Assessing the Capabilities of LLMs in Humor: A Multi-dimensional Analysis of Oogiri Generation and Evaluation

Ritsu Sakabe [1] **Hwichan Kim** [1,2] **Tosho Hirasawa** [1,2] **Mamoru Komachi** [1]
[1]Hitotsubashi University
[2]Tokyo Metropolitan University
ritsu@scl.sds.hit-u.ac.jp, kim@scl.sds.hit-u.ac.jp, tosho@scl.sds.hit-u.ac.jp, mamoru.komachi@r.hit-u.ac.jp

Abstract

Computational humor is a frontier for creating advanced and engaging natural language processing (NLP) applications, such as sophisticated dialogue systems. While previous studies have benchmarked the humor capabilities of Large Language Models (LLMs), they have often relied on single-dimensional evaluations, such as judging whether something is simply "funny." This paper argues that a multifaceted understanding of humor is necessary and addresses this gap by systematically evaluating LLMs through the lens of Oogiri, a form of Japanese improvisational comedy games. To achieve this, we expanded upon existing Oogiri datasets with data from new sources and then augmented the collection with Oogiri responses generated by LLMs. We then manually annotated this expanded collection with 5-point absolute ratings across six dimensions: Novelty, Clarity, Relevance, Intelligence, Empathy, and Overall Funniness. Using this dataset, we assessed the capabilities of state-of-the-art LLMs on two core tasks: their ability to generate creative Oogiri responses and their ability to evaluate the funniness of responses using a six-dimensional evaluation. Our results show that while LLMs can generate responses at a level between low- and mid-tier human performance, they exhibit a notable lack of Empathy. This deficit in Empathy helps explain their failure to replicate human humor assessment. Correlation analyses of human and model evaluation data further reveal a fundamental divergence in evaluation criteria: LLMs prioritize Novelty, whereas humans prioritize Empathy. We release our annotated corpus to the community to pave the way for the development of more emotionally intelligent and sophisticated conversational agents.

Introduction

Humor is a crucial aspect of human cognition: it facilitates social bonding (Savage et al. 2017) and correlates with general intelligence (Greengross and Miller 2011). In an effort to realize this complex and human activity with AI, recent studies have begun to assess the humor comprehension, evaluation, and generation abilities of large language models (LLMs) (Hessel et al. 2023; Romanowski, Valois, and Fukui 2025; Chen et al. 2024). However, much of this existing research focuses on the "performance improvement" aspect—that is, how to make LLMs generate or evaluate more humorous responses. On the other hand, there has been insufficient

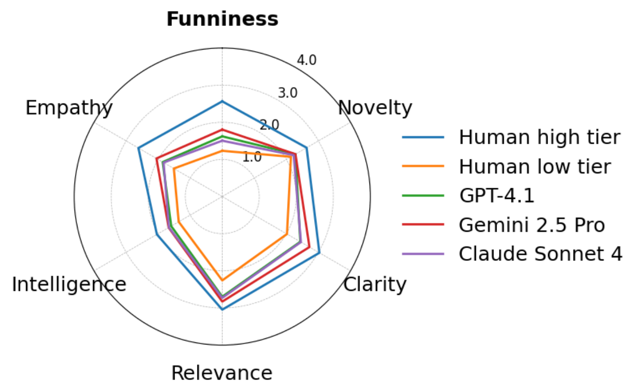

Figure 1: Multi-dimensional comparison of Oogiri responses from humans and LLMs, as evaluated by humans. "Human high tier" and "human low tier" refer to the highest- and lowest-rated human responses, as determined by manual annotation. (Funniness refers to Overall Funniness.)

analysis of how the structure of funniness as perceived by LLMs fundamentally differs from that of humans—in other words, the qualitative differences in their "humor sensibilities." Understanding this difference in sensibilities is essential for discerning whether LLMs interpret humor in the same way as humans, or if they possess their own unique preferences for what constitutes humor.

To bridge this sensibility gap and create more natural and engaging conversational agents, two fundamental abilities must be addressed. The first is humor generation (RQ1): the ability to produce contextually appropriate and creative humor. Developing this ability, however, relies on robust evaluation methods. While human assessment is the most trusted benchmark, the recent "LLM-as-a-Judge" paradigm offers a scalable alternative. This gives rise to the second, equally critical challenge: humor evaluation (RQ2). Before we can trust LLMs to guide the development of humorous AI, we must first understand if they can evaluate humor with human-like criteria and sensibilities.

