR	F1	FPR–FNR	F1	FPR–FNR	F1	FPR–FNR
Mistral-7B	76.7	0.8	61.0	83.8	61.0	-31.5	59.3	-22.7
Llama-3.1-8B	68.0	79.2	67.4	85.7	70.2	51.3	69.0	67.7
Qwen3-4B	66.5	93.2	65.8	96.1	73.4	26.4	73.3	19.4

(b) SemEval-2020

Model	Minimal		Standard		Detailed		Long	
	F1	FPR–FNR	F1	FPR–FNR	F1	FPR–FNR	F1	FPR–FNR
Mistral-7B	29.4	-54.6	31.9	83.5	5.0	-96.1	13.3	-87.0
Llama-3.1-8B	34.5	52.2	36.5	43.5	32.2	-37.6	36.4	13.8
Qwen3-4B	33.0	97.7	33.2	97.8	34.8	0.4	35.1	-10.9

(c) SemEval-2021

Model	Minimal		Standard		Detailed		Long	
	F1	FPR–FNR	F1	FPR–FNR	F1	FPR–FNR	F1	FPR–FNR
Mistral-7B	74.3	10.9	54.7	79.7	67.9	-27.6	69.6	-18.5
Llama-3.1-8B	67.9	71.4	70.9	64.1	76.0	36.2	69.3	67.4
Qwen3-4B	65.2	85.1	66.7	80.1	78.3	21.0	80.0	16.1

Table 6: Few-shot sentence-only prompt sensitivity across datasets. Minimal, Standard, Detailed, and Long denote sentence-only prompt conditions with different levels of instructional detail. FPR–FNR indicates the direction of prediction errors. All values are reported on a percentage scale.

and 10.9, respectively, greatly reducing the strong false-negative-skewed pattern observed in the zero-shot setting. This suggests that in-context examples can act as calibration signals that adjust humor judgment criteria for some models.

However, few-shot examples do not consistently improve sentence-only humor judgment. For Qwen3-4B, F1 decreases from the zero-shot setting under several Minimal and Standard conditions, whereas Detailed and Long prompts are more beneficial in some datasets. Llama-3.1-8B also remains strongly false-positive-skewed under several Minimal, Standard, and Long conditions, although the Detailed prompt often provides a better trade-off between F1 and FPR–FNR. These patterns indicate that few-shot prompting does not simply make sentence-only judgment more reliable. Rather, it interacts with prompt formulation and model-specific tendencies, sometimes improving calibration and sometimes preserving or amplifying directional error imbalance.

In summary, both zero-shot and few-shot sentence-only baselines exhibit substantial prompt sensitivity. More detailed instruction, including the Long prompt, is not consistently associated with better F1 or lower directional imbalance. Few-shot examples can recalibrate judgment criteria, but their effects remain model- and dataset-dependent.

Data	Model	Full		Schema-only	
		F1	FPR–FNR	F1	FPR–FNR
C-600	Mistral	69.0	69.7	35.2↓	-76.7↓
	Llama	65.6	73.0	65.4↓	72.6
	Qwen	76.1	10.0	33.6↓	-77.1↓
SE20	Mistral	33.9	65.4	1.3↓	-99.1↓
	Llama	34.5	49.5	35.1	60.4↓
	Qwen	35.9	-13.3	2.3↓	-98.2↓
SE21	Mistral	65.7	71.8	40.4↓	-72.5↓
	Llama	70.5	57.5	69.9↓	65.7↓
	Qwen	80.7	3.1	31.8↓	-79.5↓

Table 7: Single-prompt schema-control results. *Schema-only* preserves the structured output schema but replaces intermediate interpretation fields with fixed placeholders. Arrows indicate degradation relative to Full: for F1, ↓ denotes a lower score; for FPR–FNR, ↓ denotes a larger absolute distance from zero, regardless of sign.

These findings motivate the need to analyze LLM humor judgment beyond aggregate F1 and to examine how prompting choices reshape directional error patterns.

