OOGIRI-MASTER: Benchmarking Humor Understanding via Oogiri

Anonymous ACL submission

Abstract

Humor is a salient testbed for human-like creative thinking in large language models (LLMs). We study humor using the Japanese creative response game Oogiri, in which participants produce witty responses to a given prompt, and ask the following research question: *What makes such responses funny to humans?* Existing datasets provide only limited signals for funniness. Thus, we introduce OOGIRI-MASTER and OOGIRI-CORPUS, a benchmark and dataset for Japanese Oogiri, where each prompt is paired with approximately 100 candidate responses and funniness is measured by user votes. Using OOGIRI-CORPUS, we analyze linguistic features associated with funniness (e.g., *length*, *ambiguity*, and *incongruity resolution*). Finally, we benchmark a range of LLMs and a human baseline in OOGIRI-MASTER, showing that state-of-the-art models approach the human baseline in accuracy on humor-judgment tasks and that insight-augmented prompting improves accuracy.

1 Introduction

Endowing large language models (LLMs) with human-like creative thinking capabilities is a major challenge that extends beyond problem-solving abilities. Humor understanding is one of such key capabilities. Understanding and generating humor as humans require more than pattern matching; they necessitate creative reasoning that incorporates context and cultural nuances to produce witty and unexpected responses (Loakman et al., 2025). This study addresses humor as an instance of creative thinking in LLMs by focusing on the specific case of *Oogiri* (大喜利). *Oogiri* is a Japanese creative response game that involves improvising humorous responses to a given prompt, as shown in Figure 1, making it an ideal testbed for creativity and wit. This raises the central question:

Prompt: Worst commit message ever.
Response: "It works on my machine."

Figure 1: Oogiri prompt–response example.

What exactly makes Oogiri responses funny to humans? The starting point of our study is to answer this question. Few studies have aimed to capture the human perception of funniness using objective metrics and to analyze its components quantitatively. This absence poses a significant barrier to the evaluation of humor understanding in LLMs.

We address two key challenges in evaluating the humor understanding of LLMs. First, the constituent elements of a funny response remain insufficiently understood. Humor is a subjective construct arising from a complex interplay of factors such as the violation of expectations and resonance. However, an objective, quantitative metric does not exist for measuring funniness itself. Consequently, we lack a principled basis for explaining why an Oogiri-style response is funny, which hinders the systematic improvement of LLM humor understanding. The second challenge is the low reliability of existing datasets for such analysis. For example, the Oogiri-GO dataset (Zhong et al., 2024) was collected from Bokete,[1] a caption-contest platform on which users upvote funny responses to prompts. Although this social-voting signal is useful at this scale, it introduces two methodological limitations. First, the fairness of the evaluation process is not guaranteed: making the popularity of each response visible to other raters may introduce popularity bias and compromise objectivity. Second, the dataset exhibits structural bias. With only approximately eight candidate responses per prompt, on average, raters are likely to select *a relatively better* option rather than an intrinsically humorous one.

Therefore, in this study, we propose OOGIRI-

[1] https://bokete.jp/

MASTER, a benchmark that evaluates the humor understanding of LLMs using the Oogiri task. Specifically, we address the two challenges outlined above by constructing a novel dataset and conducting a quantitative analysis of the funniness components, with which we assess the current capabilities and pave the way for improvements. First, we construct OOGIRI-CORPUS, a dataset that ensures reliability and objectivity.[2] On average, each prompt is paired with approximately 100 diverse candidate responses, and funniness is measured via user votes on a platform that hides vote counts during voting. This design mitigates the issues of fairness and data bias observed in existing datasets. Second, using this dataset, we quantitatively analyze the linguistic features that constitute funniness. We identify common lexical and structural patterns in high-rated responses, transforming the ambiguous notion of funniness into measurable, objective metrics. This enables explanations of *why a response is funny* based on data-driven evidence, rather than subjective intuition. Finally, we present the multifaceted benchmark results on OOGIRI-MASTER. We benchmark humans and various LLMs to clarify the current state of the art in the humor understanding of LLMs.

The contributions of this study can be summarized as follows:[3] First, we constructed and release a large-scale reliable dataset, OOGIRI-CORPUS, which serves as a novel foundation for evaluating humor understanding in LLMs. Second, through quantitative analysis of this dataset, we identified the constituent components of funniness, demonstrating that features such as *response length*, *perspective shift*, and *ambiguity* are strongly correlated with high-rated responses. Third, we propose a novel benchmark, OOGIRI-MASTER, and experimentally demonstrated that (1) state-of-the-art LLMs such as GPT-5 show performance approaching human performance; (2) our analytical insights into the constituent components of humor can contribute to performance improvements in humor judgment; (3) instructing LLMs to leverage these insights only when uncertain improves their performance; and simultaneously, (4) continued pretraining on the target-language corpus enhances performance on Oogiri humor-judgment tasks.

[2]We distinguish the dataset, OOGIRI-CORPUS, which underpins our analyses, from the benchmark, OOGIRI-MASTER, which builds on it to evaluate LLMs.

[3]The dataset and the benchmark will be provided under the CC BY-NC-SA 4.0 license upon acceptance.

2 Related Work

Background on Computational Humor Computational humor is a relatively new area, and humor understanding/generation remains a challenging problem in natural language processing (Loakman et al., 2025). One obstacle is defining "humor" appropriately. Consequently, many studies have narrowed the scope to specific forms (e.g., puns, Oogiri, satire) to make the problem tractable (Amin and Burghardt, 2020). Among these, pun generation has a particularly long history and is a central task (Ritchie, 2005; Yu et al., 2018; Luo et al., 2019)

Oogiri as a Testbed for Humor Understanding We target Oogiri as our testbed for humor understanding. Oogiri is a creative response game in which one provides a witty response to a prompt. Although the most common setup is a text-to-text format in which a textual prompt is paired with a textual response, modal variants exist (e.g., image-to-text one-liners; image&text-to-text fill-in-the-blank) (Zhong et al., 2024). These formats resemble *memes* (Sharma et al., 2023; Nguyen and Ng, 2024); we regard memes as a multimodal variant of Oogiri. However, we focus on text-to-text Oogiri for two reasons. First, abundant web resources exist. Oogiri is widely popular in TV programs and social media, and large platforms such as Bokete and Oogiri Sogo host substantial data. Because analyzing humor components requires diverse and numerous samples, Oogiri is suitable from a data perspective. Second, the text-to-text format is unimodal, making semantic understanding more straightforward than with multimodal variants.

