BAHAHA at SemEval-2026 Task 1: Multi-Style Humor Generation via Comedian-Inspired Prompting and LLM Self-Assessment

Utsav Arora[1] **Michael Olaolu Arowolo**[2] **Andrew Hoblitzell**[1]
[1]Purdue University [2]Xavier University of Louisiana
{arora252, ahoblitz}@purdue.edu marowolo@xula.edu

Abstract

While humor detection and classification have received sustained attention through shared tasks and benchmark datasets, the generation of original jokes, particularly across multiple languages, is far less explored. This paper describes the BAHAHA system for SemEval-2026 Task 1: MWAHAHA, which requires generating original jokes given either a news headline or a pair of rare words. Our approach uses a generate-then-rank pipeline that combines multi-style candidate generation via comedian-inspired few-shot prompting with quality assessment from a smaller model fine-tuned on synthetic rating data produced by the generation model. Specifically, we produce up to 50 candidates per input across 15 stylistic templates and score each on humor, relevance, and novelty before surfacing a ranked shortlist for human selection. The system generates jokes natively in each target language rather than through translation, preserving culturally specific humor mechanisms such as homophonic puns and character-based wordplay. Among 305 participants and 180 submissions, our system ranks 2nd on Subtask A Chinese, and 5th on Subtasks B1 and B2. Our analysis reveals that novelty lags behind humor and relevance in automated scoring, suggesting that stylistic diversity in generation is necessary but not sufficient for humor.

1 Introduction

Humor is one of the most complex forms of human communication, requiring the simultaneous coordination of semantic incongruity, pragmatic expectation violation, and cultural common ground (Raskin, 1985; Attardo and Raskin, 1991). While natural language processing has made significant progress on humor detection and classification through shared tasks such as detecting humor in edited news headlines (Hossain et al., 2020) and rating offense in jokes (Meaney et al., 2021), the generation of original humor remains a largely open problem. Prior computational approaches to humor generation relied primarily on template-based methods and hand-crafted rules that could produce structurally recognizable jokes but lacked creativity and diversity (Amin and Burghardt, 2020). More recent neural approaches improved flexibility but still struggled to produce outputs that humans would identify as genuinely funny (Yu et al., 2018). Even with the advent of large language models, generating humor that goes beyond memorized patterns has proven difficult, as studies have shown that over 90% of LLM-generated jokes collapse into roughly 25 recurring templates (Jentzsch and Kersting, 2023).

SemEval-2026 Task 1, MWAHAHA (Models Write Automatic Humor And Humans Annotate), is the first SemEval shared task focused on computational humor *generation* rather than detection or classification (Castro et al., 2026). The task requires systems to produce original jokes under controlled constraints: given either a news headline or a pair of rare words, participants must generate humorous text that satisfies the given constraint. Systems are evaluated via pairwise human preference judgments in an arena-style setting inspired by Chatbot Arena (Chiang et al., 2024). The constrained nature of the task, particularly the word-pair subtask where both words must appear verbatim, connects to prior work on constrained humor generation such as Mittal et al. (2022), who showed that humor can emerge from carefully constructed ambiguous context rather than from training on existing jokes. The task also raises a broader question about the relationship between humor generation and hallucination: humor requires semantic deviations, including unexpected connections, intentional incongruity, and violations of pragmatic expectations, that hallucination detection systems typically flag as errors (Ji et al., 2023). MWAHAHA provides a setting in which to study *controlled* creative deviation in large language models.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 785–792

A key challenge in this space is evaluation. Generate-then-rank pipelines have shown promise for humor, with He et al. (2019) demonstrating that ranking pun candidates by a surprisal metric tripled success rates over uncurated neural generation. More broadly, hybrid systems that combine LLMs with structured humor algorithms have shown that prompted generation alone is insufficient; Toplyn (2023) demonstrated that pairing an LLM with explicit joke-writing mechanisms produced outputs judged as jokes 44% of the time, substantially outperforming a baseline where the same model was simply prompted to be funny. However, automating the ranking step introduces its own biases: LLM-based judges exhibit position bias, verbosity bias, and self-enhancement bias that can distort quality assessments (Zheng et al., 2023). Our system addresses these challenges through a three-stage pipeline. First, we generate a large pool of stylistically diverse candidates across 15 comedian-inspired templates using few-shot prompting (Brown et al., 2020). Second, a smaller model scores each candidate on humor, relevance, and novelty. Third, a human operator selects the final output from a ranked shortlist. This mixed-initiative design draws on the finding from Mirowski et al. (2023) that human-machine co-creativity works best when automated generation is paired with human judgment for final selection. We participated in all subtasks: Subtask A (text-based humor generation) for English, Spanish, and Chinese, and Subtask B (multimodal humor generation) for B1 and B2. For Subtask B, captions for input GIFs are capped at 20 words; B1 conditions on the image alone, while B2 also receives a text prompt.