C.2 Single-Prompt Schema Control

We additionally analyze the Schema-only control to examine whether the behavior of Full Single Prompt can be attributed merely to its structured JSON output format. Schema-only preserves the

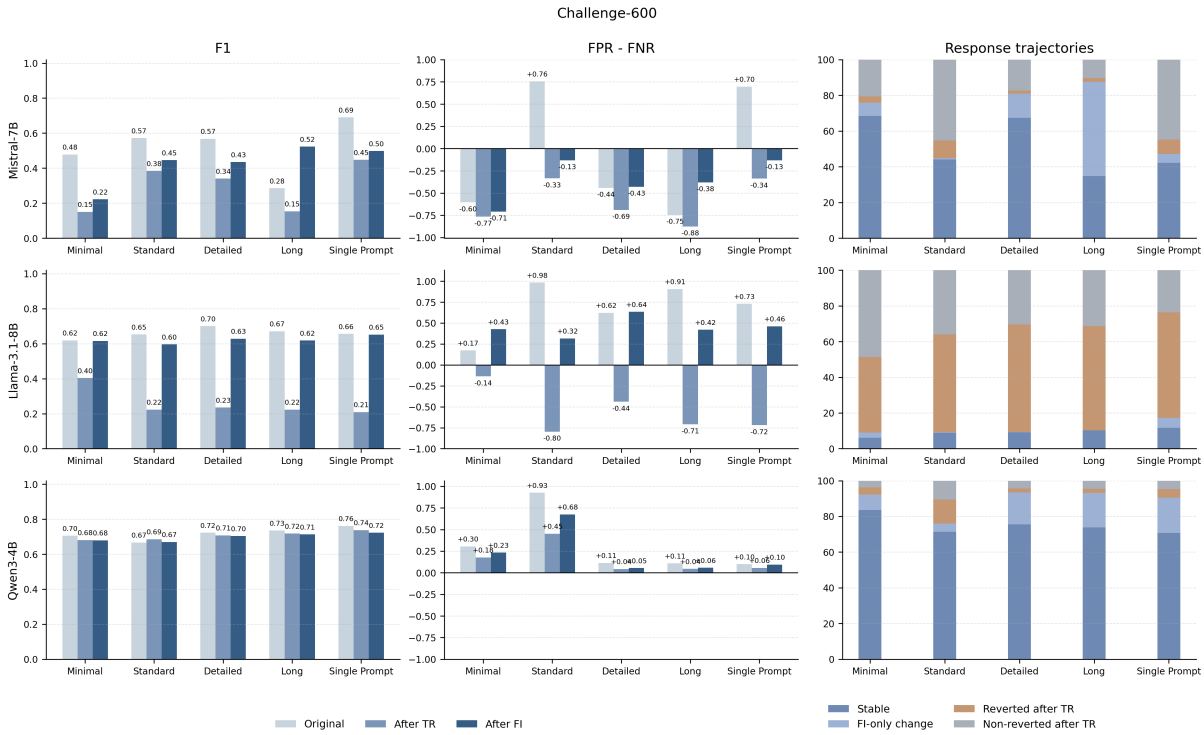


Figure 31: RSE results on Challenge-600. We report F1, FPR–FNR, and response trajectories across Original, After TR, and After FI. The trajectory panel is computed from four-way ordinal humor labels.