In this study, we conduct detailed analyses of the humor capabilities of state-of-the-art LLMs such as GPT-4.1. We begin by collecting datasets of "Oogiri," a uniquely Japanese comedic format in which participants are given a topic (e.g.,

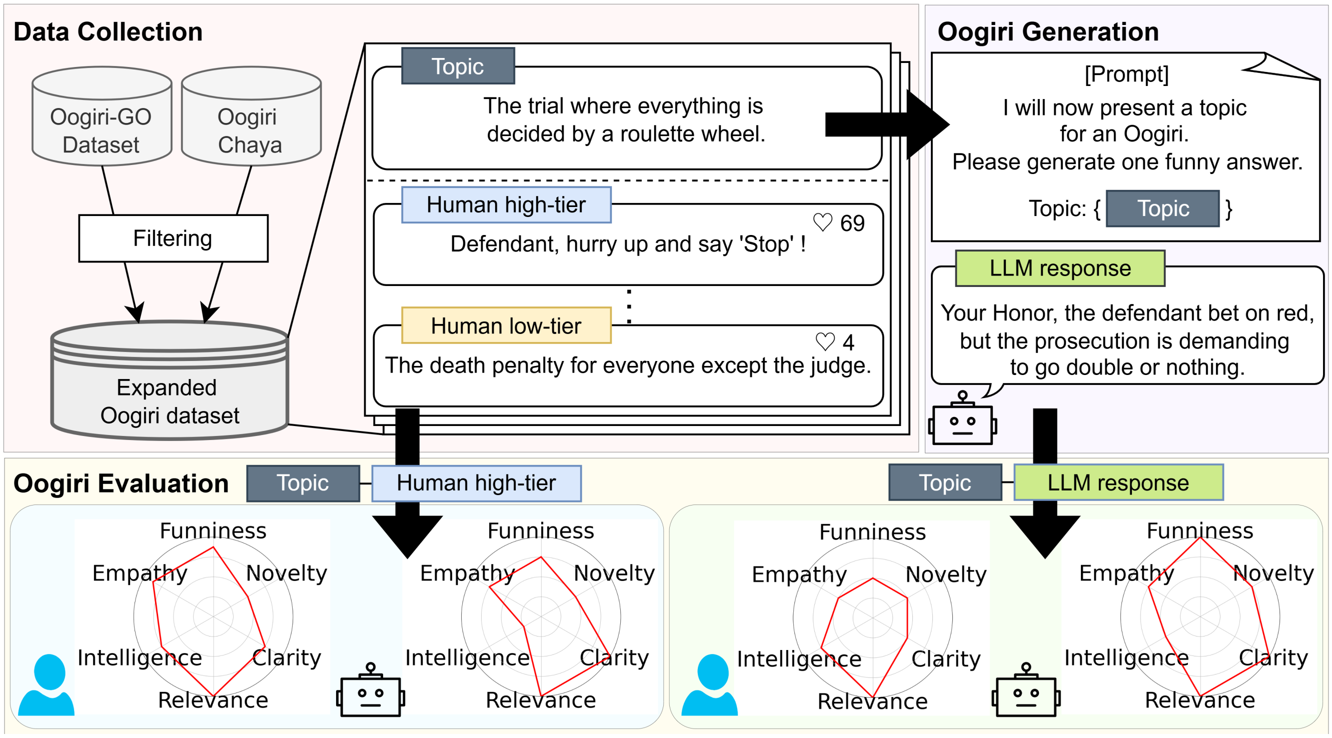

Figure 2: Overview of the experimental process for Oogiri generation and evaluation.

'A taxi driver who has awakened') and respond with a humorous punchline (e.g., 'I can make this traffic jam "disappear." Should I?'). Our primary resource is the Oogiri-GO dataset (Zhong et al. 2024), collected from Bokete, an online Oogiri competition platform. To broaden the range of topics and human-written responses, we additionally collect data from another platform, "Oogiri-Chaya." Using these resources, we address two core research questions regarding the humor capabilities of LLMs (Figure 2).

Our first research question asks: *Can LLMs produce humorous responses that humans judge as funny?* To address RQ1, we prompted several LLMs with the collected topics and then evaluated both the LLM-generated and human-generated responses through crowd annotations. Previous studies have mainly relied on binary judgments of whether a response is funny (Romanowski, Valois, and Fukui 2025; Zhong et al. 2024), but we argue that funniness is multifaceted, encompassing relevance, empathy, and other qualities (Hampes 2001). We therefore conducted a manual annotation along six dimensions: Novelty, Clarity, Relevance, Empathy, Intelligence, and Overall Funniness. Each dimension is rated on a 0–4 Likert scale (0 = not at all, 4 = excellent). This fine-grained evaluation allows us to determine not only whether LLMs can produce humor, but also which specific dimensions they excel at or struggle with.

Our second research question asks: *Can LLMs judge humorous responses that humans judge as funny?* To address RQ2, we provide the LLMs with each topic–response pair and instruct them to perform an evaluation using the same six-dimensional rubric described previously. We then compute the correlation between the human annotations and LLMs' scores. These correlation coefficients allow us to verify how reliably LLMs can assess Oogiri responses compared to human judges, and which evaluation dimensions LLMs prioritize when evaluating humor.

In summary, our contributions are three-fold:

Creation of a Multi-Dimensional Annotated Corpus for Oogiri Humor Research: To provide a reliable resource for humor research, we constructed a new, large-scale annotated corpus. The core of this contribution lies in the novel annotation of this data collection by both human annotators and the latest LLMs, applying 5-point ratings across our six proposed dimensions (Figure 1). We are releasing these annotations and our benchmark design to the community to provide a reliable resource for multifaceted humor analysis. [1]

Assessment of State-of-the-Art LLMs' Humor Generation Capabilities: We multi-dimensionally analyze the humor generation abilities of state-of-the-art LLMs (GPT-4.1, Gemini 2.5 Pro, Claude Sonnet 4) using manual, six-dimensional scoring of their generated responses. Our findings show that LLMs possess an Oogiri generation ability equivalent to that of low- to mid-tier humans, with Gemini 2.5 Pro being the most proficient among the models tested (Figure 1). An analysis across the evaluation dimensions reveals that the primary weakness of LLMs compared to humans is on the Empathy dimension.