2 Background

2.1 Task Description

In Subtask A, each input is either a news headline or a pair of rare words. Headlines require the joke to relate to the story. Word pairs require both words to appear verbatim. The task spans three languages: English, Spanish, and Chinese, with character limits of 900 for English and Spanish and 300 for Chinese. Evaluation is through pairwise human preference judgments (Castro et al., 2026).

2.2 Related Work

Humor generation builds on decades of linguistic theory. Raskin (1985) proposed the Script-based Semantic Theory of Humor (SSTH), arguing that humor arises when two semantic scripts overlap and oppose each other. Attardo and Raskin (1991) built on this with the General Theory of Verbal Humor (GTVH). The theory lays out six knowledge resources that underlie jokes. These are script opposition, logical mechanism, situation, target, narrative strategy, and language. Kao et al. (2016) later formalized pun humor through two information-theoretic measures, ambiguity and distinctiveness, showing that the funniest puns maximize semantic divergence between their two interpretations. Their distinctiveness measure parallels the novelty dimension in our quality assessment, where we reward candidates that create unexpected semantic contrasts rather than predictable associations.

Until recently, computational humor meant classification. Researchers worked on detecting humor in edited headlines (Hossain et al., 2020) and flagging offense (Meaney et al., 2021). Actually *generating* humor? Much less explored. Prior attempts focused mainly on puns and caption contests (Hessel et al., 2023). The JOKER shared tasks at CLEF tackled humor translation (Ermakova et al., 2023). We approach the problem differently, treating humor generation as a search over stylistic variations.

Amin and Burghardt (2020) catalogued the pre-LLM humor generation landscape: mostly templates and hand-crafted rules. Earlier, Ritchie (2005) formalized pun generation as a constraint-satisfaction problem within NLG, identifying the core challenge that output must simultaneously serve two semantic interpretations. Winters (2021) later characterized humor generation as an AI-complete problem and showed that generate-and-test architectures, where candidates are filtered through quality assessment, can match hand-tuned systems with fewer built-in assumptions, a paradigm our pipeline extends to the LLM era. Our system does not use any of these. Instead, it relies entirely on in-context learning to produce diverse candidates.

Generate-then-rank is not new for humor. Yu et al. (2018) used constrained decoding to generate pun candidates, and He et al. (2019) introduced a surprisal principle that treats surprise as both a way to generate puns and a way to rank them. Toplyn (2023) demonstrated a related hybrid approach, combining an LLM with explicit joke-writing algorithms derived from professional comedy writing, with outputs judged as jokes 44% of the time versus much lower rates for prompt-only baselines. Our

pipeline follows a similar philosophy but replaces hand-crafted humor rules with diverse prompt engineering across multiple comedic styles. Our novelty scoring works along similar lines to He et al.'s surprisal.

Why generate so many candidates? Because Jentzsch and Kersting (2023) found that over 90% of LLM jokes collapse into roughly 25 memorized templates. Their finding that ChatGPT can apply comedic templates but loses genuine comedy in the process directly motivates our high-volume candidate generation strategy: by producing 40 to 50 variants per input across diverse stylistic prompts, we force the model beyond its memorized defaults, and the assessment stage then filters out whatever still feels template-driven.

Chinese humor poses unique challenges for LLMs. He et al. (2024) showed that homophonic puns and character-based visual wordplay are especially hard. Our strongest result was in Chinese, possibly because generating natively avoids the losses that come with cross-lingual transfer on top of an already difficult task.

