ICT-NLP at SemEval-2026 Task 1: Humor Generation via RAG-based Augmentation and Multi-LLM Internal-External Voting

Wutao Shen[1,2,3], Liyuan Huang[1,2,3], Jiawei He[1,2], Lin Li[1,2], Jin Zhang[1,2]
[1]State Key Laboratory of AI Safety,
[2]Institute of Computing Technology, Chinese Academy of Sciences,
[3]University of Chinese Academy of Sciences
{shenwutao25e,huangliyuan25e,lilin2020,jinzhang}@ict.ac.cn
hepisces@qq.com

Abstract

This paper presents the system we developed for SemEval-2026 Task 1: Humor Generation. The task focuses on developing systems capable of generating genuinely humorous content under various constraints. In this work, we propose using a Retrieval-Augmented Generation approach to preprocess news headlines and obtain summaries of news content. Furthermore, we employ a unified humor generation mode to adapt to the two types of generation constraints. Finally, we conduct an internal-external voting process to produce the final optimal joke output. Our approach achieves competitive performance in this task: it ranks 1st (tied) among all participating teams in the Chinese track of Subtask A.

1 Introduction

Humor is a widely practiced creative form in artistic creation. Computer scientists and linguists have also devoted considerable effort to computational humor research. However, previous studies have often focused on humor recognition (Weller and Seppi, 2019; Kalloniatis and Adamidis, 2024) and humor understanding (Hwang et al., 2025). Humor generation remains a more challenging frontier field that is largely unexplored (Amin and Burghardt, 2020; Jentzsch and Kersting, 2023; Turgeman et al., 2025).

In this paper, we introduce the joke generation method designed for the Chinese track of Subtask A in SemEval 2026 Task 1 (Castro et al., 2026). Subtask A has two constraints. First, create a relevant joke based on the news headline. Second, generate a joke that incorporates the given keyword pair. Our method simplifies the two generation modes in Subtask A into a unified generation pipeline based on keywords and associations (Chen et al., 2023; Tikhonov and Shtykovskiy, 2024; Dubey, 2025).

The system adopts a pipeline workflow: first, we perform data augmentation on the news headline data to obtain a summary of the corresponding news content; next, we extract one keyword from both the news headline and the content summary respectively, serving as the two keywords in the keyword-based joke generation task, thereby unifying the different generation modes. We then generate associations for each of the two keywords and use multiple large language models to create candidate jokes based on multi-source semantic constraints. Finally, multiple large language models are employed to perform a two-stage internal–external voting process. For each set of candidate jokes, the LLMs involved in the generation stage first conduct a hard voting procedure as internal voting, where the joke receiving the highest number of votes is selected as the provisional optimal joke. If the vote count of this joke falls below a predefined threshold, a third-party LLM is introduced to perform a second-stage arbitration. The decision made by this external model serves as the final output. Our method ranks 1st (tied) among all participating teams in the Chinese track of Subtask A.

2 Related Work

Existing research on humor generation can be roughly divided into three paradigms: template-based, hybrid, and neural approaches (Amin and Burghardt, 2020). From the perspective of methodological evolution, the mainstream trend has gradually shifted from template-driven generation to neural generation. Early humor generation systems almost entirely relied on external resources such as templates, lexicons, and knowledge bases, constructing punchlines explicitly through lexical replacement, ambiguity, punning, and semantic reversal. It was not until 2018 that generation methods based on neural networks such as LSTMs emerged. For instance, a pun generation study was proposed by Yu et al. (2018). In recent years, large

