Memes-as-Replies: Can Models Select Humorous Manga Panel Responses?

Ryosuke Kohita
CyberAgent
kohita_ryosuke@cyberagent.co.jp

Seiichiro Yoshioka
CyberAgent
yoshioka_seiichiro@cyberagent.co.jp

Abstract

Memes are widely used not only as static artifacts but also as interactive replies in online conversations. While computational research has focused on analyzing intrinsic properties of memes, the dynamic and contextual use of memes to create humor remains largely unaddressed. We introduce the **Meme Reply Selection** task and present MAME-RE,[1] a benchmark of 100,000 human-annotated context–meme pairs built from openly licensed Japanese manga panels. Our analysis reveals three key findings: (1) although large language models and embedding methods exhibit different selection strategies, neither reliably outperforms a simple meme-prior baseline in the full-candidate setting; (2) visual information does not consistently improve performance, revealing a gap between understanding visual content and effectively using it for contextual humor; (3) while models approach the human selection baseline in a simplified setting with clearly distinguishable candidates, they struggle to distinguish subtle differences in wit among semantically similar candidates. These findings establish contextually humorous reply selection as a challenge that current models have yet to solve.

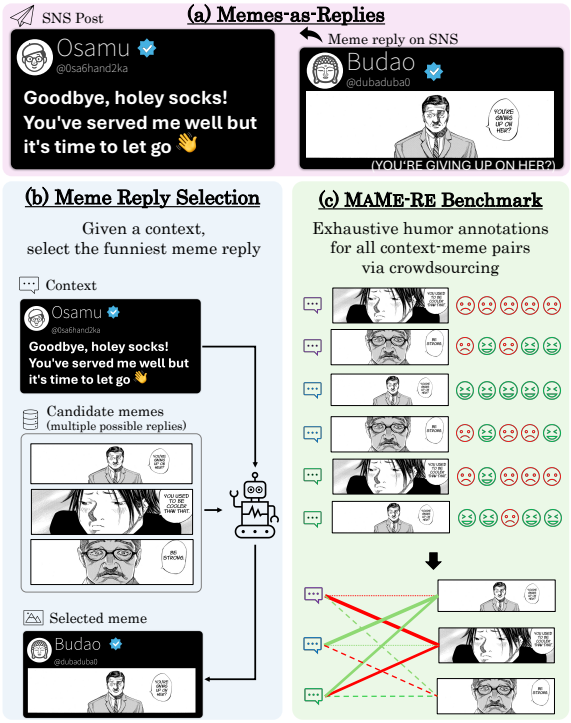


Figure 1: Overview of memes-as-replies. (a) Example of meme use on SNS. (b) Visualization of the Meme Reply Selection task. (c) MAME-RE benchmark with crowdsourced humor labels.

1 Introduction

Memes are popular in modern web communication (Joshi et al., 2024), used not just as static content but as interactive conversational replies (Wang et al., 2019), enabling creative expressions difficult to convey through text alone (Grundlingh, 2018). For example, in Figure 1(a), a user posts about their worn-out socks. Another user replies with a manga panel: "*YOU'RE GIVING UP ON HER?*" The humor arises from treating the socks as a person. Such replies span playful teasing to irony, relying on a deep understanding of context and implicit cultural knowledge (Sharma et al., 2022). While humans can intuitively grasp the social cues required for such humor, current computational approaches remain limited (Nguyen and Ng, 2024). This gap between widespread human practice and limited computational understanding motivates our work.

Most computational research treats memes as static content for analyzing intrinsic properties such as harmfulness detection or humorous captioning (Sharma et al., 2020; Loakman et al., 2025). We instead focus on the interactive function: in Figure 1(b), we define the *Meme Reply Selection* task, where the goal is to choose the funniest meme given a conversational context. Humor here is not intrinsic to the meme but emerges from its interac-

[1] https://github.com/CyberAgentAILab/mame-re. Pronounced as /meɪm riː/ (Maym-Re).

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 44207–44222
October 24-29, 2026 ©2026 Association for Computational Linguistics

tion with context. This distinguishes our task from dialogue media selection for emojis or stickers, which typically prioritize semantic relevance (Lu et al., 2023). We argue that modeling this dynamic and interactive use is essential to fully understand the role of memes in online communication.

We lay the foundation by formalizing the Meme Reply Selection task, making humorous reply selection computationally tractable. To support this task, we introduce the MAME-RE benchmark illustrated in Figure 1(c). Quality and scale are ensured through a rigorous crowdsourcing process, yielding annotations from five workers for each of the 100,000 pairs. Its exhaustive pairing design, where every context is matched with every meme, allows for a unique analysis of how a single image's humorous potential changes with context. Its openly licensed foundation ensures free redistribution and modification, fostering future research.

Our experiments reveal that current models fall short on this task. Although LLMs and embedding methods exhibit different selection strategies, no model reliably outperforms a simple meme-prior baseline that always picks the single most popular meme. Visual information does not consistently help, with performance sometimes degrading when it was provided. In a simplified setting with clearly distinguishable candidates, models approach the human selection baseline, but they struggle with the more realistic task of selecting the best reply from semantically similar candidates. These results indicate that selecting contextually humorous replies remains beyond the reach of current models.

The main contributions of this paper are as follows:

- We propose treating memes as a form of interactive web communication, broadening the scope from the analysis of static artifacts to the dynamics of online conversation.
- We provide the foundational infrastructure for this paradigm: the formal Meme Reply Selection task and MAME-RE, a benchmark featuring 100,000 human-annotated context–meme pairs, built entirely with openly licensed materials to ensure broad reusability.
- We present the first comprehensive analysis of this benchmark, identifying fundamental challenges for current models in understanding nuanced, multimodal humor.

2 Related Work

Web Communication and Multimodal Signals. Early studies explored how users adapt text-based media to convey socioemotional information despite the lack of non-verbal cues (Walther, 1992, 1996; Dresner and Herring, 2010). With social media, non-textual elements became integral to digital conversations (Highfield and Leaver, 2016; Hu et al., 2017; Jiang et al., 2018). Memes represent a more advanced form of this practice (Fei et al., 2021), used for purposes ranging from humor (Wu et al., 2025; Zhong et al., 2024) to political commentary (Kulkarni, 2017) and hate speech detection (Kiela et al., 2020). Of particular interest to our work is their use as humorous conversational replies, which remains computationally understudied.

From Static Content to Interactive Replies. Computational research has predominantly treated memes as static artifacts for content analysis (Pramanick et al., 2021; Hossain et al., 2022; Hwang and Shwartz, 2023). This paradigm has advanced harmful content detection (Zhuang et al., 2025), meme context understanding (Sharma et al., 2023b), and humorous captioning (Zhang et al., 2024). However, this focus overlooks how memes function as interactive replies (Davison, 2012; Yus, 2018). The meaning and humorous effect are not contained within the meme itself, but are generated by the interaction between the meme and its conversational context.

Humor Mechanisms in Memes and Dialogue. Memes range from reaction GIFs (Wang and Jurgens, 2021) to photographic templates with user-generated captions (Davison, 2012). A central theory in humor studies is *incongruity resolution* (Suls, 1972; Attardo and Raskin, 1991; Mihalcea et al., 2010): humor arises when an unexpected element is resolved through reinterpretation. In interactive replies, this maps to *recontextualization* (Bauman and Briggs, 1990) — placing a fixed-meaning medium in a new context to create surprising juxtaposition. LLM evaluations show that models can identify broad humor categories but fine-grained funniness prediction remains challenging (Hessel et al., 2023; Jentzsch and Kersting, 2023). Dialogue systems for selecting visual replies like stickers have focused on semantic relevance (Wang and Jurgens, 2021) rather than humor.