Existing Oogiri Datasets and Their Limitations Although progress has been hampered by limited datasets, interest has recently increased with the advent of LLMs and the concomitant need for evaluation resources. Oogiri-specific datasets remain relatively scarce; adjacent resources include English caption datasets collected from the New Yorker Caption Contest (Hessel et al., 2023) and various meme datasets (Liu et al., 2022; Hwang and Shwartz, 2023; Hossain et al., 2022). Oogiri-GO, which was built using Bokete and social media, is a representative Oogiri dataset. However, it faces two issues: (1) fairness concerns: Voter interfaces display others' popularity, inviting conformity and potentially compromising objectivity. (2) structural bias: Many prompts have few candidate responses

(approximately eight on average); hence, raters may select responses that are merely "less bad," rather than intrinsically funny. In this study, we construct a novel Oogiri dataset, OOGIRI-CORPUS, which addresses these issues and serves as a foundation for evaluating LLM humor understanding, thereby improving reliability.

Quantitative Analyses of Humor Components Although studies have been conducted on generation, understanding, and explanation in computational humor (Amin and Burghardt, 2020; Loakman et al., 2025), quantitative analyses of the constituent components of "funniness" remain underexplored. To fill this gap, using OOGIRI-CORPUS, we analyze how diverse linguistic features, such as perspective shift, ambiguity, surprisal, sentence length, and part-of-speech (POS) ratios, relate to humor, with the aim of identifying objective, quantitative indicators. Furthermore, using our benchmark experiments, we outline how these insights can improve LLM humor understanding.

3 Dataset Construction

Motivated by the second challenge mentioned in §1, we present OOGIRI-CORPUS and provide details on its construction process and descriptive statistics. We collected data from a public Japanese Oogiri competition platform, Oogiri Sogo[4]. On this platform, each prompt proceeds through an answer phase, a voting phase, and a final leaderboard announcement. During the answer phase, users submit responses within a fixed time window (e.g., 12 h). This phase then transitions to the voting phase, in which users vote for the responses that they find funny among all submissions. Each user casts up to three votes per prompt, and can assign multiple votes to the same response. Thus, users do not provide an exhaustive rating for every response; instead, they allocate a small number of votes to selected responses. Unlike other platforms (e.g., Bokete), vote counts are not displayed during the voting phase, which helps to mitigate popularity bias and supports fairer evaluation. Finally, the platform announces a leaderboard based on the total votes.

Dataset construction comprised two steps: web crawling[5] and quality filtering. First, we collected 2,165 prompts from the platform.[6] Each prompt is associated with many responses, and each response has a vote count indicating its perceived funniness. We applied vote-based filtering to ensure reliability: we excluded prompts for which the total number of votes was fewer than 100. This threshold reduces the variance owing to rater subjectivity and chance when the vote pool is small. In total, 908 prompts remained. We refer to this 908-prompt dataset as OOGIRI-CORPUS, and used it for the subsequent analyses and benchmark construction.

Statistic	Value
Prompts	908
Responses per prompt (avg.)	95.9
Votes per prompt (avg.)	171.6
Votes per response (avg.)	1.8
Votes per top-1 response (avg.)	10.3
Prompt length in characters (avg.)	20.4
Response length in characters (avg.)	16.4

Table 1: Summary statistics of OOGIRI-CORPUS.

OOGIRI-CORPUS consists of prompts, responses, and vote counts. Across the 908 prompts, each prompt has approximately 96 responses and 172 votes, on average. The total number of prompt–response pairs is 82,536. This is approximately seven times larger than that of Oogiri-GO (Zhong et al., 2024) and, to the best of our knowledge, is the largest Japanese Oogiri dataset to date.[7] Moreover, although Oogiri-GO averages approximately eight responses per prompt, our dataset offers approximately 96 responses, yielding a far more diverse candidate set per prompt. This breadth enables raters to select responses that are genuinely funny rather than merely "less bad" within a limited pool. Dataset statistics are presented in Table 1.

4 Linguistic Feature Analysis

We address the first challenge mentioned in §1: elucidating the components that constitute a "funny response." "Funniness" is subjective and complex; for example, it involves expectation violations and relatability. However, a generally accepted quantitative metric remains lacking. Accordingly, our analysis aims to explain and analyze why an Oogiri response is funny based on a variety of quantitative linguistic features. Through this analysis, we seek to identify objective and quantitative indicators for understanding humor and to pave the way for improving the ability of LLMs to understand humor.

[4] https://chinsukoustudy.com/

[5] The site explicitly permits web crawling.

[6] Prompt IDs 87–2254 were available when accessed.

[7] Compared with 11,842 Japanese Oogiri instances in a text-to-text setting.

4.1 Dataset for Analysis

We quantitatively examined the linguistic features that constitute "humor," using OOGIRI-CORPUS as the foundation. Although the dataset links an average of 96 responses to each prompt, we did not use all responses for the analysis. This is because many responses have zero votes, creating a pronounced imbalance between high-rated responses with many votes and low-rated responses with no votes, which makes the analysis challenging.

Accordingly, we first narrowed down the responses under analysis and balanced the high- and low-rated responses. Specifically, for each prompt, we defined the top three responses by vote count as "high-rated responses" and the bottom three as "low-rated responses." On average, high-rated responses received approximately 8.5 votes, whereas all low-rated responses had zero votes. Given this low-rated nature, we considered them as reasonable representatives of "unfunny responses." This yielded 5,448 responses for the analysis, with 908 prompts $\times$ 6 responses.

4.2 Analysis Methodology

We examined the relationships between linguistic features and response humor. Specifically, for each response, we quantitatively measured a range of linguistic features and analyzed the relationship of these feature values to response humor (i.e., differences between the high- and low-rated groups). We defined and quantified various aspects of linguistic features by borrowing ideas from theories of humor, such as incongruity theory (Morreall, 2024). These include basic linguistic features, such as sentence length, as well as higher-order features, such as resolution of incongruity (see details in §4.3). We considered that, when a feature exhibits a significantly higher or lower value in high-rated responses, it may constitute a component of humor.

We reported these relationships using an independent two-sample Student's t-test (two-sided, assuming equal variances) (Fisher, 1925) and Cohen's d (Cohen, 1988); we reported Cohen's d as an effect-size measure because the t-test is sensitive to large sample sizes. The conventional benchmarks interpret $d = 0.2$, 0.5, and 0.8 as small, medium, and large effects, respectively. The formula and notation for Cohen's d are given in Appendix A.