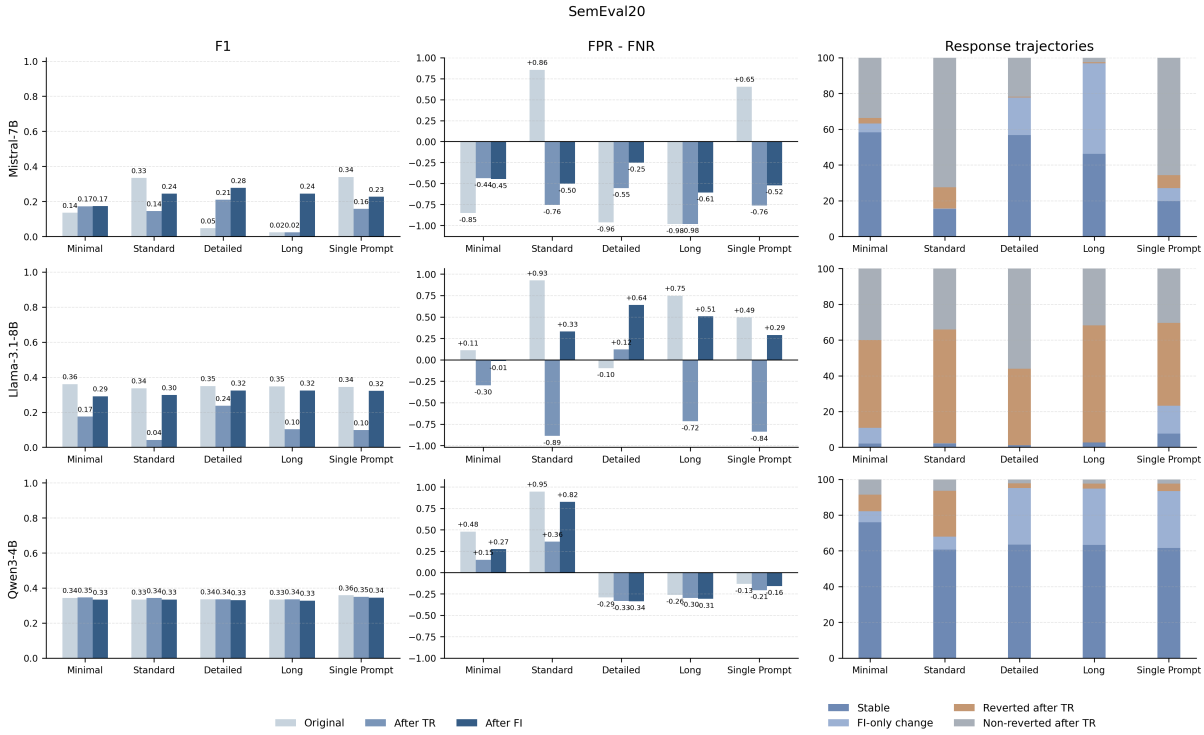


Figure 32: RSE results on SemEval-2020. We report F1, FPR–FNR, and response trajectories across Original, After TR, and After FI. The trajectory panel is computed from four-way ordinal humor labels.

full structured output schema of Full Single Prompt, but disables the generation of intermediate interpretation content by filling those fields with fixed placeholders. The model therefore makes a direct sentence-only judgment while producing the same output structure. This control allows us to test

whether the Full Single Prompt behavior is driven by the output format itself, or by the intermediate interpretation process encoded in the structured prompt.

Table 7 compares Full Single Prompt with the Schema-only control. Schema-only does not reproduce the behavior of Full Single Prompt. Preserving the structured output format without generating intermediate interpretations leads to substantial F1 degradation for Mistral-7B and Qwen3-4B, together with strongly negative FPR–FNR values. For example, on Challenge-600, Mistral-7B drops from 69.0 to 35.2 F1, and Qwen3-4B drops from 76.1 to 33.6 F1, while FPR–FNR shifts to -76.7 and -77.1, respectively. Similar false-negative-skewed collapses are observed on SemEval-2020 and SemEval-2021.

These results indicate that the effect of Full Single Prompt cannot be explained by the structured output schema alone. If the schema itself were the main driver, Schema-only would be expected to show a judgment pattern closer to Full Single Prompt. Instead, removing the intermediate interpretation content produces a qualitatively different error pattern, often shifting the model toward overly conservative judgments. This suggests that the intermediate interpretation process, rather than the output schema alone, is central to the behavior observed in Full Single Prompt.

We apply this control only to the single-prompt implementation. In the multi-turn Competing Pipeline, intermediate outputs are passed to subsequent stages as inputs, so replacing them with placeholders would invalidate the pipeline rather than isolate an output-schema effect. Thus, Schema-only is used as a single-call control for the structured output format.

C.3 Additional RSE Results

Figure 31 shows the additional RSE results on Challenge-600. Overall, the results are consistent with the SemEval-2021 findings in the main text: sentence-only baselines are sensitive to Targeted Rebuttal (TR), and this sensitivity appears in both F1 and FPR–FNR. The pattern is particularly clear for Llama-3.1-8B. Despite its relatively high initial F1 scores, the model shows substantial drops after TR: under the Standard, Detailed, Long, and Single Prompt conditions, F1 decreases from 0.65, 0.70, 0.67, and 0.66 to 0.22, 0.23, 0.22, and 0.21, respectively. Although F1 increases again after Follow-up Interaction (FI), this pattern should be interpreted as another interaction-induced judgment shift rather than as stable judgment maintenance.

FPR–FNR further shows that TR changes the direction of model errors, rather than merely reducing performance. For Llama-3.1-8B, the Standard, Long, and Single Prompt conditions initially show strong positive FPR–FNR values, but shift to large negative values after TR. This indicates that the model moves from humor over-prediction toward a false-negative-skewed pattern after user rebuttal. A similar shift is observed for Mistral-7B, where the Standard and Single Prompt conditions move from positive FPR–FNR values in the Original stage to negative values after TR. These directional reversals show that rebuttal sensitivity is not limited to isolated label changes, but affects the overall error pattern of the judge.