3 Problem Formulation

Task Definition. Let the context be a natural-language utterance $c \in \mathcal{C}$, where $\mathcal{C}$ is the set of arbitrary contexts in natural language. Given a set of memes $\mathcal{M} = \{m_1, \ldots, m_{|\mathcal{M}|}\}$, the task is to select the funniest meme for a given context: $\hat{m}(c) = \arg\max_{m \in \mathcal{M}} s(c, m)$, where $s(c, m)$ is a scoring function that measures the funniness of meme m given the context c.

Dataset Requirements. We require a dataset $\mathcal{D}$ consisting of contexts $\mathcal{C}_\mathcal{D} = \{c_1, \ldots, c_{|\mathcal{C}_\mathcal{D}|}\} \subset \mathcal{C}$, a finite set of memes $\mathcal{M}_\mathcal{D} = \{m_1, \ldots, m_{|\mathcal{M}_\mathcal{D}|}\} \subset \mathcal{M}$, and a real-valued scoring function $r : \mathcal{C}_\mathcal{D} \times \mathcal{M}_\mathcal{D} \to \mathbb{R}$, which assigns a funniness score to each context–meme pair. Thus, the dataset is represented as a set of annotated triples $\mathcal{D} = \{(c, m, r(c, m)) \mid c \in \mathcal{C}_\mathcal{D}, m \in \mathcal{M}_\mathcal{D}\}$.

Evaluation Metric. We define Score@1 as the average reference funniness score of selected memes: $\text{Score@1} = \frac{1}{|\mathcal{C}_\mathcal{D}|} \sum_{c \in \mathcal{C}_\mathcal{D}} r(c, \hat{m}(c))$, where $r(c, m)$ denotes the reference funniness score from the dataset.

4 MAME-RE: Manga Meme Reply Benchmark

To ensure data consistency and reusability, we focused on a single meme domain: Japanese manga panels from *Black Jack ni Yoroshiku*.[2] The benchmark comprises 250 synthetic social media contexts and 400 manga panels, forming 100,000 unique context-meme pairs. Each pair was evaluated by five independent crowdworkers on a binary funniness scale, yielding 500,000 annotations.

Content Collection and Curation. We extracted 400 panels covering diverse characters and scenes from *Black Jack ni Yoroshiku* by Shuho Sato, provided under a free-use license (Appendix A). For each panel, in-panel speech text was transcribed and a visual description was generated by GPT-4.1. We also prepared 250 synthetic social media contexts via GPT-4.1 to mimic posts on X (formerly Twitter); synthetic contexts ensure privacy and open licensing. We chose manga because its humor arises from recontextualizing original scenes without alteration, enabling focused study of emergent meme–context humor.

Funniness Annotation. 2,325 annotators on Yahoo! Crowdsourcing[3] each judged context–meme pairs as *funny* or *not funny*. Quality-control questions filtered unreliable submissions; no personal information was collected. The funniness score for each pair is the proportion of annotators who judged it funny. We adopted the binary format after preliminary experiments showed that fine-grained scales degraded annotation quality (see Appendix A for interface, ethics, compensation, and scale comparison details).

Dataset Statistics. The dataset comprises 100,000 context-meme pairs with five annotations each (500,000 total). Overall Fleiss' κ is -0.022; because genuinely funny pairs are rare, agreement is dominated by negative cases, which deflates κ despite 65.1% of pairs reaching high raw consensus ($\geq$4 of 5 raters agreeing). The vote distribution nonetheless provides a useful graded signal for evaluating how often selected replies are perceived as funny (Appendix A).

5 Reply Selection Methods

We investigate two strategies: (i) similarity-based selection, which relies on indirect semantic matching, and (ii) preference-based selection, which directly models funniness given the context.

Similarity-based Selection (`sim-select`). This approach selects memes by cosine similarity in embedding space, using a text encoder f_c for the context and text or multimodal encoders f_m for the memes: $\hat{m}(c) = \arg\max_{m \in \mathcal{M}} \cos\big(f_c(c), f_m(m)\big)$. We implement f_m with four model types: text embeddings (Text-Emb), LLM embeddings (LLM-Emb), multimodal embeddings (Multi-Emb), and vision-language model embeddings (VLM-Emb). Text inputs use speech text alone or with the panel description; Multi-Emb uses only the image; VLM-Emb uses both speech text and panel image.

Preference-based Selection (`pref-select`). This approach directly estimates funniness via a preference function f_p that selects the meme judged funniest: $\hat{m}(c) = f_p(c, \mathcal{M})$. We implement f_p using LLMs prompted to choose the funniest meme in a multiple-choice format (Hendrycks et al., 2021; Robinson and Wingate, 2023) (Figure 9), using either speech text alone or with the panel description.

[2] https://densho810.com/free/. All data and annotations are in Japanese; English translations provided for clarity.

[3] https://crowdsourcing.yahoo.co.jp/

6 Experiments

Our experiments form a sequential investigation of the core challenges. We begin in Exp1 with a broad comparison of `sim-select` and `pref-select` methods to establish baseline performance. Finding `pref-select` promising but limited, we test a retrieve-and-rerank architecture in Exp2, which fails to improve performance. To understand this failure, Exp3 simplifies the task to test if models can select a clearly humorous meme from a set of non-humorous ones. After confirming that the models can, Exp4 directly tests our final hypothesis: that performance drops when models must choose from a pool of semantically similar candidates. A case study qualitatively analyzes model strategies.

6.1 Overall Settings

We used Score@1 as our primary metric, which measures the average proportion of annotators who judged the selected context–meme pair as funny. We also report nDCG for ranking evaluation; nDCG is available only for similarity-based methods, which produce a full ranking over all candidates, whereas preference-based methods output a single selected candidate in our prompting setup. For similarity-based selection, we evaluated multiple configurations across four model types, focusing on Japanese-optimized models (see Appendix B for the full set). We report top-performing models per category: Sarashina-Text-Emb and OpenAI-Text-Emb for Text-Emb; Calm2-LLM-Emb and Qwen2.5-LLM-Emb for LLM-Emb; LAION-CLIP-Multi-Emb and Jina-CLIP-Multi-Emb for Multi-Emb; EvoVLM-VLM-Emb and Sarashina-VLM-Emb for VLM-Emb. For preference-based selection, we used six production-grade LLMs: GPT-5, GPT-5-mini, GPT-OSS, Gemini 2.5 Pro, Gemini 2.5 Flash, and Claude 4 Sonnet (see Appendix B for model and compute details).

To anchor the evaluation, we define three parameter-free baselines. **Random** selects a meme uniformly at random (expected Score@1 $= 0.255$). **Prior** always selects the single meme with the highest population-level mean funniness (Score@1 $= 0.344$). **Oracle** selects, per context, the meme with the highest reference funniness (Score@1 $= 0.871$), serving as an upper bound.

6.2 Exp1: Main Results

Figure 2(a) shows that the `pref-select` method using GPT-5 achieved the highest Score@1 of 0.325 with panel descriptions. Among `sim-select` methods, text-based models (Sarashina-Text-Emb: 0.320, Calm2-LLM-Emb: 0.306) outperformed multimodal ones (Jina-CLIP: 0.282, EvoVLM: 0.263), indicating that visual information does not consistently help. Notably, all evaluated configurations remain below Prior (0.344) and above Random (0.255), far below the Oracle upper bound (0.871), indicating that no model reliably exploits context beyond global meme-level popularity. The top models from each family—GPT-5 and Sarashina-Text-Emb—achieve near-identical scores despite being fundamentally different approaches.