4.3 Linguistic Features

To capture humor from multiple perspectives, we defined 26 linguistic features,[8] organized into four groups: 11 basic features (seven response-independent and four prompt–response relative), four semantic/textual entailment features, three surprisal/PMI features, and eight LLM-scored higher-order features (Table 2), and measured them quantitatively. Inspired by the theories of humor (Morreall, 2024) and prior research on humor and other creative domains (Zhong et al., 2024; Murakami et al., 2025), we selected these features as plausible constituents of humor. Precise definitions and computation details for all features, including the prompt templates and models used, are provided in Appendix B.

Basic Linguistic Features We defined basic linguistic features comprising (i) response-independent measures computed from the response alone (e.g., length, character-type and POS ratios) and (ii) prompt–response relative measures computed by comparing each response with its prompt (e.g., length ratio, lexical novelty, character-type changes).

Semantic Distance and Textual Entailment Inspired by incongruity theory (McDonald, 2013; Morreall, 2024), we used (i) semantic distance and (ii) textual entailment. We introduced these features to quantify how far a response departs semantically from the prompt and whether it preserves, contradicts, or reframes the prompted situation.

Surprisal and Pointwise Mutual Information We added surprisal (Shannon, 1948) and normalized PMI (nPMI) (Fano, 1961) to capture deviation from expectation. We introduced these measures to quantify unexpectedness both in the response itself and in the strength of association between the prompt and response.

LLM-Scored Higher-Order Features To capture higher-order cues beyond surface cues (e.g., length) and probabilistic or embedding-based signals (e.g., surprisal), we scored each prompt–response pair with GPT-5 on a 1–5 scale across eight higher-order aspects (Table 2): ambiguity exploitation, associative distance, benign violation (McGraw and Warren, 2010), coherence, expectedness, incongruity resolution (Ritchie, 2009),

[8] We do not explicitly model or annotate culture-dependent mechanisms (e.g., references to social norms, historical events, or subcultural knowledge) that can shape intended meaning and audience interpretation; see Limitations.

Feature names	Feature values		Cohen's d
	High	Low	
Basic Features			
Response-independent			
length*†	14.12	16.40	**-0.28**
unique chars*†	13.24	15.32	**-0.30**
hiragana ratio*	0.46	0.44	0.11
katakana ratio*†	0.14	0.16	-0.11
noun ratio*	0.42	0.45	-0.13
verb ratio*	0.16	0.14	0.10
symbol ratio*†	1.91	2.24	-0.07
Prompt-response			
length ratio*†	0.76	0.90	**-0.27**
lexical novelty*	0.80	0.93	**-0.21**
hiragana changes*	-0.04	-0.06	0.10
katakana changes*	0.02	0.05	-0.10
Semantic / NLI			
semantic distance*	0.73	0.72	0.16
contradiction*	0.28	0.27	0.06
entailment*	0.17	0.14	0.18
neutral*	0.55	0.59	-0.15
Surprisal / PMI			
nPMI*	0.12	0.14	-0.14
surprisal$^{*}_{\text{response-independent}}$	5.17	5.08	0.08
surprisal$^{*}_{\text{prompt-response}}$	4.66	4.51	0.13
LLM-Scored Features			
ambiguity exploitation*†	2.10	1.61	**0.42**
associative distance*†	4.38	3.90	**0.33**
benign violation*†	4.73	4.49	**0.27**
coherence*	4.11	3.95	0.15
expectedness	2.68	2.78	-0.08
incongruity resolution*†	3.71	3.35	**0.36**
metaphor use*†	1.54	1.31	**0.24**
perspective shift*†	2.40	1.87	**0.50**

Table 2: Comparison of linguistic features between high- and low-rated responses. $*$ indicates statistical significance ($p < 0.05$). Bold values in the Cohen's d indicate a small or medium effect size ($|d| \geqq 0.2$). † indicates features that are employed in the benchmark experiments (§5).

metaphor use, and perspective shift. We expected these scores to provide an interpretable characterization of the components of funniness.

4.4 Analysis Results

We report on the relationships between each linguistic feature and response humor. Table 2 presents the mean of each feature for the high- and low-rated groups, p-value of the t-test, and Cohen's d. Our analysis yielded the following findings:

High-Rated Responses Tend to be Shorter Length-related features such as the length and prompt–response length ratios were significantly lower in the high-rated group than in the low-rated group, with small effect sizes. This suggests that brevity contributes to humor.

Appropriate Vocabulary Diversity is Beneficial Interestingly, the high-rated group showed significantly lower values for the unique character count (unique chars) and the rate at which vocabulary that is not in the prompt appears in the response (lexical novelty), with small effect sizes. This indicates that, relative to the low-rated group, high-rated responses had a lower tendency to use new vocabulary and may benefit from selecting appropriate words without straying far from the prompt.

Higher-Order Linguistic Features are Effective Ambiguity exploitation, associative distance, benign violation, incongruity resolution, metaphor use, and perspective shift were significantly higher in the high-rated group, with small-to-medium effect sizes. Among these, perspective shift and ambiguity showed relatively larger effects, indicating particular importance for humor. Incongruity resolution, grounded in incongruity-resolution theory (Ritchie, 2009), also showed a relatively large effect size, suggesting its contribution to humor.

Other Features Have Limited Impact Semantic distance, textual entailment, surprisal, nPMI, and other linguistic features (e.g., POS ratio) showed statistically significant differences, but the effect sizes were below small, suggesting limited contributions to humor. Notably, textual entailment and surprisal captured similar aspects to coherence and expectedness in higher-order linguistic features, but their effect sizes were below small, consistently suggesting their limited role in constituting humor.

5 Oogiri Understanding Benchmark

We propose a novel benchmark, OOGIRI-MASTER. The aim of this benchmark is to measure the ability of an LLM to understand and judge "humor" in Oogiri from different perspectives. Specifically, we propose five tasks that can be broadly grouped into two categories: four relative-judgment tasks using multiple-choice question answering (MCQA) and one absolute-judgment task using binary classification. Standardized prompt templates and strict evaluation criteria were used to ensure reproducibility and comparability. In the experiments, we tested the insights from our analysis results in §4 and reflected the multiple linguistic features into prompt templates, seeking the performance gains of LLMs (§5.3). Our goal was to clarify the current state of LLM humor understanding and outline a path for further improvement.