The response trajectory panel provides a more fine-grained view of these changes. For Llama-3.1-8B, the proportions of *Reverted after TR* and *Non-reverted after TR* are large across most conditions, indicating that many judgments either return to the initial label only after another interaction turn or remain changed after TR. This explains why the apparent F1 recovery after FI should not be treated as evidence of stability. Mistral-7B also shows prompt-dependent trajectories: Minimal and Detailed preserve many judgments through FI, whereas Standard, Long, and Single Prompt exhibit larger proportions of interaction-induced changes. In contrast, Qwen3-4B shows substantially larger *Stable* proportions, and its F1 and FPR–FNR changes are relatively small, especially under the Detailed, Long, and Single Prompt conditions. These results show that high initial F1 does not guarantee rebuttal robustness and that RSE requires trajectory-level analysis in addition to final performance.

Figure 32 shows the additional RSE results on SemEval-2020. Although SemEval-2020 has generally low initial F1 scores, judgment shifts after TR are still clearly observed. Mistral-7B achieves initial F1 scores of 0.33 and 0.34 under the Standard and Single Prompt conditions, respectively, but these scores decrease to 0.14 and 0.16 after TR. Under the Minimal and Detailed conditions, F1 increases after TR; however, this should not be interpreted as stable improvement. Instead, it reflects a shift in judgment criteria under rebuttal, given that the initial judgments are already strongly false-negative-skewed in these conditions.

Changes in FPR–FNR are particularly pro-

Table 8: Decoding variability for the primary IRJ conditions at temperature 0.2. Values are means ± sample standard deviations across three random seeds and are reported in percentage points, with the percentage symbol omitted.

Dataset	Model	Competing		Single Prompt	
		F1	FPR–FNR	F1	FPR–FNR
Challenge-600	Mistral-7B	58.8 ± 0.8	75.5 ± 2.4	68.2 ± 0.1	72.2 ± 2.2
	Llama-3.1-8B	67.4 ± 0.3	46.4 ± 2.0	67.5 ± 0.4	71.5 ± 1.4
	Qwen3-4B	71.9 ± 0.5	41.4 ± 2.7	75.9 ± 1.0	12.2 ± 1.8
SemEval20	Mistral-7B	32.1 ± 0.4	86.6 ± 1.0	33.2 ± 0.1	64.7 ± 1.5
	Llama-3.1-8B	32.8 ± 1.2	-8.6 ± 3.1	34.2 ± 0.5	49.0 ± 1.1
	Qwen3-4B	34.6 ± 0.5	29.8 ± 0.9	36.6 ± 0.6	-12.4 ± 0.7
SemEval21	Mistral-7B	59.4 ± 0.2	83.4 ± 0.5	65.9 ± 0.1	72.7 ± 0.7
	Llama-3.1-8B	69.8 ± 0.4	45.1 ± 0.5	70.6 ± 0.2	57.2 ± 0.6
	Qwen3-4B	76.4 ± 0.2	20.4 ± 0.5	80.6 ± 0.4	2.7 ± 0.2

nounced on SemEval-2020. The Standard and Single Prompt conditions for Mistral-7B initially show large positive FPR–FNR values of 0.86 and 0.65, respectively, but both shift to -0.76 after TR. This indicates that user rebuttal can nearly reverse the model's error direction from a false-positive-skewed pattern to a false-negative-skewed pattern. A similar pattern appears for Llama-3.1-8B: the Standard and Single Prompt conditions shift from 0.93 and 0.49 to -0.89 and -0.84 after TR. Although these values move back in the positive direction after FI, this reflects another judgment shift in response to the user's subsequent stance, rather than stable judgment maintenance.

The response trajectory panel further illustrates this trajectory-level instability. Llama-3.1-8B shows large proportions of *Reverted after TR* and *Non-reverted after TR* across nearly all conditions, while *Stable* remains very limited. This indicates that the model is highly sensitive to both TR and FI, with many judgments moving across interaction turns. Mistral-7B also shows substantial instability under the Standard and Single Prompt conditions, where the *Non-reverted after TR* proportion is especially large. In contrast, Qwen3-4B exhibits relatively stable behavior: its F1 scores mostly remain within a narrow range across Original, After TR, and After FI, and its FPR–FNR changes are relatively limited under the Detailed, Long, and Single Prompt conditions. However, under the Standard condition, FPR–FNR decreases from 0.95 to 0.36 after TR and then increases again to 0.82 after FI, indicating that Qwen3-4B is not completely insensitive to interaction. Overall, the SemEval-2020 results show that RSE captures not only performance degradation, but also error-direction reversals and repeated judgment shifts under user rebuttal.