[1] https://github.com/SDS-NLP/Oogiri-Eval

Assessment of State-of-the-Art LLMs' Humor Evaluation Capabilities: By comparing response evaluations across six dimensions from both humans and LLMs, we assess the degree of alignment between the two. This allows us to determine the extent to which LLMs can serve as proxies for human humor judges and to clarify which factors LLMs prioritize when evaluating humor. Our analysis reveals that while Claude Sonnet 4 showed the highest correlation with human evaluations among the models, overall agreement between humans and LLMs is limited. Furthermore, mirroring a tendency observed in the generation results, we found that while humans prioritize the Empathy dimension most when evaluating humor, LLMs prioritize the Novelty dimension. This finding suggests that LLMs may possess their unique "sense of humor," highlighting the complexity of aligning an AI's humor sensibilities with those of humans.

Related Work

Computational Humor

Humor has long been studied as a challenging aspect of AI and NLP. Early theories suggest that humor creation and appreciation require advanced cognitive abilities, such as the ability to recognize incongruities and shared context. For instance, a prominent cognitive framework, Suls (1972) two-stage model, posits that humor is appreciated when a perceiver first detects an incongruity in a joke and then successfully resolves it by finding a cognitive rule that reconciles the punchline with the setup. Empirical research by Greengross and Miller (2011) even showed that humor ability is correlated with human intelligence and mating success, implying that humor might be an inherent component of what we consider "intelligence." In the context of AI, humor understanding and generation test a model's semantic and pragmatic knowledge, creativity, and cultural awareness.

Recent NLP studies, using materials like jokes, puns, and satirical headlines, have tackled tasks such as humor detection (determining if a given text is humorous) and humor understanding/explanation (identifying why something is funny or choosing the correct punchline for a setup). Baranov, Kniazhevsky, and Braslavski (2023) examined the robustness of humor detection models and found that their performance can degrade when jokes are phrased differently or in new domains, suggesting that humor detection systems lack true general understanding. (Hessel et al. 2023) introduced a benchmark from the New Yorker Caption Contest to evaluate humor understanding by having models match cartoons with the funniest caption; even strong language models struggled with this task, often failing to predict which caption humans found funniest. In short, despite the impressive fluency of modern LLMs, truly "getting" complex human nuances, whether humor or "expressivity" (Tint et al. 2024), remains an open challenge, relating to the "Empathy" gap our study identifies.

Humor Generation

In the realm of humor generation, the challenge for AI is even more pronounced. Producing original humor has proven extremely difficult, as jokes often require creativity, world knowledge, and a fine sense of timing or surprise. For instance, Jentzsch and Kersting (2023) demonstrated that while models like ChatGPT can be amusing in conversation, they frequently miss the mark on humor. Their outputs tend to be formulaic or overly safe, lacking the spark that makes humans laugh. Several other studies have specifically addressed this difficulty.

One interesting approach by Horvitz et al. (2024) involved using LLMs to manipulate humor in text. They showed that an LLM can effectively remove the humor from a satirical news headline (making it "unfunny"), but doing the reverse (adding humor to a serious headline or creating a new joke) is much more challenging. This asymmetry indicates that LLMs, while good at understanding literal meaning, struggle with the creative leap required for humor. The limitations of LLM humor generation have been observed across multiple studies: models often rely on puns, clichés, or shallow wordplay and may fail to surprise the audience—an essential element of humor. These findings underscore that humor generation is a high-bar task pushing the boundaries of what LLMs can do, and improvements likely require advances in knowledge representation and perhaps incorporating cultural or common-sense reasoning.

Humor Evaluation

With the increasing capabilities of LLMs, a paradigm known as "LLM-as-a-Judge," which uses LLMs themselves as evaluators to substitute for human assessment, has garnered significant attention. Various attempts have also been made in humor research.

Romanowski, Valois, and Fukui (2025) have proposed a new metric that involves transcribing stand-up comedy, inputting it into an LLM, and evaluating whether the model can accurately extract the funny parts. Their evaluation of major LLMs shows that the accuracy remains around 50 %. It is also reported that various prompting techniques, such as assigning a persona, do not lead to notable improvements.

Furthermore, Hwang, West, and Shwartz (2025) have proposed an LLM-based method to understand and explain the funniness of multimodal humor (e.g., memes and cartoons) composed of images and captions. Their method uses the information bottleneck principle to extract relevant knowledge from a Vision-Language Model (VLM) and iteratively refine it. This research highlights the importance of world knowledge in humor comprehension.

Oogiri

The Oogiri format of humor has gained attention as a challenging test of creativity. Nakagawa et al. (2019) conducted one of the first analyses of Oogiri from a computational perspective. They used crowdsourcing to annotate Oogiri responses with various humor-related attributes and found that three factors (relevance to the topic, clarity, and novelty) were the most important contributors to a response's funniness. This suggests that a good Oogiri response must be on-topic (relevant), easy to understand, yet offer a fresh or unexpected twist.