5.1 Task Design

Relative Judgment Tasks In the MCQA setting, the model selects the most humorous response to a given prompt from several candidate responses. We defined four types of tasks: two binary-choice tasks, a three-choice task, and a four-choice task. In all tasks, the high-rated response for each prompt served as the positive example, and the negatives were constructed differently for each task. For the two binary-choice tasks, we constructed negatives in two ways: (i) we paired the positive with one low-rated response from the same prompt ($\text{Binary}_{\text{same}}$) and (ii) we paired the positive with one high-rated response for a different prompt ($\text{Binary}_{\text{diff}}$). The latter evaluates whether the model can judge funniness as a response to the given prompt, rather than merely ranking responses within the same prompt, following Hessel et al. (2023). For the three- and four-choice tasks, we used one low-rated same-prompt response and one or two high-rated different-prompt responses as negatives, respectively.

Absolute Judgment Task In the binary classification setting, the model decides whether a response to a prompt is "funny" or "not funny." For each prompt, we used the high-rated response as the positive and the low-rated response as the negative, measuring the ability of the model to evaluate funniness in absolute terms. Prompt examples for the relative- and absolute-judgment tasks are provided in Appendix C.

5.2 Dataset Construction

OOGIRI-MASTER is built on OOGIRI-CORPUS. For the MCQA setting, we sampled 100 prompts per task from OOGIRI-CORPUS, and selected positives and negatives according to each task design, yielding 400 items across the four tasks. For binary classification, we sampled 100 prompts from OOGIRI-CORPUS, pairing one high-rated response and one low-rated response per prompt for 200 items. In total, OOGIRI-MASTER comprised 600 items.[9]

5.3 Benchmark Experiments

5.3.1 Experimental Setup

We evaluated a range of LLMs listed in Table 3, from proprietary (e.g., GPT-5) to open-source (e.g., DeepSeek-R1), on five tasks in OOGIRI-MASTER. We report the accuracy as an evaluation metric. For API-based models, we averaged results over three trials. During inference, we set the temperature parameter to zero for all models.

We compared two prompting strategies when instructing the LLMs to solve each task. (1) a *baseline prompt* that simply instructs the model to select options, (2) an *insight-augmented prompt* that incorporates features computed from given prompt–response pairs based on the findings of our data analysis. For reproducibility, we provide the prompt templates in Appendix C. To keep the prompts concise, we included only a small set of features selected with reference to the observed effect sizes in Table 2. Specifically, we used five basic features: length, unique character count, prompt–response length ratio, symbol ratio, and katakana ratio; and six LLM-scored features: ambiguity exploitation, associative distance, benign violation, incongruity resolution, metaphor use, and perspective shift. The basic features were precomputed and inserted directly into the prompt. LLM-scored features followed a two-step procedure: first, for each prompt–response pair, the target LLM computed scores for each aspect (e.g., metaphor use); second, these scores were included as context when instructing the model to select the options for each task.

To validate the human performance on this benchmark, we recruited Japanese-speaking crowdworkers from Yahoo! Crowdsourcing, a major Japanese crowdsourcing platform[10] and asked them to solve each item using the same baseline prompt that was shown to the LLMs. Each item was answered by 21 workers, and the final labels were determined by majority vote. We included attention checks with unambiguous answers and aggregated the results only for the 21 workers who passed the checks for each item. Inter-annotator agreement among the 21 crowdworkers, measured by Fleiss' κ, is reported in Appendix C.3.

5.3.2 Results and Discussion

Table 3 lists the benchmark results. We compared two prompting strategies: a baseline prompt and an insight-augmented prompt.

Baseline Prompt When averaging the accuracy across the five tasks, Claude-Opus-4 performed the best (68.7%), followed by GPT-5 (67.6%) and

[9] To prevent data contamination, we sampled different data points from the analysis dataset in §4.

[10] https://crowdsourcing.yahoo.co.jp/. See Appendix C.2 for the task instructions shown to crowdworkers.

Models	Features	Absolute.	Relative.				Ave.	ΔAve.
		$\text{Binary}_{\text{class}}$	$\text{Binary}_{\text{diff}}$	$\text{Binary}_{\text{same}}$	Triple	Quad	Accuracy	Accuracy
Open LLMs								
gpt-oss-20b	–	50.5	64.0	45.0	33.0	37.0	45.9	–
gpt-oss-20b	✓	54.0	52.0	57.0	27.0	22.0	42.4	-3.5
DeepSeek-R1-14b	–	48.5	56.0	43.0	31.0	28.0	41.3	–
DeepSeek-R1-14b	✓	46.0	57.0	49.0	24.0	31.0	41.4	+0.1
$\text{DeepSeek-R1-14b}_{\text{ja}}$	–	52.0	61.0	42.0	38.0	30.0	44.6	–
$\text{DeepSeek-R1-14b}_{\text{ja}}$	✓	50.0	59.0	53.0	44.0	24.0	46.0	+1.4
$\text{LLM-jp-3.1-13b}_{\text{ja}}$	–	47.0	80.0	45.0	39.0	38.0	49.8	–
$\text{LLM-jp-3.1-13b}_{\text{ja}}$	✓	50.5	58.0	45.0	30.0	28.0	42.3	-7.5
Proprietary LLMs								
Claude-Opus-4	–	57.2	83.0	**70.0**	63.0	**70.3**	68.7	–
Calude-Opus-4	✓	50.8	72.7	68.0	53.0	51.3	59.2	-9.5
Gemini-2.5-Pro	–	51.3	62.0	61.7	46.3	45.7	53.4	–
Gemini-2.5-Pro	✓	50.8	58.7	66.3	51.3	47.0	54.8	+1.4
GPT-5	–	**61.7**	89.7	65.3	62.3	59.0	67.6	–
GPT-5	✓	60.0	93.3	69.0	**69.0**	62.0	**70.7**	**+3.1**
human	–	54.5	**95.0**	59.0	67.0	68.0	68.7	

Table 3: Results of benchmark experiments. The best results for each column are **bolded**. "Ave. Accuracy" indicates the average accuracy (%) across five tasks, and "ΔAve. Accuracy" indicates the difference in average accuracy (%) when using features from our analysis.

Gemini-2.5-Pro (53.4%). Open LLMs lagged behind these proprietary LLMs; even the strongest, $\text{LLM-jp-3.1-13b}_{\text{ja}}$, reached only 49.8%. Additionally, with the same instructions as those provided to the LLMs, the 21 crowdworkers achieved 68.7%, which is comparable to that of Claude-Opus-4. One possible reason that the human performance was relatively low compared with our expectations is the demographic mismatch between crowdworkers and users of the Oogiri platform.[11] Humor is subjective, and differences in age and interests can yield different judgments of funniness. Future studies will include analyses that account for annotator attributes and evaluations using more diverse raters.