C.4 Decoding Variability

IRJ uses a temperature of 0.2 when generating intermediate interpretations. To assess variability due to stochastic decoding, we repeat the Competing Pipeline and Single Prompt evaluations with three random seeds. We keep all other settings fixed. Table 8 reports the mean and sample standard deviation of F1 and FPR–FNR for each condition. The standard deviation of F1 across the 18 combinations of models, datasets, and conditions ranges from 0.06 to 1.17 pp, with a median of 0.37 pp. The standard deviation of FPR–FNR ranges from 0.19 to 3.08 pp, with a median of 1.05 pp. The sign of FPR–FNR remains unchanged across the three seeds in all 18 combinations.

For Llama-3.1-8B, Single Prompt produces lower F1 and higher FPR–FNR than Best ZS across all datasets and seeds. For Llama-3.1-8B, Competing Pipeline reduces directional imbalance relative to Best ZS across all seeds on Challenge-600 and SemEval-2020. For Mistral-7B, Single Prompt exceeds Best ZS in F1 across all seeds on Challenge-600 and SemEval-2021. On SemEval-2020, all three seeds are below Best ZS. For Qwen3-4B, Single Prompt reduces directional imbalance relative to Best ZS across all seeds on SemEval-2020 and SemEval-2021. On Challenge-600, directional imbalance is not lower than Best ZS. The F1 ordering relative to Best ZS varies across seeds on SemEval-2021. This analysis measures variability across decoding runs. The paired bootstrap in Section 7 measures uncertainty in metric differences due to evaluation sample composition.

D Bootstrap Confidence Intervals

In Section 7, the paired bootstrap results are summarized as the proportion of comparisons whose

Analysis	Comparison	ΔF1	ΔD	$\Delta\lvert D\rvert$
Prompt	Minimal − Standard	7/9 (4↑, 3↓)	9/9 (9↓)	8/9 (8↓)
Prompt	Detailed − Standard	6/9 (5↑, 1↓)	9/9 (9↓)	9/9 (8↓, 1↑)
Prompt	Long − Detailed	6/9 (1↑, 5↓)	8/9 (5↑, 3↓)	8/9 (7↑, 1↓)
IRJ	Competing − Sentence-only prompts	20/36 (15↑, 5↓)	30/36 (18↑, 12↓)	25/36 (12↑, 13↓)
IRJ	Single Prompt − Sentence-only prompts	21/36 (19↑, 2↓)	33/36 (17↑, 16↓)	31/36 (9↑, 22↓)
IRJ	Single Prompt − Long	5/9 (5↑)	8/9 (4↑, 4↓)	6/9 (6↓)
RSE	After TR − Original	27/36 (2↑, 25↓)	34/36 (5↑, 29↓)	32/36 (11↑, 21↓)
RSE	After FI − After TR	27/36 (23↑, 4↓)	30/36 (30↑)	30/36 (12↑, 18↓)

Table 9: Paired bootstrap validation of prompt, IRJ, and RSE effects. $D = \mathrm{FPR} - \mathrm{FNR}$. Entries count comparisons whose 95% bootstrap confidence intervals exclude zero, with arrows indicating the direction of change. Differences are computed as left minus right. IRJ comparisons pool the four sentence-only prompts across model and dataset pairs. For $|D|$, ↓ indicates reduced directional error imbalance.

95% confidence intervals exclude zero. This section further examines the direction of these changes based on Table 9. For each predefined comparison, we compute the paired bootstrap distribution of the metric difference using the same resampled instances. A change is treated as supported by the bootstrap analysis when the 95% paired bootstrap confidence interval for this difference excludes zero. Its sign follows the subtraction direction shown in the comparison column. Thus, the arrows in the table indicate the direction of the supported change. Here, $D = \mathrm{FPR} - \mathrm{FNR}$, and a downward arrow for $|D|$ indicates that the left condition reduces directional error imbalance relative to the right condition.