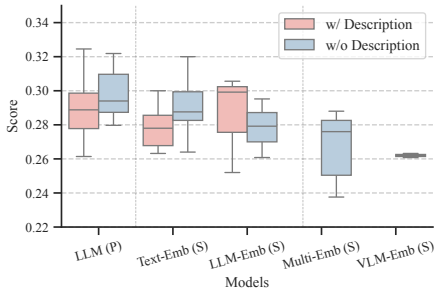
Figure 2b confirms this pattern: adding panel descriptions did not help `pref-select` or Text-Emb (scores slightly decreased), only LLM-Emb showed minor benefit, and score ranges largely overlapped. Among `pref-select` LLMs (Figure 2c), GPT-5 and GPT-5-mini avoided zero-score memes (<10% selection rate) and selected funny ones (>0.6) more often, while Claude 4 showed the opposite pattern. Among `sim-select` models (Figure 2d), Text-Emb and LLM-Emb avoided zero-score memes better than multimodal models, whose frequent selection of such memes suggests difficulty grasping contextual relevance from visual information. These results suggest that current models struggle to use visual information effectively, whether as text or pixels. We also conducted in-depth analyses including threshold sensitivity, significance tests, and feature-level profiling of selection differences (Appendix B and C).

6.3 Exp2: Retrieve-and-Rerank Approach

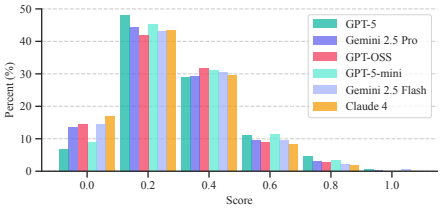
Settings We examined a hybrid retrieve-and-rerank approach, combining the `sim-select` and `pref-select` methods. This method first utilizes `sim-select` to retrieve the top-k similar candidates for a given context. A `pref-select` model then identifies the most humorous option from this set. For the retrieval step, we used Sarashina-Text-Emb, the top-performing similarity-based model from Exp1. For the selection step, we evaluated three large language models chosen from our Exp1 results to represent a range of top-performing proprietary and open models: GPT-5, Gemini 2.5 Pro, and GPT-OSS. To interpret their performance, we established two baselines: Random, which randomly selects a meme from the k candidates, and Oracle, which serves as an oracle upper bound by always choosing the candidate with the highest

Model	Desc	Score@1	nDCG@5
GPT-5 (P)	**Y**	**0.325 (±0.014)**	-
GPT-5 mini (P)	N	0.322 (±0.014)	-
Sarashina-Text-Emb (S)	**N**	**0.320 (±0.022)**	0.317
GPT-5 (P)	N	0.314 (±0.014)	-
OpenAI-Text-Emb (S)	N	0.307 (±0.024)	0.321
Calm2-LLM-Emb (S)	**Y**	**0.306 (±0.024)**	0.284
GPT-5 mini (P)	Y	0.300 (±0.013)	-
Qwen2.5-LLM-Emb (S)	Y	0.299 (±0.023)	0.278
Gemini 2.5 Flash (P)	N	0.296 (±0.014)	-
Calm2-LLM-Emb (S)	N	0.295 (±0.023)	0.294
Gemini 2.5 Pro (P)	Y	0.294 (±0.014)	-
GPT-OSS (P)	N	0.292 (±0.014)	-
Gemini 2.5 Pro (P)	N	0.286 (±0.014)	-
OpenAI-Text-Emb (S)	Y	0.285 (±0.024)	0.314
Sarashina-Text-Emb (S)	Y	0.285 (±0.023)	0.294
GPT-OSS (P)	Y	0.284 (±0.014)	-
Jina-CLIP-Multi-Emb (S)	**N**	**0.282 (±0.025)**	0.294
LAION-CLIP-Multi-Emb (S)	N	0.280 (±0.025)	0.289
Claude 4 (P)	N	0.280 (±0.013)	-
Qwen2.5-LLM-Emb (S)	N	0.279 (±0.023)	0.289
Gemini 2.5 Flash (P)	Y	0.276 (±0.014)	-
EvoVLM-VLM-Emb (S)	**N**	**0.263 (±0.022)**	0.262
Claude 4 (P)	Y	0.261 (±0.013)	-
Sarashina-VLM-Emb (S)	N	0.261 (±0.023)	0.280
Random	-	0.255 (±0.001)	0.271
Prior	-	0.344	-
Oracle	-	0.871	-

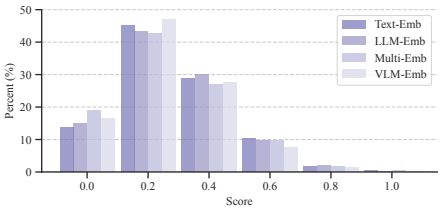
(a) Performance ranking (Exp1).



(b) By description



(c) By LLMs



(d) By embeddings

Figure 2: Main experimental results for Exp1. (a) **Table** showing the performance ranking across models and methods. S/P: similarity/preference-based; Y/N: with/without descriptions; values in parentheses denote 95% confidence intervals. (b)–(d) **Plots** of score distributions categorized by panel descriptions, LLMs, and embedding models respectively.

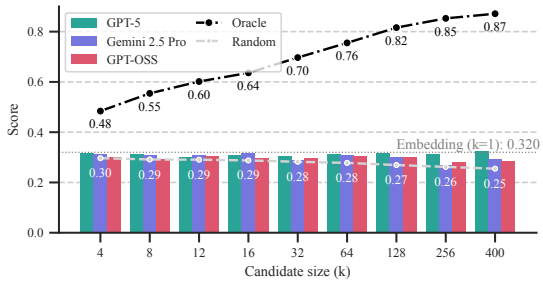


Figure 3: Performance of the retrieve-and-rerank approach (Exp2).

ground-truth funniness score within the set.

Results Figure 3 shows the results of our retrieve-and-rerank approach. There is a notable gap between the oracle and the `pref-select` models. The rising oracle indicates that retrieval successfully included high-quality candidates even for small k; for example, Score@1 reached 0.48 at $k = 4$ and 0.55 at $k = 8$. However, the models did not effectively leverage this advantage.

All `pref-select` models remained near or below the standalone `sim-select` baseline of 0.320. GPT-5 was stable around 0.31 regardless of k, while Gemini 2.5 Pro and GPT-OSS degraded at $k \geq 256$. GPT-5's stability contrasts with the typical *lost-in-the-middle* effect (Li et al., 2024; Liu et al., 2024), while the decline of others indicates context-length sensitivity. Overall, the limited improvement likely stems not from retrieval quality but from difficulty distinguishing subtle humor differences among similar candidates.

6.4 Exp3: Distinguishable Candidate Selection

Settings To understand why the retrieve-and-rerank approach from Exp2 was limited, we designed a controlled experiment with a simplified task. The goal was to test two hypotheses: (1) that models can reliably select a distinctly humorous meme from a set of less funny alternatives, and

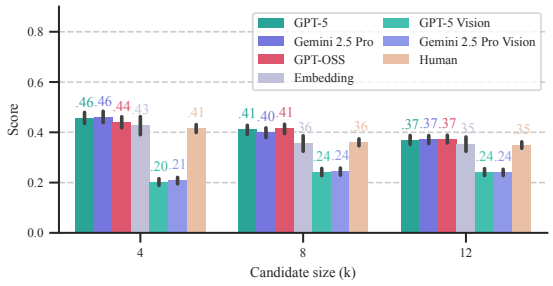


Figure 4: Performance on selecting from distinguishable candidates (Exp3). Human and vision-language models are included as baselines.

k	Model	Candidate Similarity			Δ
		Low	Mid	High	
4	GPT-5	0.457	0.459	0.437	-0.020
	Gemini 2.5 Pro	0.467	0.468	0.444	-0.023
	GPT-OSS	0.457	0.456	0.385	-0.072
8	GPT-5	0.396	0.416	0.379	-0.017
	Gemini 2.5 Pro	0.389	0.400	0.375	-0.014
	GPT-OSS	0.402	0.386	0.348	-0.054
12	GPT-5	0.365	0.365	0.356	-0.009
	Gemini 2.5 Pro	0.370	0.371	0.356	-0.014
	GPT-OSS	0.371	0.355	0.341	-0.030

Table 1: LLM performance under different candidate similarities. Δ represents the performance change between the Low and High similarity conditions. (Exp4).