Insight-Augmented Prompt With feature incorporation, four models, namely GPT-5, Gemini-2.5-Pro, DeepSeek-R1, and $\text{DeepSeek-R1}_{\text{ja}}$, improved their average accuracy across the five tasks. Notably, GPT-5 increased from 67.6% to 70.7% (+3.1%), surpassing both human performance and Claude-Opus-4 in the baseline setting. This supports the effectiveness of the linguistic features that reflect the components of humor in improving Oogiri understanding. However, three models, namely Claude-Opus-4, gpt-oss-20b, and $\text{LLM-jp-3.1-13b}_{\text{ja}}$, degraded. One possible factor is differences in the reasoning ability. Compared with the baseline, the insight-augmented prompt was longer and more complex because of the added features and instructions. Stronger reasoners (e.g., GPT-5) could correctly interpret these complex prompts and benefits, whereas weaker models (e.g., $\text{LLM-jp-3.1-13b}_{\text{ja}}$) tended to misinterpret them and overrely on feature magnitudes. For example, given the insight that funnier responses tend to be shorter, weaker models over-selected very short responses. This suggests that when reasoning is limited, instructing models to consider features can introduce overfitting problems and reduce performance.

5.3.3 Analysis

Effectiveness of Continued Pretraining on Japanese Corpus We compared the two models in Table 3, namely DeepSeek-R1 and $\text{DeepSeek-R1}_{\text{ja}}$, which share the same architecture and parameter count; the only difference is the pretraining data. $\text{DeepSeek-R1}_{\text{ja}}$ continues pretraining DeepSeek-R1 on a Japanese corpus.[12] $\text{DeepSeek-R1}_{\text{ja}}$ improved the average accuracy across the five tasks from 41.3% to 44.6% in the baseline setting (+3.3 points) and from 41.4% to 46.0% in the insight-augmented setting (+4.6 points). As our benchmark is based on Japanese Oogiri, these results suggest that continued pretraining on a Japanese corpus can improve performance on Oogiri humor-judgment tasks. Although prior work

[11] Because neither the crowdsourcing service nor the Oogiri platform discloses detailed user attributes, we could not perform a precise comparison; however, some differences in user populations are plausible.

[12] https://huggingface.co/cyberagent/DeepSeek-R1-Distill-Qwen-14B-Japanese

Models	Features		Ave.	ΔAve.
	Basic.	LLM-scored.	Acc.	Acc.
Gemini-2.5-Pro	–	–	53.4	–
Gemini-2.5-Pro	✓	–	**57.1**	**+3.7**
Gemini-2.5-Pro	–	✓	54.5	+1.1
Gemini-2.5-Pro	✓	✓	54.8	+1.4
GPT-5	–	–	67.6	–
GPT-5	✓	–	69.8	+2.2
GPT-5	–	✓	68.0	+0.4
GPT-5	✓	✓	**70.7**	**+3.1**

Table 4: Ablation study on feature types. "ΔAve. Acc." represents the difference in average accuracy compared to the model without any features. The best accuracy for each model is **bolded**.

Models	Features	Uncertain	Ave. Accuracy
GPT-5	–	–	67.6
GPT-5	✓	–	68.9
GPT-5	✓	✓	70.7

Table 5: Ablation study on instruction styles. "Uncertain" indicates whether the uncertain instruction style is used. "Ave. Accuracy" indicates the average accuracy (%) across five tasks.

has shown that continued pretraining on Japanese corpora improves performance on Japanese cultural-knowledge benchmarks (e.g., knowledge of folktales) (Tsutsumi and Jinnai, 2025), our findings suggest that such continued pretraining also helps with the higher-level language understanding needed to judge funniness in Japanese Oogiri.

Ablation Study of Feature Groups Table 4 presents the average accuracy over the five tasks for GPT-5 and Gemini-2.5-Pro under four settings: introducing only basic linguistic features, introducing only LLM-scored higher-order features, introducing both, and using the baseline with no features. In all cases, incorporating features into a prompt improved the average accuracy over the baseline. For GPT-5, using both feature groups yielded the best results. For Gemini-2.5-Pro, introducing only basic linguistic features (e.g., length and character-type ratios) performed the best. Notably, when introducing only basic linguistic features, both Gemini-2.5-Pro and GPT-5 improved more than when introducing higher-order features alone (e.g., +3.7 and +2.2 points, respectively). Response length was already identified in our analysis as a constituent component of humor, and the benchmark results empirically confirm that such simple heuristics can be effective criteria for evaluating funniness. These findings suggest that exploring a broad range of linguistic features is a promising direction for enhancing the humor understanding of LLMs further.

Effect of Instruction Style for Feature Use We also examined the influence of instruction style on performance when incorporating features into prompts, that is, how we should tell the model to use the features. We considered two styles: (1) instructing the model to use the features when judging funniness, and (2) instructing the model to consult the features *only when uncertain*. In our preliminary experiments, we first attempted style (1) and observed an over-reliance on feature magnitudes, which motivated the proposal of style (2). Table 5 shows the average accuracy of GPT-5 over the five tasks for the no-feature baseline and the two instruction styles. Here, the "Uncertain" column corresponds to style (2). In both styles, incorporating features improved over the baseline; notably, style (2) yielded the highest performance, improving the average accuracy by 3.1 points over the baseline. This indicates that asking the model to consider features only when uncertain helps to prevent over-dependence on feature magnitudes and enables more appropriate use of the features. The results highlight instruction design as an important lever for improving the humor understanding of LLMs, and the value of exploring more effective instruction styles in future studies.

6 Conclusion

We presented a systematic study of humor on OOGIRI-CORPUS, and introduced OOGIRI-MASTER, a benchmark covering relative and absolute judgments. Our analysis showed that multiple linguistic features, such as length and ambiguity, correlated with high-rated responses. In the benchmark experiments, we showed that incorporating these features into prompts improves average accuracy across the five benchmark tasks. Furthermore, we demonstrated that continued pretraining on a Japanese corpus further boosts accuracy and instructing models to consider features only when uncertain mitigates over-reliance on heuristics. Future work will include exploring other effective linguistic features and refining prompt design, scaling human evaluations with annotator attributes, and extending the method to other languages and multimodal settings.