The prompt comparisons show consistent changes, but increasing instructional detail does not produce monotonic improvements. In both the Minimal–Standard and Detailed–Standard comparisons, changes in D are supported across all 9 comparisons, and all supported changes are negative. Because each comparison subtracts the Standard prompt from the Minimal or Detailed prompt, this result indicates that the Standard prompt produces higher FPR–FNR values than the Minimal and Detailed prompts. Changes in $|D|$ are supported in 8 of 9 comparisons for Minimal–Standard and in all 9 comparisons for Detailed–Standard, mostly in the decreasing direction. In contrast, the Long–Detailed comparison shows supported F1 changes in 6 of 9 comparisons, most of which are negative. Changes in $|D|$ are supported in 8 of 9 comparisons, mostly in the increasing direction. These results suggest that prompt formulation produces consistent changes in performance and directional error patterns, while longer instruction does not necessarily yield better sentence-only humor judgment.

The IRJ comparisons depend on the model and prompt baseline. When the Competing Pipeline is compared with the full set of sentence-only prompts, supported changes are observed in 20 of 36 comparisons for F1, 30 of 36 for D, and 25 of 36 for $|D|$. Most supported F1 changes are positive, while changes in $|D|$ are mixed. The Competing Pipeline therefore often changes performance without consistently reducing directional error imbalance. Single Prompt shows a more favorable pattern in F1 and $|D|$. Supported F1 changes are observed in 21 of 36 comparisons, with most changes in the positive direction. Changes in $|D|$ are supported in 31 of 36 comparisons, with more decreases than increases. The direct comparison between Single Prompt and Long shows supported F1 improvements in 5 of 9 comparisons and supported reductions in $|D|$ in 6 of 9 comparisons. This suggests that the effect of Single Prompt is not reducible to prompt length alone, since Long is the longest sentence-only prompt. Overall, the results characterize IRJ as a structured judgment framework whose effects on performance and directional error patterns vary across models and prompt baselines.

In the RSE comparisons, a large proportion of changes are supported by the bootstrap analysis, and their directions are also informative. In the After TR–Original comparison, supported changes are observed in 27 of 36 comparisons for F1, 34 of 36 for D, and 32 of 36 for $|D|$. Most supported F1 changes are negative, indicating that Targeted Rebuttal often reduces performance. Changes in D are also mostly negative, indicating that user rebuttals move the error direction toward lower FPR–FNR values and, in many cases, toward judgments with more false negatives. Changes in $|D|$ are also supported in most comparisons. These

results show that TR systematically changes directional error imbalance beyond isolated label flips.

Supported changes are also frequent in the After FI–After TR comparison. F1 changes are supported in 27 of 36 comparisons, D changes in 30 of 36, and $|D|$ changes in 30 of 36. Most supported F1 changes are positive, showing that Follow-up Interaction often partially recovers performance after TR. All supported changes in D are positive, indicating that FI shifts judgments back in the opposite direction after the movement caused by TR. Apparent recovery in F1 or labels after FI can therefore still reflect another shift caused by the subsequent stance of the user.

Overall, the bootstrap results show that prompt formulation produces changes in directional error patterns that remain supported under resampling. The effect of IRJ depends on the model and prompt baseline being compared. Single Prompt often improves F1 and more frequently reduces $|D|$ when compared with the full sentence-only prompt family, including the Long prompt, but the effect is not uniform across all settings. RSE shows repeated shifts across TR and FI. These shifts also remain supported under bootstrap resampling.

E Future Work

Future work can extend this analysis beyond humor to broader subjective and interpretation-dependent judgment tasks, such as sarcasm, indirect speech, pragmatic implication, and opinionated debate. These tasks similarly require evaluators to choose among competing interpretations, rely on implicit context, and remain sensitive to how the judgment frame is presented. Applying prompt sensitivity, structured interpretation, and rebuttal stability analyses to these settings would clarify whether the directional error shifts and trajectory-level instability observed in humor judgment reflect task-specific behavior or a broader challenge in LLM-based subjective evaluation.

F LLM Usage

We used large language models only as assistive tools for polishing, stylistic refinement, and code-writing support. All substantive content, including the problem formulation, methodology, experimental design, analysis, and conclusions, was written and verified by the authors. No unpublished or non-public data was provided to external AI systems. When AI tools suggested text, code, or references, the authors manually inspected, revised, and validated the outputs before inclusion.