Without humor-aware prompting, baseline LLMs tend to handle jokes too literally and strip out the comedic impact (Pituxcoosuvarn and Murakami, 2025). This is exactly why we opted for specialized stylistic prompts. Their finding that phonetic wordplay suffers the most in cross-lingual settings also provides grounding for why our system may underperform on languages where humor relies heavily on sound-based or orthographic constraints.

We score candidates on humor, relevance, and novelty using a smaller model. This follows the LLM-as-a-Judge paradigm (Zheng et al., 2023). While such evaluation can reach over 80% agreement with humans, it also introduces known biases around position, verbosity, and self-enhancement.

Mirowski et al. (2023) had industry professionals evaluate LLM-assisted screenwriting, with humans guiding final decisions. That mixed-initiative setup is similar to ours.

3 System Description

3.1 Overview

The pipeline has three stages. First, we generate candidates across multiple comedic styles using few-shot prompting. A smaller model then scores each candidate on three quality dimensions. Finally, a human operator picks the winner from the top-ranked shortlist. The generation model is a large instruction-tuned language model accessed through a commercial API; the assessment model is a smaller model from the same family, fine-tuned on synthetic rating data produced by the generation model. The English dataset contains 275 headlines and 25 word pairs; the Spanish and Chinese datasets mirror this distribution (300 items each). Generation used few-shot prompting with curated style exemplars and default temperature settings, with a maximum of 200 output tokens per joke. Figure 1 illustrates the pipeline and Algorithm 1 provides pseudocode.

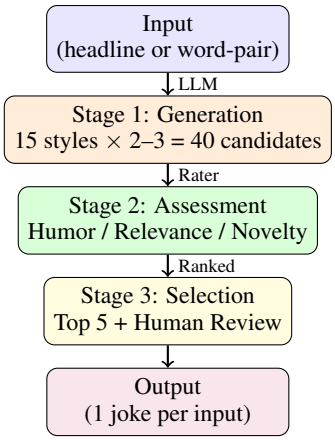

Figure 1: System pipeline. Three-stage generate-then-rank architecture with model assignments.

```
Generate ONE short joke (1–3 sentences, max 100 words)
about this news headline:
HEADLINE: {headline}
{CONTEXT: {context}}
Requirements: make it funny and original; relate directly
to the headline; keep it concise.
Output only the joke text, nothing else.
```

Figure 2: Headline-based generation prompt.

3.2 Headline-Based Generation

For headline-constrained items (∼92% of inputs), we generate 40 candidate jokes per headline distributed across 15 comedian-inspired style templates. Each template encodes a distinct comedic approach (Table 1) described via natural language style prompts rather than examples of any comedian's material.

Each joke comes from a single API call. The prompt template is shown in Figure 2; we cap output at 1 to 3 sentences.

Algorithm 1: Generate-then-Rank Pipeline

Input: Input x (headline or word-pair), styles S, k candidates per style
Output: Selected joke j^*

$C \leftarrow \emptyset$;
foreach *style* $s \in S$ **do**
  for $i \leftarrow 1$ **to** k **do**
    $j \leftarrow \text{Generate}(x, s)$; `// LLM, few-shot`
    if *satisfies_constraints*(j, x) **then**
      $C \leftarrow C \cup \{j\}$;
    end
  end
end
foreach $j \in C$ **do**
  $h, r, n \leftarrow \text{Assess}(j)$; `// Rater, 0-10`
  $j.\text{score} \leftarrow 0.5h + 0.3r + 0.2n$;
end
$C_{\text{top}} \leftarrow \text{TopK}(C, 5)$;
$j^* \leftarrow \text{HumanSelect}(C_{\text{top}})$;
return j^*

3.3 Word-Pair Generation

For word-pair constrained items (∼8% of inputs), we employ a more intensive process: 50 candidate variants per word pair, generated across 8 specialized comedy templates designed for constrained wordplay (Bureaucracy, Product Pitch, Overconfident Expert, Literal-Metaphor, Absurdist, Misdirection, Deadpan, and Status Inversion).

Both words must appear exactly as given. That is the hard constraint. After generation, we check each candidate and throw out any that fail this requirement.

3.4 Quality Assessment

All candidates undergo quality assessment using a smaller model fine-tuned on synthetic rating data generated by the larger generation model, following a knowledge distillation approach (Hinton et al., 2015). This ranking model scores each candidate on three dimensions (0 to 10 scale):

- **Humor** (w_h): How funny is the joke?
- **Relevance** (w_r): How well does it relate to the constraint?
- **Novelty** (w_n): How creative or unexpected is it?