Limitations

Limited to Japanese Oogiri Our analysis and benchmark are based on Japanese Oogiri data. Some humor depends on culture-specific knowledge (e.g., a response such as "Mount Fuji" may be funny to Japanese users because it evokes familiar shared knowledge), and similar effects may not hold in other languages or cultural contexts. We do not explicitly model or annotate culture-dependent aspects (e.g., references to social norms, historical events, or subcultural knowledge) that can shape intended meaning and audience interpretation. Therefore, our findings should be interpreted as characterizing humor judgments within the target Japanese Oogiri community rather than universal humor understanding. Moreover, our feature analysis included Japanese-specific elements (e.g., character-type ratios), which may not be directly transferred. Future work should include collecting and analyzing Oogiri-like data in other languages and cultures to better understand the cross-lingual and cross-cultural variations in humor.

Vote-Based Measurement and Sparsity Our dataset uses user votes as a proxy for perceived funniness. Voting is non-exhaustive (each user allocates only a small number of votes), and many responses receive zero votes; thus, low vote counts can reflect not only low funniness but also limited exposure or other platform dynamics that we cannot fully observe. Although we apply filtering (e.g., minimum total votes per prompt) to reduce variance, residual biases may remain.

Unknown Demographics and Potential Mismatch The platform does not provide detailed demographic information about users or voters, which prevents us from analyzing how preferences vary across groups (e.g., age and gender) and from directly controlling for demographic biases in voting. In addition, our crowdworker-based human baseline may differ from the platform user population, which can affect human–model comparisons.

Dependence on Specific Models and Prompts for Feature Measurement Several features depend on specific models and prompts (e.g., LLM-scored higher-order features) or on specific language model and tokenization choices (e.g., surprisal and nPMI). These design choices may influence feature values and downstream analyses.

Benchmark Scope Limited to Oogiri Understanding We proposed a benchmark focused on understanding "funniness" in Oogiri: four MCQA subtasks and one binary classification task. However, humor understanding is related to other capabilities such as generation and explanation (Loakman et al., 2025). Although these are beyond the scope of this study, extending the benchmark to evaluate generation and explanation is an important direction for future research.

Focus on Unimodal Settings As discussed in Related Work (§2), Oogiri can be framed as text-to-text, image-to-text, or image&text-to-text (Zhong et al., 2024). We focused on the text-to-text approach for two reasons: (1) as a first step toward measuring LLM humor understanding, a unimodal text-only setup reduces complexity relative to multimodal settings, and (2) text-to-text Oogiri data are more abundant on the web, facilitating robust dataset construction and generalizable analysis. An important next step is to extend the dataset to multimodal variants and study humor understanding involving visual information.

Ethical Considerations

Data Collection and Licensing OOGIRI-CORPUS was constructed by collecting data from the public Japanese Oogiri competition platform, Oogiri Sogo. We confirm that the site explicitly permits web crawling, ensuring the legitimacy of the data collection process in §3. To promote transparency and facilitate further research, OOGIRI-CORPUS and OOGIRI-MASTER will be made available under the CC BY-NC-SA 4.0 license.

Privacy and Content Safety The released versions of OOGIRI-CORPUS and OOGIRI-MASTER will not contain any metadata that names or uniquely identifies individuals (e.g., usernames or profile URLs). Prior to release, we will additionally screen the text for potential personal identifiers and remove them if found.

Human Evaluation on OOGIRI-MASTER We recruited Japanese-speaking crowdworkers for human baseline evaluation in §5. We used Yahoo! Crowdsourcing, a major Japanese crowdsourcing platform. In accordance with the platform's regulations, the compensation was set at 10 yen per 20 tasks. Workers were informed that the annotated results would be used for research purposes. In addition, we acknowledge that a

potential demographic mismatch between the crowdworkers and Oogiri-platform users exists as discussed in §5.3.2, suggesting that a further analysis accounting for annotator attributes is necessary to improve the evaluation reliability.

Use of AI Assistance We used an AI assistant during manuscript revision for language editing and LaTeX formatting. All changes were reviewed and validated by the authors. The AI assistant was not used to generate the dataset, human annotations, or experimental results.

Potential Risks and Mitigations OOGIRI-CORPUS and OOGIRI-MASTER are intended for research on humor judgment in Japanese Oogiri. Since humor is subjective and culturally contingent, using these resources beyond this scope (e.g., other languages, cultures, or high-stakes settings) may lead to misleading or unfair evaluations. We mitigate this risk by restricting our claims to the target setting, documenting key limitations, and recommending research-only use with human oversight.

References

Miriam Amin and Manuel Burghardt. 2020. A survey on approaches to computational humor generation. In *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 29–41, Online. International Committee on Computational Linguistics.

J. Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

R.M. Fano. 1961. *Transmission of Information: A Statistical Theory of Communication*. MIT Press Classics. MIT Press.

R.A. Fisher. 1925. *Statistical methods for research workers*. Edinburgh Oliver & Boyd.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. MemoSen: A multimodal dataset for sentiment analysis of memes. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1542–1554, Marseille, France. European Language Resources Association.

EunJeong Hwang and Vered Shwartz. 2023. MemeCap: A dataset for captioning and interpreting memes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1445, Singapore. Association for Computational Linguistics.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Barcelona, Spain. Association for Computational Linguistics.

Moritz Laurer, Wouter van Atteveldt, Andreu Salleras Casas, and Kasper Welbers. 2022. Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT - NLI. *Preprint*. Publisher: Open Science Framework.

Chen Liu, Gregor Geigle, Robin Krebs, and Iryna Gurevych. 2022. FigMemes: A dataset for figurative language identification in politically-opinionated memes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7069–7086, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tyler Loakman, William Thorne, and Chenghua Lin. 2025. Who's laughing now? an overview of computational humour generation and explanation. *Preprint*, arXiv:2509.21175. Preprint, arXiv:2509.21175.

Fuli Luo, Shunyao Li, Pengcheng Yang, Lei Li, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. Pun-GAN: Generative adversarial network for pun generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3388–3393, Hong Kong, China. Association for Computational Linguistics.

P. McDonald. 2013. *The Philosophy of Humour*. Philosophy Insights. HEB Humanities E-Books.

A Peter McGraw and Caleb Warren. 2010. Benign violations: making immoral behavior funny: Making immoral behavior funny. *Psychol. Sci.*, 21(8):1141–1149.

John Morreall. 2024. Philosophy of Humor. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Fall 2024 edition. Metaphysics Research Lab, Stanford University.

