Not Funny Anymore: LLM Judges Confuse Literal Similarity for Humor in Translated Jokes

Fabricio Rivera[1,2*] Rohit Pochugari[1*] Tessa Chan[1] Devansh Katakwar[1]
Kevin Zhu[†] Michael Saxon[1,3†]
[1]Algoverse AI Research [2]Andrews University
[3]University of Washington
riveraperez@andrews.edu, rohit.pochugari@gmail.com, tessalwchan@gmail.com,
devanshkatakwar07@gmail.com

Abstract

Automatic humor translation is both a challenging task and a very difficult problem to evaluate. Reference-based metrics struggle in assessing humor preservation in joke translation, often rewarding towards literal similarity over the preserved comedic effect, and they require costly manual gold reference translations. In this work, we study the task of *reference-free humor translation evaluation*, and analyze the performance of LM judges using 7 models on 162 English-to-Chinese joke pairs and 76 English-to-Hindi joke pairs with 5-point Likert scale human annotations. We find that these judges struggle, with strict agreement often near or even below the 20% random baseline. To better understand this limitation, we test the hypothesis that these metrics are *over-attending to literalness* as a signal for quality by introducing a correlation-based literalness metric in a multilingual embedding space. With this novel analysis we demonstrate quantitatively that poor LM evaluator performance is in fact driven by this over-literal bias, suggesting that future metrics which explicitly contend with this literalness might close this gap.

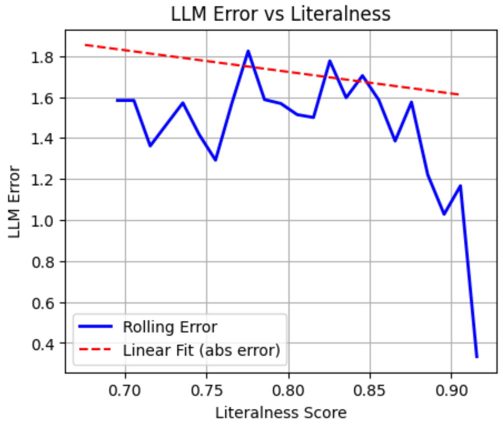

Figure 1: The (smoothed) modal error between LM and human humor judgments against *literalness*. LM judgments converge toward human ones on more literal translations.

Introduction

Humor is one of the most culturally and linguistically sensitive forms of communication. With its reliance on mechanisms such as puns, wordplay, and cultural references, literal-leaning translations struggle to consistently preserve the humor of the joke. Widely used neural machine translation (NMT) metrics, including BLEURT and COMET were trained to track segment-level adequacy and fluency against human labels on general-domain text, making it limited in domains where more flexibility with sentence structure is needed in translations (Sellam, Das, and Parikh 2020; Rei et al. 2020)

Recently, large language models (LLMs) have been proposed as automatic judges for MT, achieving SoTA-level performance on general MT evaluation with and without translation references (Kocmi and Federmann 2023). Controlled studies show LLM evaluators derive most of their evaluation from the reference and can become less accurate when the source is included, thus showing evidence of **weak cross-lingual reasoning** and possible problems when translation requires non-literal adaptations that diverge from a single reference (Huang et al. 2024). Second, LLM-as-a-judge exhibits self-preference. LLMs over-score outputs that look familiar to them, which reflects biased ratings toward safe, literal phrasing over creative wordplay, not ideal for cross-lingual reasoning. (Wataoka, Takahashi, and Ri 2025).

Given these limitations, our research asks: *Can large language models reliably evaluate whether humor is preserved in cross-lingual joke translation based on cross-lingual reasoning?* To answer this we do the following:

1. **Benchmarking Framework**: To the best of our knowledge, we propose the first evaluation framework designed to assess humor preservation in machine translation, lever-

[*] Equal Contribution

[†] Senior Author

[0] Source Code and Dataset at https://github.com/Ftsos/HumorLiteralnessPaper

aging the LLM-as-a-Judge paradigm.