More recently, the Oogiri-GO dataset by Zhong et al. (2024) provided a large-scale resource for multilingual humor research. The dataset compiled Oogiri topics and responses from the web and was used to benchmark the humor generation and evaluation abilities of LLMs. As mentioned earlier, their results showed that the models performed poorly in both generating humorous responses and comparatively evaluating humor. Our work builds on these insights by focusing specifically on Japanese-language humor and addressing some limitations of Oogiri-GO.

Dataset

This section describes the sources and curation process for our Oogiri dataset. The resulting collection of topics and responses is the foundation for the multi-dimensional manual annotation process, detailed in the Methodology section.

Sources of Oogiri Samples

To assemble a diverse collection of human-written Oogiri topics and responses, we drew on the two distinct sources.

Oogiri-GO is a dataset proposed by Zhong et al. (2024). This dataset is derived from Bokete[2], an online Oogiri competition platform. Each sample comprises three elements: a topic, a response submitted by a human competitor, and the number of votes that the response received from human judges based on its funniness. The topics are grouped into two categories: text-based and image-based. Text-based topics are expressed entirely in natural language, whereas image-based topics present an image for which participants must craft a humorous caption. [3]

Oogiri-Chaya is a dataset derived from another online Oogiri competition platform[4]. We collected all samples between 21 November 2021 and 17 July 2024 [5]. The dataset mirrors the structure of Oogiri-GO: each entry contains a topic, a participant's response, and the number of votes that response received, and is labelled as either text-based or image-based. After collecting samples, we detected topics and responses that potentially contain discriminatory, violent, sexual, or otherwise socially inappropriate content using GPT-4o. Then, we manually reviewed all flagged items and removed those deemed inappropriate.

Characteristics of the two Oogiri sources The two data sources differ significantly in terms of their evaluation systems and the potential biases that may arise from these systems. In Bokete, the data source for Oogiri-GO, the design allows users to post and vote while viewing other users' responses and their current vote counts. This system can potentially give rise to a "first-mover advantage," where earlier responses are favored. It can also cause a "conformity bias" (bandwagon effect), in which popular responses tend to attract more votes. Consequently, there is a concern that the vote counts may be influenced not only by the pure funniness of a response but also by the timing of submission and other users' evaluations.

On the other hand, Oogiri-Chaya is designed to systematically eliminate these biases. During the submission period, no other participants' responses are visible, and during the evaluation period, the vote counts for each response remain hidden. Therefore, participants can evaluate responses based on pure funniness without being influenced by others, making the evaluation data from Oogiri-Chaya a more objective and reliable indicator of funniness compared to Bokete. In particular, responses with low vote counts can be considered to have been judged as "unfunny" with a high degree of certainty, free from the influence of such biases. By incorporating samples from Oogiri-Chaya, we not only broaden the range of topics and responses but also enhance the overall reliability of our dataset with less biased evaluation data.

Ensuring the Quality of Samples

To guarantee the quality of the dataset, we applied filtering procedures to both the Oogiri-GO and Oogiri-Chaya data sources, selecting only samples that met specific participation and voting criteria.

For the Oogiri-GO dataset, we retained only those topics that satisfied two conditions: (1) the top-voted response had at least 100 votes, and (2) the topic had at least 10 submitted responses. This filtering process resulted in a final set of 1,329 topics (313 text-based and 1,016 image-based).

Similarly, for the Oogiri-Chaya dataset, we selected only topics that had 100 or more responses, yielding a total of 551 topics (425 text-based and 126 image-based).

Methodology

This section details the overall experimental design used to address our research questions. For RQ1, we investigate whether state-of-the-art LLMs can produce Oogiri responses that humans perceive as funny, and how their humor compares with that of humans. Because Oogiri humor is inherently multidimensional, assessing it along a single funniness dimension is insufficient. To capture its nuances, we evaluate each response across several complementary dimensions. We let the LLMs generate responses for each Oogiri topic and then ask human annotators to rate each response along these dimensions. The resulting scores reveal the strengths and limitations of current LLMs in humorous response generation. For RQ2, we investigate whether state-of-the-art LLMs can evaluate Oogiri responses that humans perceive as funny, and to what extent the characteristics of their humor evaluation align with those of humans.

Experimental Materials and Stimuli

To analyze the characteristics of Oogiri generation, we instructed the latest models—GPT-4.1 (gpt-4.1-2025-04-14), Gemini 2.5 Pro (gemini-2.5-pro-preview-05-06), and Claude Sonnet 4 (claude-sonnet-4-20250514-v1:0)—to generate Oogiri responses. For each Oogiri topic, we embedded it within a base prompt (see Appendix A for details) written in Japanese to solicit a humorous response. All generation was performed with default parameter settings.

From the collected datasets (Oogiri-GO and Oogiri-Chaya), we selected 200 topics (100 text-based and 100

[2]https://bokete.jp/

[3]Image-based topics consisting solely of text were transcribed using GPT-4.1 and reclassified as text-based.

[4]https://oogiri-chaya.com/

[5]We scraped this data on July 16-19, 2024.