(2) that performance would decrease as the number of candidates k increases. We constructed test sets where each contained one funny meme (score ≥ 0.8), half the candidates ($k/2$) were unfunny memes (score ≤ 0.2), and the remaining ones were somewhat funny ($0.2 <$ score < 0.8). Candidate sets were sampled with a fixed random seed and candidate order was randomized before prompting.

In addition to the models from our previous experiments, we included two additional baselines. First, we collected human performance data through a crowdsourced selection task to serve as a practical reference. Second, to assess the role of visual understanding, we evaluated vision-language models (GPT-5 Vision, Gemini 2.5 Pro Vision). For these models, all candidate images for a given context were combined into a single grid image, and the model was prompted to select the best option from this composite image.

Results Figure 4 shows that LLMs effectively identified the target funny meme (scores up to 0.46), and performance decreased with k (GPT-5: 0.46, 0.41, 0.37 for k=4, 8, 12). In this controlled setting, the `pref-select` models achieved reference scores comparable to the human selection baseline. Vision-language models performed poorly, indicating that visual humor understanding remains a challenge. However, this simplified setting limits generalizability: models can approach the human selection baseline when the humorous option is obvious, but their ability to handle more realistic scenarios remains undemonstrated.

6.5 Exp4: Sensitivity to Candidate Similarity

Settings In Exp3, we observed that LLMs performed well when funny memes were clearly distinguishable from less funny ones. Building on this observation, we examined how LLM performance changes when candidates become more similar, extending the controlled setting of Exp3. This follows the implication from Exp2 that the retrieve-and-rerank approach may have been limited by the high similarity among retrieved candidates, creating a *hard negative* scenario that makes it difficult for the reranker to identify the single best option.

To test this hypothesis, we modified the evaluation sets from Exp3. We retained the funny candidates (score ≥ 0.8) and resampled the other candidates based on their similarity to the funny one, using the similarity threshold t. We selected only candidates whose similarity score with the funny meme exceeded this threshold. We used three thresholds (t={-2,0,2}), based on the standardized cosine similarity from Sarashina-Text-Emb embeddings, denoting these conditions as Low, Mid, and High, respectively. The Low condition contains a diverse set of candidates, whereas the High condition requires the model to distinguish the funny meme from a pool of semantically similar yet unfunny alternatives. Contexts for which the required number of candidates could not be sampled under a given threshold were excluded from all conditions to keep evaluation sets matched.

Results Table 1 confirms that LLM performance decreased with candidate similarity, most pronouncedly at k=4; the effect diminished at larger k, suggesting that pool size itself becomes the dominant challenge. GPT-OSS was most affected, consistent with its lower capacity. These results explain the limited retrieve-and-rerank performance in Exp2: the retrieval step likely created highly similar candidate pools where LLMs could not discern the most humorous option.

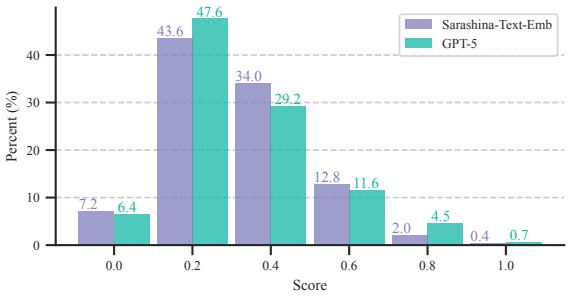


Figure 5: Score distributions of Sarashina-Text-Emb and GPT-5.

6.6 Case Study

This section analyzes how the `sim-select` and `pref-select` approaches create humor by examining their patterns of success and failure. We focus on the outputs of Sarashina-Text-Emb and GPT-5, the best-performing models from each method. An analysis of their overall performance in Figure 5 reveals two distinct behavioral patterns. Sarashina-Text-Emb is reliable but conservative; it tended to select memes that are moderately funny but semantically close to the context. In contrast, GPT-5 showed a high-risk, high-reward pattern, more frequently selecting very funny replies. A χ^2 homogeneity test (Appendix B) confirms that these two models reach statistically indistinguishable Score@1 ($p = 0.748$) via highly disjoint selection distributions (Cramér's $V = 0.832$, large effect). Appendix C provides a detailed feature-level analysis of this divergence. To understand the mechanisms that produce these different outcomes, we examine specific examples in Table 2.

The `sim-select` model, Sarashina-Text-Emb, succeeds when its keyword-matching mechanism produces unexpectedly complex humor like personification or irony. In case (A), a simple match for "*fast*" selects the meme "*HE'S JUST SO FAST...*" The humor arose not from the word match itself, but because the reply personifies the Wi-Fi with an overly dramatic, human-like reaction. This pattern continued in case (B), where a more subtle match for "*lost*" selects "*LATELY I DON'T UNDERSTAND MYSELF.*" This reply unexpectedly reinterprets a simple problem as a deep philosophical one, producing humor through its sudden exaggeration. Finally, in case (C), a literal match for "*I give up*" resulted in a blunt, agreeing reply ("*GIVE UP.*"). The humor here comes from its cold, dismissive tone, the opposite of the supportive reaction one would expect.

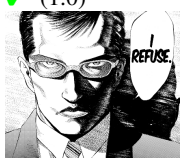
Context	Sarashina (Similarity-based)	GPT-5 (Preference-based)
(A) *The Wi-Fi at my new place is seriously fast!*	✓ (0.8) 	✓ (0.8) 
(B) *I think I'm lost...*	✓ (1.0) 	✗ (0.2) 
(C) *I'm done. I give up.*	✓ (0.6) 	✗ (0.2) 
(D) *Just found a 10-year-old pudding in the back of my fridge...*	✗ (0.2) 	✓ (0.8) 
(E) *Rumor has it that chocolate makes you 3x more productive when studying.*	✗ (0.2) 	✓ (0.8) 
(F) *Maybe life without a smartphone isn't so bad after all.*	✗ (0.2) 	✓ (1.0) 
(G) *Whoops, I gamed straight through to the morning.*	✗ (0.2) 	✗ (0.2) 

Table 2: Examples of selected replies by Sarashina-Text-Emb (similarity-based) and GPT-5 (preference-based). Numbers in parentheses represent funniness score.

In contrast, the `pref-select` model GPT-5 excelled at understanding context beyond simple keywords. This contextual inference is clear in case (D). While the `sim-select` model failed due to a superficial connection between "*fridge*" and "*cool*," GPT-5 inferred the dangerous nature of the "*10-*

year-old pudding" and provided the fitting reply, "*THIS IS POISON.*". Another key strength of GPT-5 was its ability to create a humorous persona. This was illustrated in case (E), where the simple association made by Sarashina-Text-Emb between "*studying*" and related keywords like "*memory*," "*history*," or "*subject*" was not effective. GPT-5, however, adopted the persona of a confident expert making an overconfident claim. The core of the humor lies in its statement, "*15 TIMES.*". It directly builds on the "*3x*" in the context, absurdly escalating it with a larger, baseless number. This overconfidence made the reply far funnier. Likewise, in case (F), the direct "*I REFUSE*" to living without a smartphone created another strong, humorous personality. In these instances, the humor arose from the character the reply embodies, making it more effective than the simple associations of the `sim-select` model.