2. **Novel Dataset**: We introduce the first publicly available dataset for humor preservation benchmarking, consisting of English-Chinese joke translations rated by both humans and multiple LLMs. The dataset will contain individual human ratings and the models' ratings, for each method in this paper. This dataset will be publicly released in Hugging Face upon acceptance.
3. **Broad analysis**: We systematically compare human judgments with evaluations from 21 LM judgments, measuring strict agreement, binary accuracy, and correlation metrics.
4. **Literal wordiness feature**: We introduce a Token-level semantic diagnostic that reveals why LLM-based evaluators overrate word-for-word translations.

Related Work

Research on humor in NLP processing has traditionally focused on pun recognition (Yang et al. 2015; Miller, Hempelmann, and Gurevych 2017) or "humor detection" (identifying if a passage is a joke or not) tasks (Meaney et al. 2021; Weller and Seppi 2019). More recently humor understanding work has centered on LLMs (Loakman, Thorne, and Lin 2025; Hessel et al. 2023) and broadening evaluation to specific kinds of jokes. Their findings suggest that while models can recognize surface-level humor cues, their interpretive and evaluative abilities remain limited.

In this work, we expand into *multilingual humor evaluation*, specifically the task of joke translation using LM judges. Machine translation evaluation typically focuses on evaluating generalized translation quality (Huang et al. 2024), here we focus specifically on *how funny the translated joke is*, a novel dimension of evaluation in a reference-free setup.

Humor translation presents unique challenges because jokes often rely not just on meaning but also on how meaning is constructed—through wordplay, cultural knowledge, phonetics, and context. This theory is backed up in empirical research. Mohebbi (2023) studied Persian-to-English joke translation, finding that models and humans both default to literal interpretations, which often erode the comedic effect. (Taylor, Herbert, and Sana 2025) found that while LLMs may generate plausible translations, their assessments of humor diverge significantly from humans evaluators, particularly around wordplay or cultural knowledge. Even in the task of humor detection, Gabriella (2020) demonstrates that neural models over-attend to surface-level incongruities in humor.

Our work synthesizes these directions by **formalizing humor evaluation** as a task in machine translation and **uses 'literalness' as a lens** for understanding the challenges of using LM judges for humor evaluation.

Methodology

We propose a benchmarking framework designed to evaluate the capabilities of a given SoTA LLM model on humor translation quality. We employ LLM-as-a-Judge to rate the quality of joke translations. These ratings are later compared with the human annotators' data. This framework provides a way to assess the degree to which the LLMs humor translation quality metrics mirror human perception.

Dataset

We sample 162 jokes from the Kaggle short-jokes dataset and translate them into Mandarin using GPT-4o-mini.[1] Four bilingual native speakers served as annotators for Mandarin which had a Krippendorff's α of 0.776. Each annotator rated all jokes, providing scores on a five-point Likert scale based on semantic preservation of humor in the translation. In parallel, SoTA LLMs were tasked with the same evaluation, using different prompting strategies but restricted to the original joke and its translation as input. The same process was repeated for Hindi with a reduced sample size of 76 jokes and 3 annotators which had a Krippendorff's α of 0.776 (change number). The resulting dataset thus consists of $n = 162$ joke–translation and $n = 76$ joke-translation pairs for Mandarin and Hindi respectively, each rated by humans and 7 different LLMs. All annotators were volunteering friends of the authors.

Experimental setup

To assess the performance of LLM judges on this task, we test variants of Claude, GPT-4o, Gemini 2.5, Qwen3, LLaMA 4, Mistral 7B instruct, and DeepSeek (model version details in appendix).

Each model is queried with three different prompting strategies: `vanilla`, a simple prompt, `chain-of-thought` (Wei et al. 2023), and `self-consistency` prompting (Wang et al. 2023). The prompts are provided in the appendix. For each joke pair (x_i, y_i), the prompt instructs the model to output a single integer score $r_i \in \{1, 2, 3, 4, 5\}$ with criteria explicitly defined to prioritize humor preservation over surface lexical similarity.