	Novelty	Clarity	Relevance	Intelligence	Empathy	Overall Funniness
High	**2.630** (0.578)	**3.029** (0.691)	**3.046** (0.665)	**2.039** (0.708)	**2.615** (0.656)	**2.563** (0.648)
Mid	2.320 (0.621)	2.571 (0.822)	2.652 (0.724)	1.732 (0.669)	2.100 (0.715)	1.913 (0.675)
Low	2.142 (0.690)	2.018 (0.928)	2.258 (0.876)	1.358 (0.614)	1.510 (0.738)	1.232 (0.658)
Unrelated	2.377 (0.630)	0.898 (0.702)	0.935 (0.727)	1.032 (0.613)	0.725 (0.615)	0.681 (0.591)
GPT-4.1 (serious)	1.369 (0.826)	2.836 (0.728)	<u>3.055</u> (0.623)	1.786 (0.741)	2.023 (0.710)	0.976 (0.654)
GPT-4.1	2.282 (0.608)	2.435 (0.840)	2.696 (0.691)	1.592 (0.656)	1.858 (0.770)	1.621 (0.689)
Gemini 2.5 Pro	2.283 (0.616)	2.719 (0.734)	2.828 (0.674)	1.671 (0.638)	2.053 (0.700)	1.803 (0.700)
Claude Sonnet 4	2.227 (0.597)	2.452 (0.815)	2.729 (0.711)	1.662 (0.636)	1.830 (0.707)	1.504 (0.669)

Table 1: Mean (and standard deviation) of human evaluations of Oogiri generation.

image-based) based on the popularity of their top-voted answers. For each topic, we prepared a final evaluation set consisting of the following eight types of responses:

- **High/Mid/Low**: The response with the most/average/fewest votes on the source website
- **Unrelated**: The top-voted response from a different, irrelevant topic
- **GPT-4.1 (serious)**: A serious (non-humorous) response by GPT-4.1
- **GPT-4.1**: A (humorous) response by GPT-4.1
- **Gemini 2.5 Pro**: A response by Gemini 2.5 Pro
- **Claude Sonnet 4**: A response by Claude Sonnet 4

Evaluation Procedure

Multi-dimensional Evaluation Framework We conducted absolute evaluations for the eight response types across six dimensions: Novelty, Clarity, Relevance, Intelligence, Empathy, and Overall Funniness. Each dimension was rated on a 0–4 integer scale. The specific definitions provided to annotators are as follows:

1. **Novelty**: Measures the originality of the response to the topic. A score of 0 means "not new at all (something anyone could think of)," while a 4 means "very new (a unique perspective that others would not think of)."
2. **Clarity**: Measures how easy the response is to understand. A 0 indicates "completely incomprehensible (the meaning is unclear even after multiple readings)," and a 4 indicates "easily comprehensible (the meaning is immediately clear)."
3. **Relevance**: Measures the connection between the topic and the response. A 0 means "completely unrelated to the topic," while a 4 means "very closely related to the topic."
4. **Intelligence**: Measures the intellectual quality of the response. A 0 means "contains no intellectual elements (silly or trivial)," and a 4 means "highly intelligent (conveys intellect or sophistication)."
5. **Empathy**: Measures how relatable the response is. A 0 means "not relatable at all (unconvincing, hard to imagine the situation or feelings)," while a 4 means "highly relatable (the situation and feelings are well understood)."

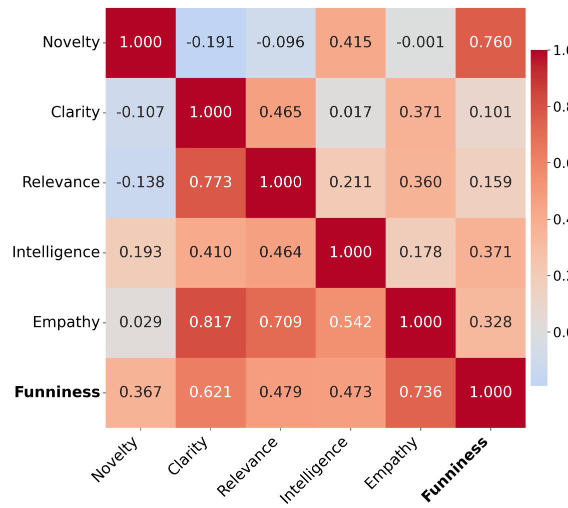

Figure 3: Spearman correlation matrix of human and GPT-4.1 evaluation. Human evaluations are in the lower triangle, and GPT-4.1 evaluations are in the upper triangle. GPT-4.1 is a representative example; see Appendix E for other models.

6. **Overall Funniness**: Measures the pure humor of the response, integrating all dimensions and factors. A 0 means "not funny at all," and a 4 means "extremely funny."

Human and LLM Annotation Process Both human and LLM annotators performed the evaluation. For human annotation, we used Lancers[6]. Each response was rated by four different native Japanese speakers, recruited without regard to age or educational background. To ensure a fair rating, we created eight distinct subsets of the data, ensuring each annotator evaluated only one response type per topic. For LLM annotation, we used the same models from the generation task (GPT-4.1, Gemini 2.5 Pro, Claude Sonnet 4) and provided them with the same instructions and evaluation framework as the human annotators. The prompt template for the LLM evaluation task is shown in Appendix A.