However, both models can fail when a contextually relevant reply is too obvious and lacks an element of surprise. This is evident in case (G), where for the context "*gamed straight through to the morning*," both models selected the most predictable meme: "*I ONLY SLEPT 2 HOURS...*". While this reply is perfectly relevant, it is also highly predictable and therefore received a low humor score. This highlights a core challenge for automated humor generation: models need to distinguish simple relevance from genuine surprise.

In summary, these case studies show that the `sim-select` method provides reliable relevance but struggles to create high-impact humor consistently. The `pref-select` approach shows greater potential for more complex, context-aware humor, but must also avoid predictable responses.

7 Discussion

Implications for Computational Humor Our results highlight three challenges for modeling humor in online communication. First, no model reliably outperforms the Prior baseline that always selects the globally most popular meme, indicating that current models fail to exploit conversational context beyond global meme-level popularity. `pref-select` LLMs and `sim-select` methods achieve similar scores through different strategies—case studies suggest LLMs capture social cues such as irony and exaggeration, while embedding methods rely on surface-level semantic matching—yet neither approach translates these capabilities into consistently context-sensitive selection. Second, visual information – whether as text descriptions or direct images – does not consistently improve performance, pointing to a gap between *understanding* and *using* visual humor. While models can classify meme sentiment or explain humor (Sharma et al., 2023a), this understanding does not translate into selecting the wittiest reply from plausible candidates. Closing this gap may require objectives that couple visual recognition with contextual inference rather than simply scaling existing multimodal encoders. Third, models struggle most when candidates are semantically similar: human-baseline-comparable scores on clearly distinguishable choices (Exp3) decrease in realistic retrieve-and-rerank scenarios (Exp2) or under high candidate similarity (Exp4). The bottleneck lies in the final reranking step, and future work should target evaluation settings that capture these distinctions.

Scope and Future Directions Our benchmark is intentionally scoped to a single domain (freely licensed Japanese manga panels) and a binary funny/not-funny scale. This controlled setting isolates selection from generation and reduces subjective variance, but limits cultural and format generalizability. Extending to other meme formats (GIFs, reaction photos) and richer annotation scales (graded funniness, humor-type labels) is a key next step, as is validating these findings in naturalistic rather than synthetic contexts. Practical deployment of humor-aware agents will further require modeling user and cultural context, handling the temporal dynamics of meme popularity, and filtering inappropriate content in sensitive settings.

8 Conclusion

We proposed a new research direction for computational meme analysis, shifting the focus from static content to dynamic, interactive communication. To support this direction, we introduced the Meme Reply Selection task and the MAME-RE benchmark, providing a systematic framework for studying contextual humor in online conversations. Our analysis revealed that no model reliably outperforms a simple meme-prior baseline, visual information does not consistently improve selection, and models struggle to distinguish subtle differences in wit among similar candidates—establishing this task as a genuine open challenge. We release MAME-RE as a benchmark to measure progress on these challenges and to advance research on how AI systems participate in humor.

Limitations

This work has several limitations. First, our benchmark uses a single manga source (*Black Jack ni Yoroshiku*) under a free-use license; humor norms and visual conventions may differ across meme formats (photographic memes, GIFs, reaction images) and cultures. Second, the binary funny/not-funny annotation scale captures consensus but discards gradations of humor intensity. Third, all 250 contexts are synthetically generated; while topic coverage is broad, they may lack the spontaneity and community-specific register of authentic social media posts, and may favor models familiar with LLM-like text distributions. Fourth, we evaluate only selection (choosing from a fixed candidate set), not generation of novel meme replies. Finally, our annotator pool consists entirely of Japanese crowdworkers, and humor perception is known to vary across cultures and demographics.

Ethical Statement

All manga panels were taken from *Black Jack ni Yoroshiku*, a publicly distributed work under proper copyright conditions, and screened to exclude inappropriate content. Funniness annotations were collected via Yahoo! Crowdsourcing with explicit consent for research use; annotators were compensated according to the reward structure of the platform and were free to accept or skip tasks. No personally identifiable information is included; all context posts were generated by language models.

We acknowledge that the automatic generation or use of memes may potentially produce offensive or harmful content if not carefully moderated, and encourage future research and deployment to include robust safety and fairness measures.

Acknowledgements

We used Claude for content review and prose editing during manuscript preparation. All scientific claims, experimental design, and analysis were conducted by the authors.

References

Salvatore Attardo and Victor Raskin. 1991. Script theory revis(it)ed: Joke similarity and joke representation model. *Humor – International Journal of Humor Research*, 4(3-4).

Richard Bauman and Charles L Briggs. 1990. Poetics and performance as critical perspectives on language and social life. *Annual Review of Anthropology*.

Patrick Davison. 2012. The language of internet memes. *The Social Media Reader*, pages 120–134.

Eli Dresner and Susan C. Herring. 2010. Functions of the nonverbal in cmc: Emoticons and illocutionary force. *Communication Theory*, 20:249–268.

Zhengcong Fei, Zekang Li, Jinchao Zhang, Yang Feng, and Jie Zhou. 2021. Towards expressive communication with internet memes: A new multimodal conversation dataset and benchmark. *arXiv preprint arXiv:2109.01839*.

Alvan R Feinstein and Domenic V Cicchetti. 1990. High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549.

Lezandra Grundlingh. 2018. Memes as speech acts. *Social Semiotics*, 28(2):147–168.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *Proc. ICLR*.

Jack Hessel, Ana Marasović, Jena D. Hwang, Lillian Lee, Jeff Da, and Rowan Zellers. 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)*. Association for Computational Linguistics.

Tim Highfield and Tama Leaver. 2016. Instagrammatics and digital methods: studying visual social media, from selfies and gifs to memes and emoji. *Communication Research and Practice*, 2(1):47–62.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. MUTE: A multimodal dataset for detecting hateful memes. In *Proc. AACL-IJCNLP SRW*, pages 32–39.

Tianran Hu, Han Guo, Hao Sun, Thuy vy Thi Nguyen, and Jiebo Luo. 2017. Spice up your chat: The intentions and sentiment effects of using emojis. In *Proc. ICWSM*, pages 102–111.

EunJeong Hwang and Vered Shwartz. 2023. MemeCap: A dataset for captioning and interpreting memes. In *Proc. EMNLP*, pages 1433–1445.

Sophie Jentzsch and Kristian Kersting. 2023. ChatGPT is fun, but it is not funny! Humor is still challenging large language models. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*. Association for Computational Linguistics.

Jialun "Aaron" Jiang, Casey Fiesler, and Jed R. Brubaker. 2018. 'the perfect one': Understanding communication practices and challenges with animated gifs. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).

Saurav Joshi, Filip Ilievski, and Luca Luceri. 2024. Contextualizing internet memes across social media platforms. In *Companion Proc. WWW*, pages 1831–1840.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: detecting hate speech in multimodal memes. In *Proc. NeurIPS*.

Anushka Kulkarni. 2017. Internet meme and political discourse: A study on the impact of internet meme as a tool in communicating political satire. *Journal of Content, Community & Communication*, 6.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2024. LooGLE: Can long-context language models understand long contexts? In *Proc. ACL*, pages 16304–16333.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguistics*, 12:157–173.

Tyler Loakman, William Thorne, and Chenghua Lin. 2025. Who's laughing now? an overview of computational humour generation and explanation. In *Proc. INLG*.

Hua Lu, Zhen Guo, Chanjuan Li, Yunyi Yang, Huang He, and Siqi Bao. 2023. Towards building an open-domain dialogue system incorporated with internet memes. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 32:721–726.