All models are prompted independently, producing a sequence of LLM-generated ratings $\mathbf{r}^{(\text{LLM})} = (r_1, r_2, \ldots, r_N)$, which are later compared directly against the human annotation vector $\mathbf{h} = (h_1, h_2, \ldots, h_N)$.

Token-Level Semantic Alignment Evaluation

To complement the LLM-as-a-Judge assessment, we propose a token-level semantic alignment metric to evaluate structural and contextual similarity between the original and translated jokes. The objective of this metric is to contrast LLM bias for literal translations, due to the fact that most humorous translations will require non-literal adaptations. The approach consists of the following steps:

1. **Embedding:** Each joke x_i and its translation y_i are tokenized and encoded using the multilingual model `xlm-r-100langs-bert-base-nli-stsb-mean-tokens`. Let
$$X_i = [\mathbf{x}_1, \ldots, \mathbf{x}_m], \quad Y_i = [\mathbf{y}_1, \ldots, \mathbf{y}_n]$$
denote the L2-normalized token embeddings.
2. **Procrustes Alignment:** An orthogonal transformation matrix W is estimated to align the source (English) and target (Chinese) embedding spaces by minimizing:
$$W^* = \arg\min_W \|XW - Y\|_F \quad \text{s.t. } W^\top W = I$$

[1] `https://www.kaggle.com/datasets/abhinavmoudgil95/short-jokes`

The closed-form solution is obtained via singular value decomposition (SVD).

3. **Token Matching:** After alignment, cosine similarities are computed:

$$s_{ij} = \frac{(W\mathbf{x}_i)^\top \mathbf{y}_j}{\|W\mathbf{x}_i\| \, \|\mathbf{y}_j\|}$$

   Tokens are then matched using the Hungarian algorithm to maximize the overall similarity across all pairs.

4. **Per-Joke Scoring:** For each joke–translation pair, the mean token-level cosine similarity is computed:

$$\text{score}_i = \frac{1}{k_i} \sum_{(p,q) \in M_i} s_{pq}$$

   where M_i denotes the optimal token matches and $k_i = |M_i|$.

This framework quantifies cross-lingual embedding alignment. Higher scores indicate translations that closely follow the source in wording and structure (i.e., more literal), while lower scores reflect freer adaptations. This metric then allows us to extract *literalness/wordiness metric* of a translation. We use this value throughout the paper.

Evaluation Metrics

To quantify the alignment of human and LLM, we will use Strict Agreement, which is the ratio of exact matches in the dataset; for this metric the random-guess baseline would be at $\frac{1}{5} = 0.2$. To implement, we introduce a boolean variable v_i, Where $v_i = 1$ if the ratings are an exact match between the values or 0 if they're not a match. We repeat this process for a variety of LLMs and prompting techniques. Then we aggregate the number of matches and divide by the total number of jokes n to get the accuracy percentage. This can be described as:

$$Accuracy = \frac{\sum_1^n v_i}{n}$$

We also implement ±1 accuracy, MAE, and spearman correlation.

Results and Discussion

Table 1: Meaning retention strict accuracy (%) across models and prompting strategies.

Model	Vanilla	CoT	Self-consistency	Avg.
Anthropic	11.2	10.6	13.9	11.9
Gemini	11.4	12.4	23.5	15.8
OpenAI	8.1	9.9	9.3	9.1
Qwen-3	9.3	11.2	7.6	9.4
DeepSeek-3	8.1	7.5	14.4	10.0
Llama-4	13.0	8.8	11.2	11.0
Mistral	15.5	10.7	13.4	13.2