Soichiro Murakami, Peinan Zhang, Hidetaka Kamigaito, Hiroya Takamura, and Manabu Okumura. 2025. AdParaphrase v2.0: Generating attractive ad texts using a preference-annotated paraphrase dataset. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15212–15230, Vienna, Austria. Association for Computational Linguistics.

Khoi P. N. Nguyen and Vincent Ng. 2024. Computational meme understanding: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21251–21267, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI. 2025. New embedding models and API updates. Accessed: 2025-10-21.

Graeme Ritchie. 2005. Computational mechanisms for pun generation. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*, Aberdeen, Scotland. Association for Computational Linguistics.

Graeme Ritchie. 2009. Variants of incongruity resolution. *Journal of Literary Theory (18625290)*, 3(2).

C. E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Shivam Sharma, Siddhant Agarwal, Tharun Suresh, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2023. What do you meme? generating explanations for visual semantic role labelling in memes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8):9763–9771.

Ayuto Tsutsumi and Yuu Jinnai. 2025. Do large language models know folktales? a case study of yokai in Japanese folktales. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16124–16146, Vienna, Austria. Association for Computational Linguistics.

Zhiwei Yu, Jiwei Tan, and Xiaojun Wan. 2018. A neural approach to pun generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1660, Melbourne, Australia. Association for Computational Linguistics.

Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. 2024. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13246–13257.

A Analysis Methodology

This section provides supplementary details for the statistical analysis used in Section 4, focusing on the definition and notation of Cohen's d.

A.1 Cohen's d

Cohen's d is the difference between the two group means divided by a pooled standard deviation:

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s_p} \tag{1}$$

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{2}$$

where $\bar{X}$, s, and n are the mean, standard deviation, and sample size for each group.

B Definition of Linguistic Features

This section defines the 26 linguistic features analyzed in Section 4, grouped into basic linguistic features, semantic/NLI-based features, LM-based features, and LLM-scored higher-order features.

B.1 Basic Linguistic Features

Response-independent measures are computed from the response alone and correspond to the table rows *length*, *unique chars*, *hiragana ratio*, *katakana ratio*, *noun ratio*, *verb ratio*, and *symbol ratio*; we used MeCab (Kudo et al., 2004) for tokenization and POS tagging. Here, *length* is the response character count, *unique chars* is the number of distinct characters in the response, and each character-type ratio is the proportion of characters of that type in the response. Prompt–response measures are computed by comparing each response with its prompt and correspond to *length ratio*, *lexical novelty*, *hiragana changes*, and *katakana changes*. Here, *length ratio* is the response-to-prompt character-count ratio, *lexical novelty* is the proportion of response words that do not appear in the prompt, and *hiragana/katakana changes* are the differences between the corresponding character-type ratios in the response and prompt.

B.2 Semantic Distance and Textual Entailment

Semantic distance is one minus cosine similarity between `text-embedding-3-large` (OpenAI, 2025) embeddings. For NLI we used `mDeBERTa-v3-base` (He et al., 2021) (fine-tuned on XNLI (Conneau et al., 2018) and multilingual-NLI (Laurer et al., 2022)); each (prompt, response) pair was fed with the prompt as premise and the response as hypothesis, with inputs truncated to 512 tokens, and we used softmax-normalized entailment/neutral/contradiction probabilities. In Table 2, these correspond to the rows *semantic distance*, *entailment*, *neutral*, and *contradiction*.

B.3 Surprisal and nPMI

Surprisal is the average negative log-probability per response token (nats/token) under `rinna/japanese-gpt2-medium`. We computed it with teacher forcing under two prefixes: "prompt: {prompt}\n response:" (prompt-conditioned) and "response:" (unconditioned). We use the natural logarithm. In Table 2, the two surprisal settings correspond to $\textit{surprisal}_{\text{prompt-response}}$ and $\textit{surprisal}_{\text{response-independent}}$, respectively. We also defined LM-based PMI and nPMI as follows:

$$\mathrm{PMI}(t, a) = \log p_\theta(a \mid t) - \log p_\theta(a) \tag{3}$$

$$\mathrm{nPMI}(t, a) = \frac{\mathrm{PMI}(t, a)}{-\log p_\theta(a \mid t)} \tag{4}$$

where $\log p_\theta(\cdot)$ is the sum of token log-probabilities of the response; this uses no co-occurrence windows or stopword removal.

B.4 LLM-Scored Higher-Order Features

Because of API cost, we scored 2,000 sampled prompt–response pairs (1,000 high-rated and 1,000 low-rated) with GPT-5 on a 1–5 scale across the following eight aspects (Table 2); higher scores indicate more of the stated property:

- *Ambiguity exploitation*: The use of lexical or structural ambiguity that enables multiple interpretations.
- *Associative distance*: A moderate and natural conceptual leap from the prompt to the response.
- *Benign violation*: Deviations framed as harmless and acceptable, grounded in benign violation theory (McGraw and Warren, 2010).
- *Coherence*: Strong discourse-level connectedness between the prompt and response.
- *Expectedness*: The ease of predicting the response given the prompt.

```
You are an expert judge of Oogiri humor.
Prompt: {prompt}
Choose the funniest response.
A: {response_A}
B: {response_B}
C: {response_C}
D: {response_D}
Important: Answer with A, B, C, or
D only. No explanation is required.
Answer:
```

Figure 2: Prompt for the relative judgment task.

```
You are an expert judge of Oogiri humor.
Prompt: {prompt}
Response: {response}
Is this response funny?
Important: Answer with either funny or not
funny only. No explanation is required.
Answer:
```

Figure 3: Prompt for the absolute judgment task.

- *Incongruity resolution*: The natural resolution of an initial mismatch by a coherent reinterpretation, grounded in incongruity-resolution theory (Ritchie, 2009).
- *Metaphor use*: The presence of metaphorical expression in the response.
- *Perspective shift*: A meaningful change in viewpoint or framing that enables a punchline.

Figure 5 provides the prompt template used to score higher-order linguistic features for prompt–response pairs.

C Benchmark Experiments

This section provides supplementary materials for the benchmark experiments in Section 5, including the prompt templates used for the relative- and absolute-judgment tasks and the insight-augmented setting.

C.1 Prompt Templates for LLMs

Figures 2 and 3 provide the baseline prompt examples for the relative- and absolute-judgment tasks, respectively. Figure 6 provides the insight-augmented prompt template used in our benchmark experiments.