Rada Mihalcea, Carlo Strapparava, and Stephen Pulman. 2010. Computational models for incongruity detection in humour. In *Proc. CICLing*, pages 364–374.

Khoi P. N. Nguyen and Vincent Ng. 2024. Computational meme understanding: A survey. In *Proc. EMNLP*, pages 21251–21267.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Detecting harmful memes and their targets. In *Findings of ACL*, pages 2783–2796.

Joshua Robinson and David Wingate. 2023. Leveraging large language models for multiple choice question answering. In *Proc. ICLR*.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 task 8: Memotion analysis— the visuo-lingual metaphor! In *Proc. SemEval*, pages 759–773.

Shivam Sharma, Siddhant Agarwal, Tharun Suresh, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2023a. What do you meme? generating explanations for visual semantic role labelling in memes. In *Proc. AAAI*.

Shivam Sharma, Firoj Alam, Md. Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022. Detecting and understanding harmful memes: A survey. In *Proc. IJCAI*, pages 5597–5606.

Shivam Sharma, Ramaneswaran S, Udit Arora, Md. Shad Akhtar, and Tanmoy Chakraborty. 2023b. MEMEX: Detecting explanatory evidence for memes via knowledge-enriched contextualization. In *Proc. ACL*, pages 5272–5290.

Jerry M. Suls. 1972. A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis. *The Psychology of Humor: Theoretical Perspectives and Empirical Issues*, pages 81–100.

Joseph B. Walther. 1992. Interpersonal effects in computer-mediated interaction: A relational perspective. *Communication Research*, 19(1):52–90.

Joseph B. Walther. 1996. Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication Research*, 23(1):3–43.

Xingyao Wang and David Jurgens. 2021. An animated picture says at least a thousand words: Selecting gif-based replies in multimodal dialog. In *Findings of EMNLP*, pages 3228–3257.

Yuan Wang, Yukun Li, Xinning Gui, Yubo Kou, and Fenglian Liu. 2019. Culturally-embedded visual literacy: A study of impression management via emoticon, emoji, sticker, and meme on social media in china. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Zhikun Wu, Thomas Weber, and Florian Müller. 2025. One does not simply meme alone: Evaluating co-creativity between llms and humans in the generation of humor. In *Proc. IUI*, pages 1082–1092.

Francisco Yus. 2018. Identity-related issues in meme communication. *Internet Pragmatics*, 1(1):113–133.

Jifan Zhang, Lalit K Jain, Yang Guo, Jiayi Chen, Kuan Lok Zhou, Siddharth Suresh, Andrew Wagenmaker, Scott Sievert, Timothy T. Rogers, Kevin Jamieson, Bob Mankoff, and Robert D Nowak. 2024. Humor in AI: Massive scale crowd-sourced preferences and benchmarks for cartoon captioning. In *NeurIPS Datasets and Benchmarks Track*.

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 *Proc. CVPR*, pages 13246–13257.

Yong Zhuang, Keyan Guo, Juan Wang, Yiheng Jing, Xiaoyang Xu, Wenzhe Yi, Mengda Yang, Bo Zhao, and Hongxin Hu. 2025. I know what you meme! understanding and detecting harmful memes with multimodal large language models. In *Proc. NDSS*.

A MAME-RE Dataset Details

Source Materials and Licensing. All manga images are sourced from *Black Jack ni Yoroshiku* (*Give My Regards to Black Jack*, SHUHO SATO), provided by Sato Manga Works Ltd. under a free-use license.[4] This license allows both commercial and non-commercial reproduction, public transmission, and a broad range of derivative uses, provided that attribution of the original title is maintained. The MAME-RE benchmark itself—contexts, panel metadata, and funniness annotations—is released under CC BY 4.0; the panel images remain governed by the free-use license above.

Annotation Procedure and Ethics. All human annotation was conducted on Yahoo! Crowdsourcing.[5] The annotation workflow consisted of three steps: (1) the annotator was shown a context–meme pair with the context text, meme image, and in-panel speech; (2) the annotator made a binary judgment (*funny* / *not funny*); and (3) the response was submitted. The interface (Figure 6, top) displayed these elements, while the full instruction text is shown in Figure 6, bottom. Quality control was ensured by including test questions with explicit instructions. Annotators were informed that their responses would be used for academic research, and consent was recorded through the platform interface. No personal information was collected; only aggregate, anonymized responses were used. We offered 10 JPY per batch of 50 annotations (average completion time: three minutes per batch). The reward was set with reference to comparable microtasks on the same platform.

Annotation Scale Choice. We initially considered a multi-level Likert scale for graded humor ratings. However, pilot annotations revealed that distinguishing fine-grained degrees of funniness was difficult for annotators, resulting in inconsistent and noisy labels. The higher cognitive load required for ranking also appeared to degrade annotation quality. In contrast, the binary judgment of funny or unfunny is intuitive and easy to answer, and aligns well with our goal of selecting a funny reply. Given these findings and the exploratory nature of this benchmark, we opted for the simpler binary approach.




```
[Overview]
- This task involves evaluating the humor of image replies
  to SNS posts.
- You will be shown an SNS post and an image reply to that post.
- Judge whether the image reply is funny or not as a response
  to the SNS post.
- Please answer intuitively, without overthinking.
- Please answer as instructed for quality control questions.
- If your answers differ from the instructions on these
  questions, your work may not be approved.

(Example)

[Agreement]
${CONSENT_STATEMENT}
```

Figure 6: Crowdworker annotation interface and full task instruction for the funniness scoring task in MAME-RE. Top: interface screenshot. Bottom: instruction text shown to annotators.

Data Statistics. Section 4 gives an overview of dataset scale and agreement; here we report additional distributional details. Table 3 lists extended statistics. The contexts and in-panel speech texts are concise, while the generated panel descriptions provide detailed visual cues (Figure 7).

The distribution of funniness scores for individual context-meme pairs is heavily skewed toward the low end (Figure 8, left), with a mode of 0.2, indicating that funny combinations are rare. In contrast, the per-context (center) and per-meme (right) averages both form bell-shaped distributions, suggesting that our dataset contains a well-balanced set of source materials.

Annotation Reliability and Threshold Robustness. Fleiss' $\kappa = -0.022$ (Table 3) is low, but this does not indicate that annotators responded randomly. The vote distribution (Table 4) shows

[4] https://densho810.com/free/
[5] https://crowdsourcing.yahoo.co.jp/



Figure 7: For this panel, the following description was generated — *Expression:* "eyes shut tightly, crying out loudly, conveying intense anger"; *Background:* "white background with speed lines"; *Effect:* "impact emphasized by speed lines, emotional expression"; *Tone:* "strong atmosphere of tension."

#Contexts	250
#Memes (panels)	400
#Context–meme pairs	100,000
Annotations per pair	5 (total 500,000)
Avg. context length (chars)	19.4 ± 4.8
Avg. speech text length (chars)	18.9 ± 10.6
Avg. panel description length (chars)	148.2
Fleiss' κ	-0.022
Mean pair score	0.27 ± 0.18
Unique annotators	2,325
Avg. annotations per annotator	215.1 ± 242.1
Mean per-annotator positive rate	0.27 ± 0.26

Table 3: Composition and annotation statistics of the MAMERE benchmark, covering dataset scale (top), content length (middle), and inter-annotator agreement and annotator behaviour (bottom).

an intuitive pattern: most context–meme pairs are not funny (63.7% at $0/5$ or $1/5$), and only a small fraction reach consensus as funny (1.5% at $4/5$ or $5/5$). This heavy skew toward the negative class is expected — genuinely funny combinations are rare. In such imbalanced binary settings, κ is known to be deflated because the high base rate of agreement by chance compresses the statistic (Feinstein and Cicchetti, 1990). Coefficients that are robust to prevalence give a different picture on the same votes: Gwet's AC1 $= 0.374$ and Brennan–Prediger $\kappa = 0.224$, against a majority-vote agreement rate of 0.771.