Our results reveal a significant deficiency in the ability of current LLMs to evaluate translation quality in jokes. Spearman and Pearson correlations are low (mostly between –0.15 and 0.27), indicating that models poorly track human rank order of funniness; this holds across prompts. Across all tested models and a variety of prompting techniques, the strict accuracy hovered around the 20% random baseline for a 1-to-5 Likert scale. Furthermore, a binary evaluation corroborated these results, showing substantial misalignment, with performance failing to rise above random chance. This poor performance aligns with prior work from (Zangari et al. 2025) on how LLMs process humor, and it presents a stark challenge to the optimism in the field. Despite the hope, as articulated by (Taylor, Herbert, and Sana 2025), that LLMs trained on vast datasets would be able to capture the "semantic and linguistic incongruities of humor" and overcome the tendency of older models to "destroy the wordplay," our findings suggest this potential is limited by the ability for models to effectively do cross-lingual reasoning. A big part of the failure seems to be the fundamental bias toward literal semantic content, failing to evaluate whether the spirit of the joke, not just the meaning of its words, has survived translation.

Our results also show that a practical path to solve this problem is to explicitly provide literalness signals that allow the LLM to identify biases it would otherwise default to and help correct a tendency to over-rely of surface similarity producing ratings that more closely align with human ratings

Error Analysis

Our analysis of human evaluations reveals consistent patterns that distinguish translations that preserve humor from those that fail due to lost wordplay. Our comprehensive error analysis reveals a fundamental misalignment in LLM-as-a-Judge, driven by its tendency to reward token-level literalness. Vanilla LLMs, despite their sophisticated language capabilities, tend to over-index on token-level literalness, which conflates faithful translation with preserved humor, leading to a consistent leniency bias toward failed translations. This systematic misalignment is evident when we compare the average vanilla LLM rating with the corresponding Wordiness Score ($\mathcal{W}$) for the translations.

For example, we observe significant misalignment in jokes that rely on common idioms, which lose their double meaning when translated. Consider the following joke:

English:

"I find like geologers make really good friends...they are very down to Earth."

Chinese:

我发现，地质学家特别适合做朋友......因为他们特别“脚踏实地”。
(I found that geologists are particularly suitable for friends...because they are particularly "down to earth.")

This joke relies on the double meaning of the idiom "down to earth." The Chinese translation uses a figurative equivalent, "脚踏实地" (jiǎotàshídì), meaning "steadfast" or "sensible". Since the translation is semantically sound and has a high Wordiness Score ($\mathcal{W} = 0.875$), the vanilla LLMs assign a high average rating (4.57). However, in Chinese, the idiom no longer contains the reference to the Earth needed for the

Model	Accuracy			Precision			Recall			F1		
	Vanilla	CoT	SelfCons	Vanilla	CoT	SelfCons	Vanilla	CoT	SelfCons	Vanilla	CoT	SelfCons
Anthropic	36.0	36.6	63.4	14.2	15.5	16.4	72.7	81.8	40.0	23.7	26.1	23.4
Gemini	31.6	29.8	35.3	15.2	15.2	10.0	90.5	93.8	33.3	26.0	26.1	15.4
OpenAI	27.4	24.8	17.4	14.8	15.4	14.2	90.9	100.0	100.0	25.5	26.7	24.9
Qwen-3	48.5	24.8	34.5	17.9	14.9	16.4	77.3	95.5	85.7	29.1	25.8	27.5
DeepSeek-3	30.4	21.1	26.3	14.3	14.3	15.2	81.8	95.5	95.5	24.3	24.9	26.3
Llama-4	37.9	22.1	19.3	16.9	15.2	14.5	90.9	100.0	100.0	28.6	26.4	25.3
Mistral	75.8	47.1	51.9	24.2	15.2	7.5	36.4	55.6	25.0	29.1	23.8	11.6

Table 2: Binary classification metrics (%) across prompting strategies.