Style	Approach
Observational	Cynical, authority-questioning
Social Commentary	Honest, storytelling
Everyday	"What's the deal with…"
Contrarian	Rants, self-aware anger
Dark/Awkward	Self-deprecating situations
Absurdist	Surreal, callbacks
One-liner	Dark humor, wordplay
Intellectual	Non-sequitur
Raw/Personal	Emotion, social observation
Character-driven	Voices, animated delivery
Stream-of-consciousness	Rapid-fire associations
Smart/Cultural	References, wit
Optimistic Absurd	Character-based positivity
Anti-humor	Deadpan, misdirection
Relationship	Social issues, energy

Table 1: The 15 comedian-inspired style templates used for headline-based generation.

Candidates are ranked by composite score $s = w_h \cdot h + w_r \cdot r + w_n \cdot n$, with $w_h > w_r > w_n$ and the weights summing to 1. Humor gets the most weight. Funniness matters most. We fine-tuned the assessment model on synthetic data from the generation model, which reduces self-preference bias and keeps the ranking criteria aligned with what the generator is actually trying to do.

3.5 Selection and Multilingual Adaptation

The selection stage uses a mixed-initiative approach in which a human operator interacts with the system through a user interface. For headlines, the interface shows the top 5 candidates. For word pairs, it shows 8. Each comes with its humor, relevance, and novelty scores. The operator reviews them in context and picks one. For word-pair items, the interface enforces a diversity constraint requiring at least 2 distinct comedy styles among finalists. This human-in-the-loop design combines the breadth of automated generation with human judgment about contextual appropriateness and humor quality. For Spanish and Chinese, the same pipeline is applied with language-specific prompts; jokes are generated natively rather than translated, preserving idiomatic humor and cultural context.

4 Results

We submitted to every subtask, producing 300 jokes each in English, Spanish, and Chinese for Subtask A and also entering B1 and B2. Evaluation used pairwise human preference judgments. Annotators compared jokes head-to-head, and the organizers computed Elo ratings (Elo, 1978) with bootstrap confidence intervals, as in Chatbot Arena

Subtask	Rank	Rating	95% CI	Teams
A-Chinese	**2**	**988**	[945, 1033]	20
B1	5	949	[921, 983]	11
B2	5	957	[911, 989]	10
A-Spanish	9	927	[894, 968]	16
A-English	23	929	[903, 960]	31

Table 2: Official results via pairwise human preference judgments (Elo ratings). Rank is computed based on how many other systems have significantly higher ratings at the 95% confidence level.

Subset	N	Humor	Relev.	Novel.
EN Headlines	275	7.8	8.1	7.4
EN Word-pairs	25	8.2	7.9	8.0
ES Headlines	275	7.5	7.8	7.2
ES Word-pairs	25	8.0	7.6	7.8

Table 3: Mean self-assessment scores (0 to 10) for selected jokes. Scores reflect the rating model's evaluation, not human judgments.

(Chiang et al., 2024). See Table 2.

Our best result was rank 2 on Subtask A Chinese with an Elo rating of 988, where only 8 systems in the rank-1 cluster achieved significantly higher scores according to the 95% confidence interval analysis. Table 3 summarizes how the candidates scored on our internal quality metrics before submission.

Word-pair jokes scored higher than headline jokes on self-assessment, which seems counterintuitive given the tighter constraints. But the word-pair pipeline had more selection pressure: 50 variants narrowed to 8, versus 40 variants narrowed to 5 for headlines. More candidates per shortlist slot means better odds of finding something good.

Across 3,167 English headline candidates, humor and relevance both peaked at 8 on our 0–10 scale (Table 4). Novelty told a different story. It peaked at 7, a full point lower, and nearly 20% of candidates scored only a 6. The model handles topicality and comedic structure well enough when given stylistic variety, but producing something genuinely surprising remains difficult. Composite scores reinforce this (Figure 3). Most candidates landed between 7.0 and 7.9. Only 59 out of 3,167 cracked 9.0.