Furthermore, per-annotator positive rates (mean 0.266, SD 0.255) reveal systematic individual differences in what counts as "funny," rather than random noise. Different annotators apply different humor thresholds, producing structured disagreement that reflects the subjectivity of humor perception. We therefore retain disagreement as meaningful signal rather than adjudicating it away.

Score@1 evaluated on agreement-threshold subsets is reported in Appendix B.

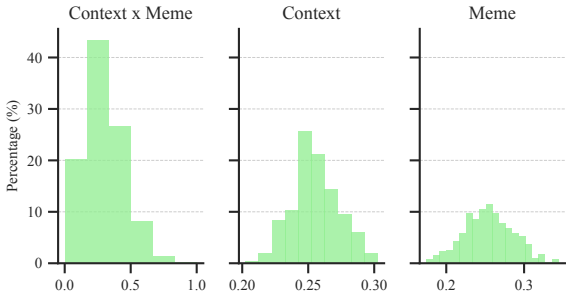


Figure 8: Funniness score distributions per context-meme pair (left), per context (center), and per meme (right).

Vote pattern (positive / 5)	#Pairs	Share
$5/5$ (unanimous funny)	109	0.11%
$4/5$ (high agreement funny)	1,369	1.37%
$3/5$ (3-vs-2 funny)	8,262	8.26%
$2/5$ (3-vs-2 not_funny)	26,600	26.60%
$1/5$ (high agreement not_funny)	43,392	43.39%
$0/5$ (unanimous not_funny)	20,268	20.27%

Table 4: Vote-pattern distribution over the 100,000 pairs.

B Experimental Details

Model Details. All models were evaluated with their official weights and default settings; no fine-tuning was performed. Similarity-based models score candidates by cosine similarity between context and meme embeddings (text, LLM, multimodal, or vision–language). Preference-based models directly select the funniest reply in a multiple-choice prompt (Figure 9). See Table 5 for the complete list of evaluated configurations. Specific model identifiers: sarashina-embedding-v1-1b and text-embedding-3-large for Text-Embs; CLIP-ViT-H-14-frozen-xlm-roberta-large-laion5B-s13B-b90k and jina-clip-v2 for Multimodal-Embs; Calm2-7b and Qwen2.5-7B for LLM-Embs; sarashina2-vision-8b and Llama-3-EvoVLM-JP-v2 for VLM-Embs; gpt-5-2025-08-07, gpt-5-mini-2025-08-07, gpt-oss-120b, gemini-2.5-flash, gemini-2.5-pro, and claude-sonnet-4-20250514 for LLMs. Preference-based selection used commercial APIs (OpenAI, Google, Anthropic); similarity-based models were run on a single NVIDIA A100 GPU (80 GB). API calls were made in April–May 2026. Post-hoc analyses used Python 3.11 with `sentence-transformers`, `transformers`, and `torch`; exact versions are pinned in the released code. All preference-based experiments used Japanese prompts; the English translation is shown in Figure 9.

```
You are skilled at creating funny meme replies (replies
using meme images) on Twitter.

[Instructions]
- You are given a post and meme candidates.
- Choose the funniest reply to the post from the candidates.
- Each meme is given in CSV format with columns: {item_fmt}
- Choose the single funniest candidate.

[Output format]
- Output only the ID.
- Do not include any extra text or markup.

[Candidates]
{meme_candidates}

---------------
Post: {context}
```

Figure 9: Prompt template (English translation). `{item_fmt}` is "*id, speech*" or "*id, speech, description*"; `{meme_candidates}` contains the candidates in the corresponding CSV format.

Confidence Intervals. All reported 95% confidence intervals are percentile bootstrap intervals over the 250 shared contexts: contexts are resampled with replacement ($B = 10{,}000$, seed 42) and the interval spans the 2.5th and 97.5th percentiles of the resampled Score@1 values.

Human Selection Baseline (Exp3). The human reference in Figure 4 was collected on Yahoo! Crowdsourcing over the same 249 evaluation contexts, the same candidate sets, and the same scoring table used for the models. For each $k \in \{4, 8, 12\}$, 125 workers made 1,250 selections (five votes per item), with the candidate panels presented as images. Score@1 is 0.414 $[0.395, 0.433]$ for $k = 4$, 0.361 $[0.346, 0.376]$ for $k = 8$, and 0.349 $[0.336, 0.362]$ for $k = 12$.

Score@1 by Agreement Threshold. To verify that model rankings do not depend on the chosen agreement cut-off, we recompute Score@1 on three pair subsets: (i) the *full* set of 100,000 pairs over 250 contexts; (ii) a *high-agreement* subset of 65,138 pairs (65.1%) whose ground-truth funniness label is supported by ≥ 4 of 5 raters (0/5 + 1/5 + 4/5 + 5/5 patterns); and (iii) a *unanimous* subset of 20,377 pairs (20.4%; 0/5 + 5/5). Table 6 reports Score@1 of the best preference-based model (GPT-5 with description) and the best similarity-based model (Sarashina-Text-Emb, text condition) on each subset. The per-model n_{ctx} (number of contexts in which the model has at least one prediction landing in the subset) varies with the subset.

Model	Desc	Score@1 (95% CI)
gpt-5-2025-08-07 (P)	Y	0.325 (±0.014)
gpt-5-mini-2025-08-07 (P)	N	0.322 (±0.014)
sarashina-embedding-v1-1b (S)	N	0.320 (±0.022)
gpt-5-2025-08-07 (P)	N	0.314 (±0.014)
text-embedding-3-large (S)	N	0.307 (±0.024)
calm2-7b (S)	Y	0.306 (±0.024)
gpt-5-mini-2025-08-07 (P)	Y	0.300 (±0.013)
amber-large (S)	Y	0.300 (±0.023)
Qwen2.5-7B (S)	Y	0.299 (±0.023)
plamo-embedding-1b (S)	N	0.297 (±0.023)
gemini-2.5-flash (P)	N	0.296 (±0.014)
calm2-7b (S)	N	0.295 (±0.023)
gemini-2.5-pro (P)	Y	0.294 (±0.014)
gpt-oss-120b (P)	N	0.292 (±0.014)
GLuCoSE-base-ja-v2 (S)	N	0.288 (±0.025)
CLIP-ViT-B-32-xlm-roberta (S)	N	0.288 (±0.025)
GLuCoSE-base-ja-v2 (S)	Y	0.288 (±0.024)
amber-large (S)	N	0.287 (±0.023)
gemini-2.5-pro (P)	N	0.286 (±0.014)
llm-jp-clip-vit-large-patch14 (S)	N	0.286 (±0.024)
text-embedding-3-large (S)	Y	0.285 (±0.024)
sarashina-embedding-v1-1b (S)	Y	0.285 (±0.023)
gpt-oss-120b (P)	Y	0.284 (±0.014)
ruri-v3-310m (S)	N	0.283 (±0.022)
jina-clip-v2 (S)	N	0.282 (±0.025)
LaBSE (S)	N	0.281 (±0.024)
CLIP-ViT-H-14-frozen-xlm-roberta (S)	N	0.280 (±0.025)
claude-sonnet-4-20250514 (P)	N	0.280 (±0.013)
Qwen2.5-7B (S)	N	0.279 (±0.023)
gemini-2.5-flash (P)	Y	0.276 (±0.014)
japanese-clip-vit-b-32-roberta-base (S)	N	0.272 (±0.023)
plamo-embedding-1b (S)	Y	0.271 (±0.022)
multilingual-e5-large (S)	Y	0.269 (±0.024)
LaBSE (S)	Y	0.265 (±0.021)
multilingual-e5-large (S)	N	0.264 (±0.023)
Llama-3-EvoVLM-JP-v2 (S)	N	0.263 (±0.022)
ruri-v3-310m (S)	Y	0.263 (±0.023)
claude-sonnet-4-20250514 (P)	Y	0.261 (±0.013)
Llama-3.1-Swallow-8B-v0.2 (S)	N	0.261 (±0.023)
sarashina2-vision-8b (S)	N	0.261 (±0.023)
Llama-3.1-Swallow-8B-v0.2 (S)	Y	0.252 (±0.022)
clip-japanese-base (S)	N	0.252 (±0.023)
japanese-stable-clip-vit-l-16 (S)	N	0.246 (±0.024)
clip-ViT-B-32-multilingual-v1 (S)	N	0.238 (±0.023)

Table 5: Full results for all models evaluated in Exp1 (Section 6.2). S = similarity-based; P = preference-based; Desc = panel description provided.