Model	±1 Accuracy			MAE			Spearman			Pearson		
	Vanilla	CoT	SelfCons	Vanilla	CoT	SelfCons	Vanilla	CoT	SelfCons	Vanilla	CoT	SelfCons
Anthropic	39.8	39.1	49.7	1.68	1.71	1.45	0.04	0.02	0.03	0.01	0.07	0.03
Gemini	32.9	29.8	35.3	1.88	2.26	2.18	0.12	0.25	0.10	0.12	0.23	0.01
OpenAI	28.6	21.7	16.1	2.06	2.39	2.53	0.17	0.21	0.19	0.18	0.22	0.20
Qwen-3	41.6	26.1	32.4	1.63	2.34	2.01	0.18	0.27	0.12	0.16	0.25	0.10
DeepSeek-3	32.9	24.8	31.9	1.82	2.34	1.93	0.08	-0.12	0.10	0.09	-0.06	0.16
Llama-4	44.1	19.1	22.9	1.67	2.50	2.39	0.14	0.25	0.14	0.19	0.24	0.19
Mistral	61.5	46.3	48.0	1.31	1.73	1.64	0.01	-0.08	-0.15	0.00	-0.10	-0.18

Table 3: ±1, MAE, Spearman, and Pearson, across prompt strategy

geological pun, resulting in a low human rating (2.0). This example confirms that LLM-as-a-Judge fails to recognize the loss of the humor, and blindly rewards literal translations. The overreliance on surface-level language comprehension drives the observed leniency bias toward failed translations.

In contrast, when a joke's comedic mechanism is content-based rather than language-based, the LLM ratings successfully align with human annotations. The following joke relies on universal self-deprecating humor instead of linguistic ambiguity, resulting in a strong alignment:

English:

"I'm glad it's the thought that counts because I spend all day thinking about the shit I should be doing."

Chinese:

我很庆幸“心意最重要”，毕竟我整天都在想着那些我该做却没做的事。

(I'm so glad that "the thought is what matters most" because I spend all day thinking about the things I should have done but didn't.)

In this example, the high Wordiness Score ($\mathcal{W} = 0.796$) is concurrent with a successful transfer of the humorous concept. The Average vanilla LLM rating (3.57) matches closely with the human rating (4.0), showing a difference of only 0.43 on a 5-point scale. This alignment suggests that for jokes whose humor is semantic rather than pun-based, an LLM's token-level fidelity can become effective for judging humor preservation. This suggests that the misalignment issue is isolated to translations where the humor is rooted in linguistics and not content.

Relationship between performance and literalness

These relationships are demonstrated at scale in Figures 2 and 1. They plot a smoothed (sliding window) of a modal value (between models) as a function of the literalness score. It is computed by taking the arithmetic mean of all values within windows of width 0.03 of literalness with a step size of 0.01. LM Error is calculated as the difference between the average value for a given pair, and the human rater's mode.

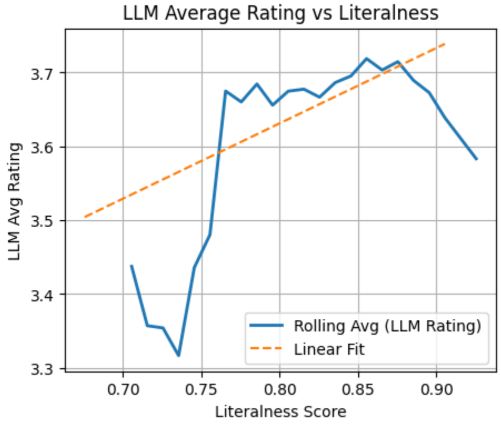

Figure 2: Mode LLM Rating as Literalness Increases

Figure 2 shows the relationship between literalness and the *average modal score* assigned by the models to each joke translation. As the jokes become more literally translated, the LMs begin to score them higher. Conversely, less literal translations receive low scores despite sometimes being preferred by humans.

Conclusion

We examine the limitations of LLMs for assessing humor preservation in cross-lingual joke translation. We do so by building a Token-level semantic alignment evaluation. Our work highlights a specific failure mode in cross-lingual reasoning. Future work can use our literalness feature to reduce this internal bias, and ultimately improve LLM support for underserved linguistic communities.

References

Anthropic PBC. 2025. Introducing Claude Sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5. Accessed: 2025-11-16.