5 Discussion

We generated over 12,000 candidates per language. A human picked the final joke from a scored shortlist. Three things stood out. First, style diversity worked. No template dominated. Absurdist humor won for trade policy; deadpan worked for sports; the pattern shifted with every headline across all 275 inputs. Second, novelty lagged behind the other two dimensions throughout the experiments, with the model reliably producing jokes that were funny and on-topic but only occasionally surprising (Table 4). Third, generating Chinese jokes natively rather than translating them contributed more to our rank-2 Chinese placement than we had expected going in.

Style diversity as exploration. A contrarian rant about a politician can be hilarious. Apply that same tone to a cooking headline and it falls flat. We had no way to predict which style would work for which topic, so we used 15 and let the scoring decide. The results justified the approach. Across 275 English headlines, no individual style accounted for more than about 10% of the top-5 selections, a distribution too flat to be explained by one or two dominant strategies.

We placed 2nd in Chinese. Part of the gap is competition size: 20 teams entered Chinese versus 31 for English. But native-language generation likely helped too, especially for a language where humor relies on phonological tricks that break in translation. He et al. (2024) showed that homo-

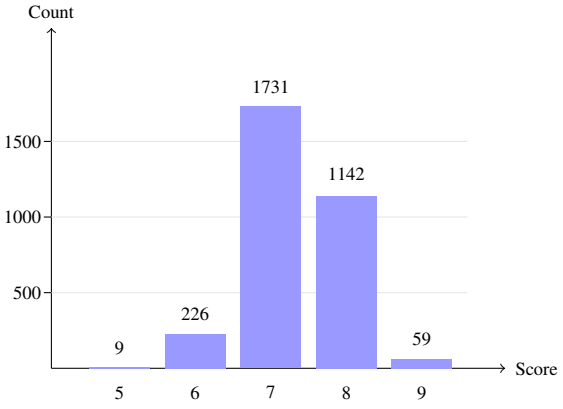

Figure 3: Distribution of composite self-assessment scores across all 3,167 rated English headline jokes.

Dim.	5	6	7	8	9	10
Humor	0.0%	1.2%	30.0%	**54.6%**	14.1%	0.0%
Relev.	0.3%	4.1%	26.0%	**45.6%**	23.3%	0.7%
Novel.	1.9%	19.6%	**37.9%**	26.5%	13.8%	0.3%

Table 4: Score distributions (%) by dimension across all 3,167 rated English jokes. Novelty skews lower (mode at 7) vs. humor and relevance (mode at 8).

phonic puns and character-based visual wordplay are fundamentally hard to transfer across languages, and Pituxcoosuvarn and Murakami (2025) found the same pattern for phonetic wordplay in English-to-Thai settings. We generated all Chinese jokes from scratch in Chinese.

Self-assessment limitations. LLMs overrate their own jokes. They also prefer familiar structures. We tried to address this by using a separate, smaller model for scoring instead of letting the generator judge itself, but that only fixes the self-enhancement part of the problem. Zheng et al. (2023) identified positional bias and verbosity bias as additional failure modes that our design does not address, and both probably still affect rankings in our pipeline. The deeper question is whether these automated scores predict what humans actually find funny, something the MWAHAHA pairwise evaluation was built to test.

Connections to hallucination research. Our system deliberately produces what hallucination detectors penalize: novel entities, unexpected connections, violations of world knowledge. A medical chatbot that invents a drug interaction has failed. A joke generator that invents an absurd scenario has succeeded. Same behavior, different evaluation.

6 Conclusion

We presented BAHAHA, a generate-then-rank system for multilingual humor generation that produces thousands of joke candidates per language across 15 comedic styles, scores them along humor, relevance, and novelty dimensions, and surfaces a ranked shortlist for human selection. Two findings generalize beyond this task. First, native-language generation substantially outperforms translate-then-adapt pipelines for humor, particularly in languages like Chinese where phonological and orthographic wordplay resists cross-lingual transfer. Second, stylistic diversity in candidate generation is more effective than optimizing a single prompting strategy, as no individual style accounted for more than 10% of top-ranked selections across 275 English headlines. Future work could replace the synthetic self-assessment model with one trained on human humor ratings collected during the MWAHAHA evaluation, and explore whether style selection can be conditioned on input features rather than relying on exhaustive generation across all templates.