Analysis Procedure

For RQ1, we analyze the human evaluation scores for each response type. For RQ2, we compute the Spearman's cor-

[6]https://www.lancers.jp

Responses	Emp.	Fun.
A funeral director shouldn't say things like, 'See you next time!' (High)	3.00	2.00
So, this makes it zero calories now, right? (GPT-4.1)	1.25	0.75

Table 2: Examples for the topic "Tear a coupon to pieces and say one line." illustrating discrepancies in human ratings of empathy and funniness between human and LLM responses.

relation coefficient between the human scores and the LLM scores. A high correlation would indicate that an LLM can mimic human evaluation abilities, while a low correlation would suggest a divergence in their evaluation criteria.

RQ1: Can LLMs Generate Humorous Responses That Humans Judge as Funny?

Results Human evaluation data (Table 1) reveals several key findings regarding LLM generation abilities. Firstly, current models generate Oogiri responses at a level between that of low- and mid-tier humans. For instance, the best-performing model, Gemini 2.5 Pro, averaged 1.803 in Overall Funniness, placing it below the human Mid-tier (1.913) but above the Low-tier (1.232). Secondly, the primary weakness of LLMs compared to high-tier human responses is a notable deficit in Empathy. As shown in the radar chart (Figure 1), the largest performance gap is on this dimension. Finally, among the LLMs, Gemini 2.5 Pro was the most proficient, achieving the highest scores across all six dimensions, largely due to its relative strength in Empathy and Clarity.

Discussions The results suggest that while LLMs can construct logically coherent and relevant responses, they struggle to create truly compelling humor due to a gap between pattern recognition and genuine human-like cognition. The notable deficit in Empathy points to the core of this challenge. As revealed by the heatmap of human evaluation dimensions (lower triangle of Figure 3), Empathy is the dimension most strongly correlated with Overall Funniness for humans. This indicates that the ability to resonate with shared human social contexts, cultural norms, and emotions is a foundational element of high-quality humor, and it is precisely this area where current LLMs fall short (Table 2). This finding can be effectively interpreted through the lens of incongruity-resolution theories of humor (Suls 1972). Our results suggest that LLMs are relatively capable of the first stage of this process: creating incongruity by generating responses with high Novelty while maintaining some Relevance to the topic. However, they fail at the crucial second stage—the resolution. This is because resolving the incongruity satisfyingly requires Empathy, the ability to ground the unexpected punchline in a shared, relatable human context. The LLMs' notable deficit in Empathy prevents them from achieving this humorous resolution, even when their responses are novel and relevant. Therefore, to generate responses that humans consistently find funny, future development must focus not only on creativity (Novelty) but critically on improving the models' capacity for empathetic reasoning.

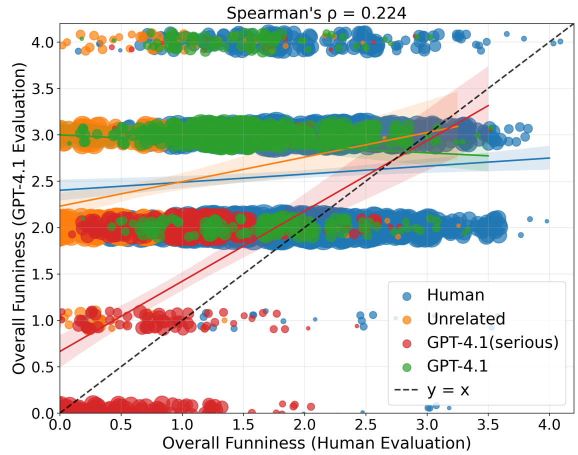

Figure 4: Human vs. LLM ratings: Scatter plot of Overall Funniness. Solid colored lines show the linear regression trend for each response category, with shaded areas indicating the 95 % confidence interval. The dashed line ($y = x$) represents perfect agreement. GPT-4.1 is a representative example; see Appendix D for results from all models.

RQ2: Can LLMs Judge Humorous Responses That Humans Judge as Funny?

Results An analysis of the correlation between human and LLM evaluations (Table 3) indicates a limited degree of agreement. The overall Spearman correlation for Overall Funniness is low for all models, with Claude Sonnet 4 showing a marginally higher correlation ($\rho = 0.266$) than GPT-4.1 ($\rho = 0.224$) and Gemini 2.5 Pro ($\rho = 0.169$). The scatter plots (Figure 4) visually confirm this weak relationship (Figure 4 shows GPT-4.1; see Appendix D for all models).