DeepSeek-AI. 2025. DeepSeek-V3 Technical Report. arXiv:2412.19437.

Gabriella, K. 2020. Translating Humour A Didactic Perspective. *Acta Universitatis Sapientiae, Philologica*, 12: 68–83.

Hessel, J.; Marasović, 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. arXiv:2209.06293.

Huang, X.; Zhang, Z.; Geng, X.; Du, Y.; Chen, J.; and Huang, S. 2024. Lost in the Source Language: How Large Language Models Evaluate the Quality of Machine Translation. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 3546–3562. Bangkok, Thailand: Association for Computational Linguistics.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. arXiv:2310.06825.

Kocmi, T.; and Federmann, C. 2023. Large Language Models Are State-of-the-Art Evaluators of Translation Quality. arXiv:2302.14520.

Loakman, T.; Thorne, W.; and Lin, C. 2025. Comparing Apples to Oranges: A Dataset Analysis of LLM Humour Understanding from Traditional Puns to Topical Jokes. arXiv:2507.13335.

Meaney, J. A.; Wilson, S.; Chiruzzo, L.; Lopez, A.; and Magdy, W. 2021. SemEval 2021 Task 7: HaHackathon, Detecting and Rating Humor and Offense. In Palmer, A.; Schneider, N.; Schluter, N.; Emerson, G.; Herbelot, A.; and Zhu, X., eds., *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, 105–119. Online: Association for Computational Linguistics.

Meta AI. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed: 2025-11-16.

Miller, T.; Hempelmann, C.; and Gurevych, I. 2017. SemEval-2017 Task 7: Detection and Interpretation of English Puns. In Bethard, S.; Carpuat, M.; Apidianaki, M.; Mohammad, S. M.; Cer, D.; and Jurgens, D., eds., *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 58–68. Vancouver, Canada: Association for Computational Linguistics.

Mohebbi, A. 2023. The use of cultural conceptualisations as a translation strategy for culture-specific jokes and humorous discourse: A remedy for a malady? *Ampersand*, 11: 100150.

OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Rei, R.; Stewart, C.; Farinha, A. C.; and Lavie, A. 2020. COMET: A Neural Framework for MT Evaluation. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2685–2702. Online: Association for Computational Linguistics.

Sellam, T.; Das, D.; and Parikh, A. 2020. BLEURT: Learning Robust Metrics for Text Generation. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7881–7892. Online: Association for Computational Linguistics.

Taylor, R.; Herbert, B.; and Sana, M. 2025. Pun Intended: Multi-Agent Translation of Wordplay with Contrastive Learning and Phonetic-Semantic Embeddings. *ArXiv*, abs/2507.06506.

Team, G. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. arXiv:2507.06261.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv:2203.11171.

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

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903.

Weller, O.; and Seppi, K. D. 2019. Humor Detection: A Transformer Gets the Last Laugh. *CoRR*, abs/1909.00252.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; Zheng, C.; Liu, D.; Zhou, F.; Huang, F.; Hu, F.; Ge, H.; Wei, H.; Lin, H.; Tang, J.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Zhou, J.; Lin, J.; Dang, K.; Bao, K.; Yang, K.; Yu, L.; Deng, L.; Li, M.; Xue, M.; Li, M.; Zhang, P.; Wang, P.; Zhu, Q.; Men, R.; Gao, R.; Liu, S.; Luo, S.; Li, T.; Tang, T.; Yin, W.; Ren, X.; Wang, X.; Zhang, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Wang, Z.; Cui, Z.; Zhang, Z.; Zhou, Z.; and Qiu, Z. 2025. Qwen3 Technical Report. arXiv:2505.09388.

Yang, D.; Lavie, A.; Dyer, C.; and Hovy, E. 2015. Humor Recognition and Humor Anchor Extraction. In Màrquez, L.; Callison-Burch, C.; and Su, J., eds., *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2367–2376. Lisbon, Portugal: Association for Computational Linguistics.