C.2 Crowdworker Instructions

Figure 4 provides the task instructions shown to crowdworkers for the absolute- and relative-judgment tasks.

```
[Absolute judgment task (binary)]
Task: Given an Oogiri prompt and a response,
judge whether the response is funny.
Selection criterion: Is the response funny for
the given prompt?
Prompt: {prompt}
Response: {response}
—
[Relative judgment task (MCQA)]
Task: Given an Oogiri prompt and multiple
candidate responses, choose the funniest
response.
Selection criterion: Which response is the
funniest for the given prompt?
Prompt: {prompt}
A: {response_A}
B: {response_B}
C: {response_C}
D: {response_D}
(The number of options varies by task.)
```

Figure 4: Task instructions shown to crowdworkers.

C.3 Inter-Annotator Agreement

We report inter-annotator agreement among the 21 crowdworkers using Fleiss' κ. It was 0.035 for binary classification (200 items), and 0.067 ($\text{Binary}_{\text{same}}$), 0.235 ($\text{Binary}_{\text{diff}}$), 0.113 (three-choice), and 0.145 (four-choice) for the four MCQA tasks (100 items each). Agreement was highest for $\text{Binary}_{\text{diff}}$, where the negative response comes from a different prompt, which may make the choice less ambiguous. Overall agreement is low, which is expected because humor judgments are subjective and some items can be genuinely ambiguous. We therefore used attention checks and majority vote to obtain a robust aggregate label and interpret the human baseline as an approximate reference.

```
You receive a prompt and a response, and your task is to evaluate how appropriate the response is for the prompt. Please evaluate
the characteristics of the response to the prompt for the following “Oogiri” scenario on a scale of 1–5.
Prompt: {prompt}
Response: {response}
Evaluate based on the following criteria and respond in JSON format.
In doing so, please explain the reasoning for the scores.

1) ambiguity_exploitation (1–5): Use of Ambiguity
1: The response does not exploit ambiguity.
3: The response is somewhat ambiguous.
5: The response effectively exploits ambiguity.

2) associative_distance (1–5): Appropriateness of Association
1: The association between prompt and response is direct OR requires 5 or more associative leaps.
3: The association is reached in 1 step OR requires 4 steps with somewhat unnatural association.
5: The association is naturally reached in 2–3 steps.

3) benign_violation (1–5): Degree of Harmless Violation
1: The response deviates from the prompt and is extremely harmful/offensive.
3: The response deviates from the prompt and is somewhat harmful/offensive.
5: The response deviates from the prompt but is harmless.

4) coherence (1–5): Logical Coherence between prompt and Response
1: The prompt and response are not logically connected.
3: The prompt and response are somewhat logically connected.
5: The prompt and response are perfectly logically connected.

5) expectedness (1–5): Predictability of the Response
1: The response is completely unexpected and surprising relative to the prompt.
3: The response is somewhat unexpected or surprising relative to the prompt.
5: The response is very predictable or obvious relative to the prompt.

6) incongruity_resolution (1–5): Degree of Resolution of Incongruity
1: The incongruity between the prompt and response is not resolved at all.
3: The incongruity between the prompt and response is somewhat resolved.
5: The incongruity between the prompt and response is naturally resolved.

7) metaphor_use (1–5): Appropriateness of Metaphor Use
1: The response does not use metaphor regarding the prompt.
3: The response somewhat uses metaphor regarding the prompt.
5: The response uses metaphor regarding the prompt.

8) perspective_shift (1–5): Shift in Perspective
1: The response shows no shift in perspective regarding the prompt.
3: The response shows a partial shift in perspective regarding the prompt.
5: The response shows a clear shift in perspective regarding the prompt.

Output Requirements:
- All scores must be integers (1–5).
- In the reasoning field, summarize the concise basis for each score in 1–3 sentences.
- Return in JSON format.

{
“reasoning”: “Reason for the scores”,
“ambiguity_exploitation”: number,
“associative_distance”: number,
“benign_violation”: number,
“coherence”: number,
“expectedness”: number,
“incongruity_resolution”: number,
“metaphor_use”: number,
“perspective_shift”: number
}
```

Figure 5: Prompt for LLM-based scoring of higher-order linguistic features (shown in English; translated from the original prompt used in our experiments).

```
You are an expert judge of Oogiri humor. Please evaluate the funniness of the responses in two steps.
First, assign 1–5 scores for the following aspects and output JSON (for style_features, consult the measured values in
[Measurements]).
Then, use these features to decide which response is the funniest for the prompt.
If you are unsure, you may consult the core_features and style_features below; they are helpful cues but not absolute criteria.
[Scoring criteria]
core_features (higher scores indicate more of the stated property):
- perspective_shift: (criteria description)
- ambiguity_exploitation: (criteria description)
- incongruity_resolution: (criteria description)
- benign_violation: (criteria description)
- metaphor_mapping: (criteria description)
- bridge_complexity: (criteria description)
style_features (consult [Measurements] when needed):
- answer_length_chars: shorter tends to be funnier
- rel_length_ratio: lower tends to be funnier (answer shorter than prompt)
- unique_char_count: fewer tends to be funnier
- unique_char_ratio: lower tends to be funnier
- punctuation_density: lower tends to be funnier
- katakana_ratio: lower tends to be funnier
[Measurements]
Measured values for option A:
- answer_length_chars: {...}
- rel_length_ratio: {...}
- unique_char_count: {...}
- unique_char_ratio: {...}
- punctuation_density: {...}
- katakana_ratio: {...}
(Similarly for options B, C, and D.)
[Requirements]
- Output JSON only.
- All core_features scores must be integers (1–5).
- In the reasoning field, summarize the basis for the scores in 1–3 sentences.
- The decision field must be a single option label (e.g., “A”, “B”, “C”, or “D”).
- When assessing style_features, consult the measured values in [Measurements].
[Output format]
{
  “options”: {
    “A”: {
      “reasoning”: “Reason for the scores”,
      “core_features”: {
        “perspective_shift”: number,
        “ambiguity_exploitation”: number,
        “incongruity_resolution”: number,
        “benign_violation”: number,
        “metaphor_mapping”: number,
        “bridge_complexity”: number
      }
    },
    “B”: {...},
    ...
  },
  “decision”: “A” | “B” | “C” | “D”
}
Prompt: {prompt}
Choose the funniest answer among the four options.
A: {option_A}
B: {option_B}
C: {option_C}
D: {option_D}
Important: Output JSON only.
Answer:
```

Figure 6: Insight-augmented prompt template used in OOGIRI-MASTER. The model is instructed to output a JSON object with rubric-based scores and to consult measured linguistics features only when uncertain.