C How Do LLM and Embedding Selections Differ?

The case study (Section 6.6) showed that GPT-5 and Sarashina-Text-Emb achieve near-identical Score@1 yet select almost entirely different memes. Here we ask *how* they differ, examining three dimensions: visual features, text features, and humor type.

Method. We compare the memes selected by each route against a *funny-pool baseline*: the feature distribution over the 1,478 high-agreement funny pairs ($\geq 4/5$). Deviations from this baseline reveal route-specific preferences beyond what the dataset structure would produce by chance. Each meme is labelled on 34 axes by Gemini Flash Lite. The axes cover visual properties (face count, expression, gaze direction, character count, pose, attire, interaction type, image tone, background complexity, panel cropping, symbolic effects, screen-

Subset	#Pairs	GPT-5	Sarashina
Full	100,000	**0.325**	**0.320**
$\geq 4/5$	65,138	**0.230**	**0.202**
$5/5$ (unan.)	20,377	0.074	0.053

Table 6: Score@1 across agreement-threshold subsets. $\geq 4/5$ includes both positive and negative high-agreement patterns ($0/5$, $1/5$, $4/5$, $5/5$). Full and $\geq 4/5$ are ranking-stable; $5/5$ is a small-sample stress test (per-model $n_{\text{ctx}} = 19$–36), not a reporting baseline.

tone density, bubble shape, text position) and text properties (speech length, text role). Humor types are labelled on each context–meme pair with the image attached.

Q1: Do the routes differ in visual preferences? Table 7 shows clear divergence. LLMs prefer memes with serious tone, viewer-directed gaze, and minimal visual effects — a "clean, direct" visual style. Embeddings prefer memes with distress expressions, upper-body poses, and more complex compositions.

Visual feature	Base	LLM	Emb
`image_tone=serious`	.589	.747	.643
`gaze=toward_viewer`	.438	.550	.486
`symbolic_effect=none`	.702	.862	.786
`face_expression=distress`	.357	.221	.438
`character_pose=upper_body`	.474	.367	.576
`panel_cropping=cropped`	.698	.484	.419

Table 7: Visual-feature pick rates (selected). *Base* = funny-pool baseline. $n_{\text{LLM}} = 286$; $n_{\text{Emb}} = 210$.

Q2: Do the routes differ in text preferences? LLMs avoid memes with long speech and prefer monologues, while Embeddings favor long speech and avoid monologues (Table 8). This likely reflects how each route processes text: LLMs reason about content, favoring concise, punchy lines; Embeddings match surface similarity, benefiting from more text to align with the context. LLMs also strongly avoid question-marked speech (.200→.052) and self-referential text, while Embeddings over-pick mixed text roles.

Q3: Do the routes differ in humor type? Each funny pair was labelled on 7 humor-type axes (Table 9) by Gemini Flash Lite with the meme image attached.

Table 10 shows that the two routes prefer different humor types: LLMs under-pick `exaggeration` while Embeddings under-pick `deadpan`.

On the context side, the largest divergence is `failure_embarrassment`: Embeddings over-represent this topic among their funny hits (+8.4pp above baseline) while LLMs under-represent it (−10.4pp), consistent with LLMs' avoidance of deadpan and self-deprecation humor.

Text feature	Base	LLM	Emb
`text_role=dialogue`	.592	.479	.471
`speech_length=long`	.367	.238	.481
`text_role=monologue`	.357	.420	.290
`speech_length=short`	.294	.392	.243
`has_question_mark`	.200	.052	.119
`contains_self_reference`	.198	.084	.110
`text_role=mixed`	.022	.049	.200

Table 8: Text-feature pick rates. Notation as in Table 7.

Axis	Definition
`sarcasm`	Surface meaning opposite to intent.
`exaggeration`	Overstating magnitude of the event/feeling.
`absurdity`	Bizarre / non-sequitur logic.
`deadpan`	Suppressed reaction, stoic tone.
`self_deprecation`	Mocking / belittling oneself.
`parody_recontext`	Applying a serious source to a mundane context.
`relatable_humor`	Capturing a universally relatable feeling.

Table 9: Humor-type taxonomy (7 multi-label axes).

D Dataset Examples

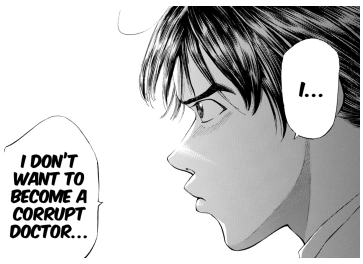
Table 11 shows examples of context-meme pairs with different funniness scores. Scores range from 0.0 (all five annotators rated not funny) to 1.0 (all rated funny). Each block corresponds to a single context, where the associated memes are arranged by their average funniness.

Humor type	Base	LLM	Emb
exaggeration	.865	.811	.905
relatable_humor	.624	.661	.543
deadpan	.431	.374	.333
absurdity	.355	.283	.333
self_deprecation	.300	.220	.305
sarcasm	.058	.077	.024

Table 10: Humor-type pick rates on the funny pool. parody_recontext ($\geq$ .99) omitted. $n_{\text{LLM}} = 286$, $n_{\text{Emb}} = 210$.

"Woke up and it's already evening."

0.0



"I DON'T WANT TO BECOME A CORRUPT DOCTOR..."

0.2



"I ONLY SLEPT 2 HOURS!"

0.4



"REALLY? THERE WAS NOTHING WE COULD DO–NOTHING!?"

0.6



"WHAT ARE YOU DOOOING!?"

0.8



"BEAT, DAMN YOU!"

1.0



"OH, MAN. THIS AIN'T GONNA WORK."

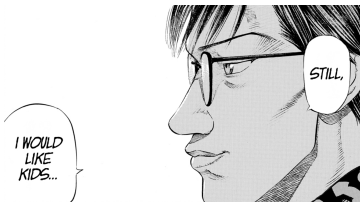
"I want to be a cat."

0.0



"THEY'RE LIKE ALIEN BEINGS..."

0.2



"STILL, I WOULD LIKE KIDS..."

0.4



"WHAT THE HELL ARE YOU DOING, ANYWAY!?"

0.6



"ENOUGH!!"

0.8



"I CAN'T STOP THIS."

1.0



"YOU'RE A LOT LIKE ME WHEN I WAS YOUNGER."

Table 11: Examples from the MAME-RE benchmark. Each block shows six memes selected for one context, arranged by funniness score (0.0 = unanimous not funny; 1.0 = unanimous funny). Scores reflect the proportion of annotators who judged the pair as funny. Speech text below each panel is the English translation of the dialogue.