A detailed look at the results of the LLM evaluations for each response type:

- For human-generated responses (combining High, Mid, and Low tiers), the LLM evaluations showed a weak positive correlation with human ratings (see Appendix B for details). The accuracy rates for correctly predicting the Overall Funniness order of High, Mid, and Low were 50.9 % for GPT-4.1, 51.9 % for Gemini 2.5 Pro, and 54.1 % for Claude Sonnet 4, respectively, indicating some ability to discern the hierarchy of funniness.
- A disagreement between human and LLM evaluators was observed for irrelevant (Unrelated) responses. Humans consistently rated these as not funny, giving them a very low mean Overall Funniness score of 0.681. In contrast, LLMs rated these same responses very highly, with mean scores of 2.411 from GPT-4.1 (Table 4), 3.291 from Gemini 2.5 Pro, and 2.183 from Claude Sonnet 4.
- Similarly, for the LLMs' own generated responses, which humans judged to be of lower quality than the Mid-tier, the LLMs rated them at a level comparable to or even higher than the High-tier human responses. Furthermore,

Model	**Novelty**	**Clarity**	**Relevance**	**Intelligence**	**Empathy**	**Overall Funniness**
GPT-4.1	0.361	0.379	**0.441**	**0.352**	**0.303**	0.224
Gemini 2.5 Pro	**0.393**	0.356	0.328	0.188	0.248	0.169
Claude Sonnet 4	0.375	**0.383**	0.315	0.275	0.278	**0.266**

Table 3: Spearman correlation coefficients of Oogiri evaluation between humans and LLMs.

	Novelty	**Clarity**	**Relevance**	**Intelligence**	**Empathy**	**Overall Funniness**
High	2.416 (0.884)	3.877 (0.445)	3.719 (0.674)	1.466 (0.884)	2.602 (0.841)	2.637 (0.787)
Mid	2.376 (0.924)	3.875 (0.448)	3.652 (0.717)	1.416 (0.852)	2.617 (0.842)	2.604 (0.785)
Low	2.429 (0.953)	3.702 (0.645)	3.514 (0.832)	1.376 (0.888)	2.286 (0.856)	2.461 (0.813)
Unrelated	**2.722** (0.964)	3.158 (1.072)	2.624 (1.065)	1.333 (0.872)	2.075 (0.987)	2.411 (0.863)
GPT-4.1 (serious)	1.195 (1.113)	<u>3.997</u> (0.050)	<u>3.932</u> (0.359)	1.659 (0.940)	2.749 (0.861)	1.404 (1.125)
GPT-4.1	2.524 (0.712)	3.987 (0.132)	**3.917** (0.348)	1.627 (0.675)	**2.955** (0.739)	2.895 (0.629)
Gemini 2.5 Pro	2.622 (0.753)	**3.992** (0.086)	3.885 (0.445)	1.619 (0.754)	2.945 (0.761)	**2.947** (0.676)
Claude Sonnet 4	2.476 (0.773)	3.987 (0.150)	3.897 (0.410)	**1.689** (0.664)	2.920 (0.756)	2.794 (0.693)

Table 4: Mean (and standard deviation) of GPT-4.1 evaluations of Oogiri generation.

the LLM evaluations for these responses had a lower standard deviation compared to their evaluations of other response types, indicating less variance in their ratings.

- For intentionally non-humorous (serious) responses, the evaluations from humans and LLMs showed a surprisingly high correlation (Figure 4).

Discussions The low overall correlation stems from a fundamental divergence in how humans and LLMs evaluate humor. The results suggest LLMs have a unique, non-human-like sense of humor. This alien "sensibility" manifests in clear evaluative biases that differ from human judgment.

One such bias is a positivity bias. The models systematically overvalued irrelevant (Unrelated) responses. While humans penalize Unrelated responses for their lack of clarity, relevance, and empathy, LLMs rate them highly. This bias is also evident elsewhere; while humans differentiate between High, Mid, and Low-tier human responses with an average score gap of over 0.6 points, LLMs evaluate them with only a slight difference in scores (Tables 1, 4). In other words, even for inadequate responses to a topic (such as Unrelated or human Low answers), the models judge them as potentially containing clarity, relevance, empathy, and funniness, and assign high scores. This evaluative leniency could stem from the models' core training objectives. Through processes like instruction and preference tuning, they are optimized to be cooperative and to generate positive responses, inadvertently predisposing them to assign favorable scores and hesitate to provide negative feedback.

Another observable bias is a self-preference bias, where LLMs tend to rate responses generated by other LLMs as comparable to high-tier human answers, whereas humans place them below the mid-tier (Tables 1, 4). This self-preference bias aligns with studies by Panickssery, Bowman, and Feng (2024) and Wataoka, Takahashi, and Ri (2025).

These biases appear to be symptoms of a deeper difference in evaluation logic. This is supported by the models' internal evaluation logic, where, in contrast to humans for whom Empathy is the strongest correlate of Funniness (lower triangle in Figure 3), Novelty is the strongest correlate for LLMs (the upper triangle in Figure 3 shows GPT-4.1 as a representative; see Appendix E for other models). This tendency to prioritize unexpectedness over contextual and empathetic resonance helps explain why LLMs are prone to the specific evaluative biases mentioned above.

In contrast to this widespread failure in evaluating positive humor, a notable pattern is that agreement is significantly higher for non-humorous content (Figure 4). While LLMs fail to grasp the subjective essence of high-quality humor, they show a strong ability to agree with humans on what is not funny (the serious responses). This suggests that "agreement on the negative" functions as an easier, objective classification task that does not require the shared subjective sensibility needed for appreciating positive humor.