Limitations

This work has several limitations. First, our pipeline relies on proprietary language models accessed through commercial APIs, which limits full reproducibility of our results. Second, the self-assessment scores produced by our ranking model may not correlate reliably with human humor judgments, as the rating model was fine-tuned on synthetic data from the generation model rather than on human annotations. Third, our 15 comedian-inspired style templates were designed primarily around Western comedic traditions and may not capture humor conventions in Chinese or Spanish-speaking cultures, potentially disadvantaging our system for non-English subtasks. This concern is consistent with broader findings that current LLMs carry systematic Western cultural biases in how they handle culturally grounded language tasks (Liu et al., 2025), which may compound the mismatch between our style templates and non-Western humor conventions. Fourth, the human-in-the-loop selection step introduces annotator subjectivity and makes the pipeline non-deterministic, meaning different operators could produce different final submissions from the same candidate pool. Finally, our analysis of score distributions and style effectiveness is based on internal metrics rather than the official pairwise human evaluations, so conclusions about which styles or dimensions drive humor quality should be interpreted cautiously. Languages where humor relies heavily on phonetic or orthographic mechanisms may pose particular challenges, as cross-lingual research has shown that sound-based wordplay suffers disproportionate losses even under optimized prompting conditions (Pituxcoosuvarn and Murakami, 2025).

Acknowledgments

We thank the SemEval-2026 Task 1 organizers for the dataset and evaluation infrastructure, and the human annotators who participated in the pairwise preference evaluation.

References

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

Salvatore Attardo and Victor Raskin. 1991. Script theory revis(it)ed: joke similarity and joke representation model. *Humor*, 4(3–4):293–348.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Santiago Castro, Luis Chiruzzo, Naihao Deng, Julie-Anne Meaney, Santiago Góngora, Ignacio Sastre, Victoria Amoroso, Guillermo Rey, Salar Rahili, Guillermo Moncecchi, Juan José Prada, Aiala Rosá, and Rada Mihalcea. 2026. SemEval-2026 task 1: MWAHAHA – models write automatic humor and humans annotate. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics. To appear.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating LLMs by human preference. *Preprint*, arXiv:2403.04132.

Arpad E. Elo. 1978. *The Rating of Chessplayers, Past and Present*. Arco Publishing, New York.

Liana Ermakova, Tristan Miller, Anne-Gwenn Bosser, Victor Manuel Palma Preciado, Grigori Sidorov, and Adam Jatowt. 2023. Overview of JOKER – CLEF-2023 track on automatic wordplay analysis. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2023*, volume 14163 of *Lecture Notes in Computer Science*, Cham. Springer.

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

Ruiqi He, Yushu He, Longju Bai, Jiarui Liu, Zhenjie Sun, Zenghao Tang, He Wang, Hanchen Xia, and Naihao Deng. 2024. Chumor 1.0: A truly funny and challenging chinese humor understanding dataset from ruo zhi ba. *arXiv preprint arXiv:2406.12754*.

Jack Hessel, Ana Marasović, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from The New Yorker caption contest. In *Proceedings of ACL*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

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

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 (WASSA)*, pages 325–340. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12).

Justine T. Kao, Roger Levy, and Noah D. Goodman. 2016. A computational model of linguistic humor in puns. *Cognitive Science*, 40(5):1270–1285.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art. *Transactions of the Association for Computational Linguistics*.

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

Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–34.

Anirudh Mittal, Yufei Tian, and Nanyun Peng. 2022. AmbiPun: Generating humorous puns with ambiguous context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Mondheera Pituxcoosuvarn and Yohei Murakami. 2025. Jokes or gibberish? humor retention in translation with neural machine translation vs. large language model. *Digital*, 5(4):49.

Victor Raskin. 1985. *Semantic Mechanisms of Humor*, volume 24 of *Synthese Language Library*. D. Reidel Publishing Company, Dordrecht.

Graeme Ritchie. 2005. Computational mechanisms for pun generation. In *Proceedings of the 10th European Workshop on Natural Language Generation (ENLG)*, pages 125–132, Aberdeen, Scotland.

Joe Toplyn. 2023. Witscript 3: A hybrid AI system for improvising jokes in a conversation. In *Proceedings of the AAAI 2023 Workshop on Creative AI Across Modalities*.

Thomas Winters. 2021. Computers learning humor is no joke. *Harvard Data Science Review*, 3(2).

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*.