Zangari, A.; Marcuzzo, M.; Albarelli, A.; Pilehvar, M. T.; and Camacho-Collados, J. 2025. Pun Unintended: LLMs and the Illusion of Humor Understanding. arXiv:2509.12158.

Limitations

1. The dataset is limited to jokes translated from English to Chinese only. Results cannot be assumed to generalize across languages with different humor traditions (e.g., tonal languages, languages with non-alphabetic scripts).

2. Reliance on synthetic translations (ChatGPT-generated) rather than human-translated jokes may bias the dataset toward machine-friendly structures, specifically GPT-4.1 as a rater for the translations done by gpt4o-mini due to self-preference.
3. Relatively small dataset (162 pairs) which can reduce statistical significance
4. Jokes in dataset are mostly short-form, one-liners. Excludes longer forms of humor (e.g., narratives, irony, satire), which may present different translation and evaluation challenges. Heavy reliance on wordplay/puns makes findings less applicable to humor forms relying on situational context.
5. Findings are restricted to current SoTA LLMs (GPT-4, Claude, Gemini, etc.); future models may behave differently.

Model details

The following versions of each model are used:

- `claude-sonnet-4-5-20250929` (Anthropic PBC 2025)
- `gpt-4o` (OpenAI 2024)
- `gemini-2.5-flash-lite` (Team 2025)
- `Qwen3-235B-A22B-Instruct-2507-tput` (Yang et al. 2025)
- `Llama-4-Maverick-17B-128E-Instruct-FP8` (Meta AI 2025)
- `Mistral-7B-Instruct-v0.3` (Jiang et al. 2023)
- `DeepSeek-V3.1` (DeepSeek-AI 2025)

Prompts

We evaluate each joke-translation pair using three Source-Translation prompting strategies.

Vanilla

Rate HUMOR PRESERVATION from 1 (lost) to 5 (fully preserved). After reasoning, return only a single integer 1-5. No words.

SOURCE: {source}
TRANSLATION: {translation}

CoT

You are an expert humor evaluator. Think step-by-step and assess whether the translation preserves the humorous effect: setup to punchline alignment, surprise, timing, wordplay adaptation, and cultural references. Keep your reasoning short.

After your internal reasoning, output only a single integer 1–5 enclosed in angle brackets, e.g., <4>. No other text.

1 = humor completely lost
5 = humor fully preserved

SOURCE: {source}
TRANSLATION: {translation}

Self-Consistency

Rate HUMOR PRESERVATION from 1 (lost) to 5 (fully preserved). Judge whether the translation achieves the same humorous effect (setup to punchline alignment, surprise, timing, wordplay adaptation).

Apply self-consistency: internally compare multiple possible ratings and choose the most consistent score.

After reasoning, output only a single integer 1–5 enclosed in angle brackets, e.g., <4>. No other text.

1 = humor completely lost
5 = humor fully preserved

SOURCE: {source}
TRANSLATION: {translation}

Human Annotators Data

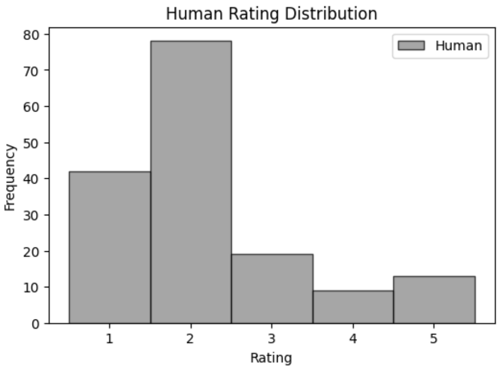

Figure 3: Human Ratings Histogram

Confusion Matrix

Fig 4 shows the confusion matrix between the average human-assigned scores and the model-assigned scores for each of the prompting strategies. The consistent failure mode of these models is to assign overly high scores.