Conclusion

In this study, we conducted a multi-dimensional analysis of the humor capabilities of LLMs to generate and evaluate Oogiri responses. This analysis is based on a comprehensive corpus, which we constructed by filtering prior datasets and integrating new data, containing responses from both humans and LLMs. We then manually annotated this corpus from a six-dimensional perspective. Regarding generation ability, our results show that LLM performance remains at a level between that of low- and mid-tier humans. This performance limitation stems primarily from a deficit in Empathy. To enhance humor generation ability, future training should focus on improving Empathy. Regarding evaluation ability, human and LLM assessments showed low agreement, revealing a fundamental divergence in evaluation criteria: humans prioritize Empathy, whereas LLMs prioritize Novelty. This divergence in criteria suggests that humans and LLMs may possess different sensibilities for evaluating humor.

Acknowledgments

This work is partly supported by JST PRESTO Grant Number JPMJPR2366, and also by JSPS KAKENHI Grant Numbers 22H03651 and 25K03178.

References

Baranov, A.; Kniazhevsky, V.; and Braslavski, P. 2023. You Told Me That Joke Twice: A Systematic Investigation of Transferability and Robustness of Humor Detection Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 13701–13715. Singapore: Association for Computational Linguistics.

Chen, Y.; Yang, C.; Hu, T.; Chen, X.; Lan, M.; Cai, L.; Zhuang, X.; Lin, X.; Lu, X.; and Zhou, A. 2024. Are U a Joke Master? Pun Generation via Multi-Stage Curriculum Learning towards a Humor LLM. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 878–890. Bangkok, Thailand: Association for Computational Linguistics.

Greengross, G.; and Miller, G. 2011. Humor ability reveals intelligence, predicts mating success, and is higher in males. *Intelligence*, 39(4): 188–192.

Hampes, W. P. 2001. Relation between humor and empathic concern. *Psychological Reports*, 88(1): 241–244.

Hessel, J.; Marasovic, A.; Hwang, J. D.; Lee, L.; Da, J.; Zellers, R.; Mankoff, R.; and Choi, Y. 2023. Do Androids Laugh at Electric Sheep? Humor "Understanding" Benchmarks from The New Yorker Caption Contest. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 688–714. Toronto, Canada: Association for Computational Linguistics.

Horvitz, Z.; Chen, J.; Aditya, R.; Srivastava, H.; West, R.; Yu, Z.; and McKeown, K. 2024. Getting Serious about Humor: Crafting Humor Datasets with Unfunny Large Language Models. arXiv:2403.00794.

Hwang, E.; West, P.; and Shwartz, V. 2025. BottleHumor: Self-Informed Humor Explanation using the Information Bottleneck Principle. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 22611–22632. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.

Jentzsch, S.; and Kersting, K. 2023. ChatGPT is fun, but it is not funny! Humor is still challenging Large Language Models. In Barnes, J.; De Clercq, O.; and Klinger, R., eds., *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, 325–340. Toronto, Canada: Association for Computational Linguistics.

Nakagawa, H.; Murawaki, Y.; Kawahara, D.; and Kurohashi, S. 2019. Analysis of Components of Humor in Oogiri using Crowdsourcing (In Japanese). In *Proceedings of the 25th Annual Meeting of the Association for Natural Language Processing*, 233–236. The Association for Natural Language Processing.

Panickssery, A.; Bowman, S. R.; and Feng, S. 2024. LLM Evaluators Recognize and Favor Their Own Generations. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 68772–68802. Curran Associates, Inc.

Romanowski, A.; Valois, P. H. V.; and Fukui, K. 2025. From Punchlines to Predictions: A Metric to Assess LLM Performance in Identifying Humor in Stand-Up Comedy. In Kuribayashi, T.; Rambelli, G.; Takmaz, E.; Wicke, P.; Li, J.; and Oh, B.-D., eds., *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 36–46. Albuquerque, New Mexico, USA: Association for Computational Linguistics. ISBN 979-8-89176-227-5.

Savage, B. M.; Lujan, H. L.; Thipparthi, R. R.; and DiCarlo, S. E. 2017. Humor, laughter, learning, and health! A brief review. *Advances in Physiology Education*, 41(3): 341–347. PMID: 28679569.

Suls, J. M. 1972. A Two-Stage Model for the Appreciation of Jokes and Cartoons: An Information-Processing Analysis. In Goldstein, J. H.; and McGhee, P. E., eds., *The Psychology of Humor: Theoretical Perspectives and Empirical Issues*, 81–100. New York: Academic Press.

Tint, J.; Sagar, S.; Taparia, A.; Raines, K.; Pathiraja, B.; Liu, C.; and Senanayake, R. 2024. ExpressivityArena: Can LLMs Express Information Implicitly? arXiv:2411.08010.

Wataoka, K.; Takahashi, T.; and Ri, R. 2025. Self-Preference Bias in LLM-as-a-Judge. arXiv:2410.21819.

Zhong, S.; Huang, Z.; Gao, S.; Wen, W.; Lin, L.; Zitnik, M.; and Zhou, P. 2024. Let's Think Outside the Box: Exploring Leap-of-Thought in Large Language Models with Creative Humor Generation . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13246–13257. Los Alamitos, CA, USA: IEEE Computer Society.