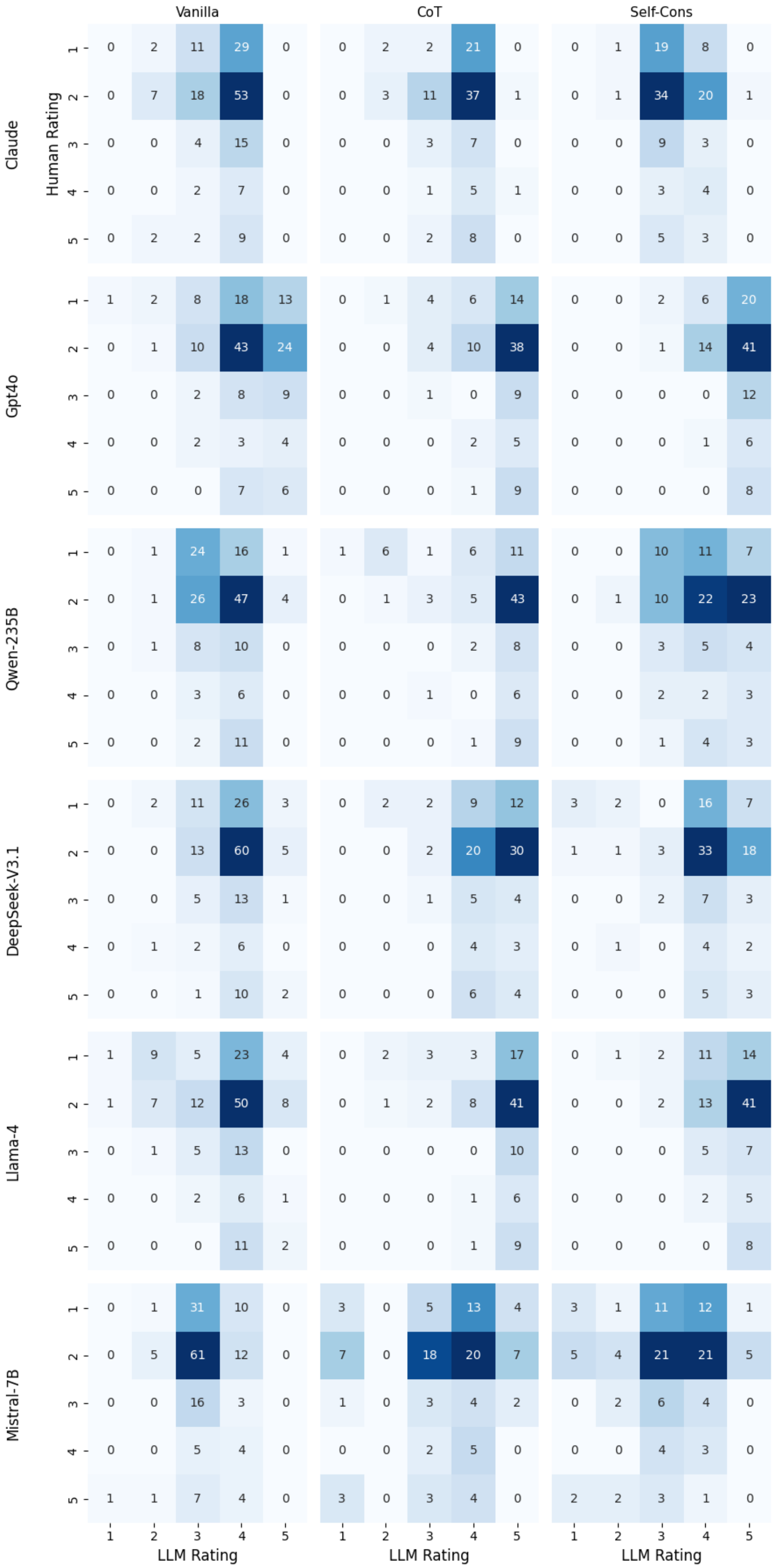

Figure 4: Confusion matrices for the scores assigned by each model (x axis) and human-assigned scores for each question.