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1965–1972

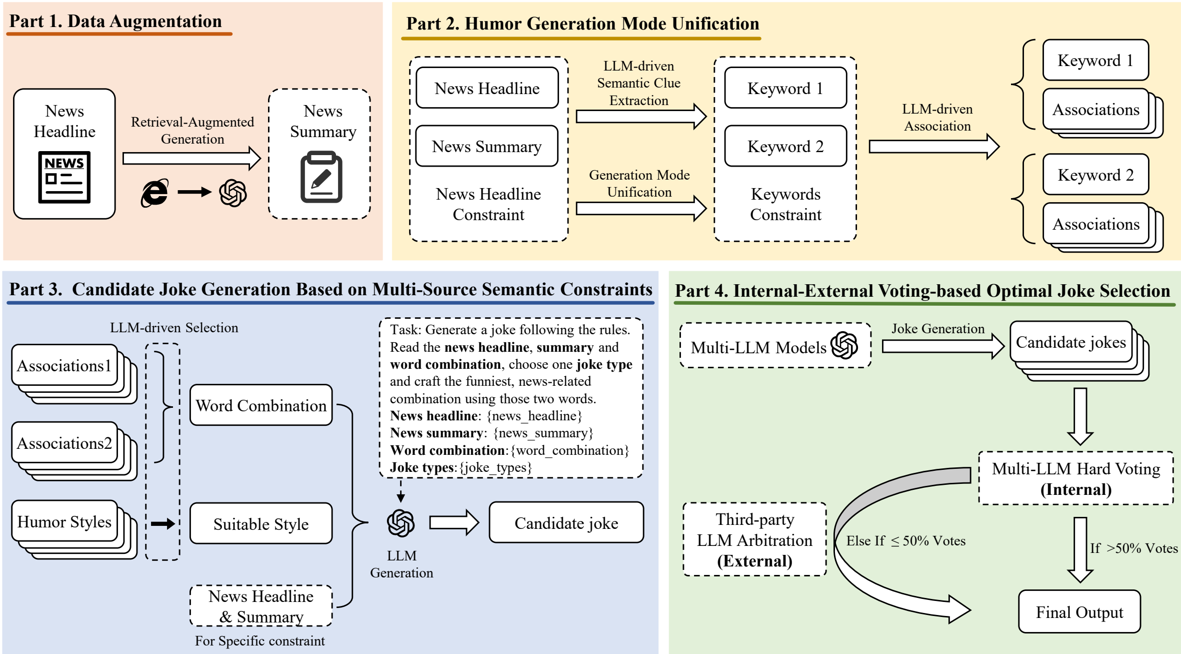

Figure 1: Humor Generation Framework. We propose a retrieval-augmented framework that expands news headlines into summaries and extracts keywords from both sources, unifying headline-based and keyword-based generation modes. An LLM performs associative expansion over keyword pairs to construct constrained word combinations, while selecting a predefined joke style. Multiple LLMs then generate candidate jokes in parallel. These candidates are anonymously voted on by peer models, as the internal voting; those exceeding a predefined threshold are selected. For low-scoring cases, an external model conducts a secondary third-party arbitration to determine the final output.

language models (LLMs) have gradually demonstrated their capabilities in artistic content creation. However, in the field of humor generation, prior research has been restricted to the collection of joke datasets (He et al., 2025; Yu et al., 2025) and the fine-tuning of models (Zhong et al., 2024; Wang et al., 2025; Evstafev, 2025), without establishing a definitive method for generating high-quality jokes using such models. Overall, although existing work has made continuous progress in fluency and controllability, there are still relatively few systems that can truly balance humor, structural complexity, and diversity.

3 System Overview

Subtask A requires us to generate jokes based on two types of provided input: either a given news headline or two given keywords. As shown in Figure 1, our system adopts a pipeline approach to generate jokes. The core technologies supporting our method include: (i) data augmentation using Retrieval-Augmented Generation (RAG); (ii) Unifying the joke generation mode to accomplish both generation tasks; (iii) Candidate joke generation based on multi-source semantic constraints; (iv) Multi-model internal-external voting-based optimal joke selection.

3.1 Data Augmentation

News headlines often contain creativity employed by journalists to attract readers, which means they are not highly condensed versions of the actual news content. Therefore, data augmentation is required for news headline data. To better generate jokes highly relevant to the news based on news headlines, we need to obtain a summary of the news content. The Retrieval-Augmented Generation (RAG) approach is suitable for this scenario: news headlines are highly time-sensitive, and corresponding news articles can be retrieved from the Internet based on the headlines, enabling us to summarize the content of these news articles.

Conventional data augmentation methods are implemented via agent application frameworks such as LangChain. To retrieve news headlines and generate content summaries, web search and content collection are first performed, and the collected web information is then fed into a large language model to obtain a summary of the news content. We adopt an equivalent and simpler approach: we use the web search functionality of the Qwen-Max model provided by the Alibaba Cloud Bailian platform to perform data augmentation in this step. With this simple implementation, we only need to provide the news headline to the Qwen-Max API to obtain a summary of the relevant news content.

3.2 Humor Generation Mode Unification

Subtask A requires us to create jokes following two modes, which can be unified into one with reference to Toplyn's joke generation theory (Toplyn, 2014, 2023; Chen et al., 2023). According to Toplyn's theory, one first summarizes the news content and extracts keywords from it as the anchors for comedy creation, then makes associations based on these anchors to construct punchlines, and finally connects the content summary and the punchline with a linking sentence to produce a joke about the news. Based on this theory, the task of creating jokes from news headlines can be transformed into a scenario of generating jokes from keywords, and the two joke creation modes are thus integrated into one. Therefore, we extract dual-source semantic cues to unify the generation mode and perform keyword association with an LLM.

Dual-Source Semantic Clue Extraction. We extract semantic clues from news headlines and summary content respectively, capturing macro-level topics and micro-level contextual details. Keywords extracted from news headlines enable the joke creation to better meet task requirements; moreover, news headlines embody the ingenuity of journalists, from which creative ideas for jokes can be drawn. In contrast, extracting keywords from news content summaries avoids the thinking limitations of taking two keywords solely from news headlines, expands the source of ideas for subsequent joke creation, and also makes the joke content more relevant to the actual news. At this point, for the subtask of generating jokes based on news headlines, after obtaining the two keywords, the subsequent generation process can follow the same mode as the task of generating jokes from two keywords. The two subtasks adopt a unified generation mode.

LLM-Driven Keyword Association. Through the previous step, we have obtained two keywords. Direct joke generation using only these keywords has limitations. We use large language models to perform further associations for the keywords, deriving three related words for each keyword. We set three associated words for each keyword because fewer words would limit diversity, while excessive words introduce redundant noise and reduce the controllability of constraints. At this point, we have a total of eight words, forming two groups with four words in each group.

3.3 Candidate Joke Generation Based on Multi-source Semantic Constraints

This task focuses on the genuine process of humor creation. Given the diversity of humor and the difficulty of joke evaluation, no annotated training data is provided. Our generation pipeline also aims to produce more diverse and humorous jokes, granting large language models sufficient room for autonomous selection, and attempting to unleash their greater capacity for independent decision-making and creation, rather than confining them to limited generation based on pre-existing templates.

LLM-driven Word Combination. Through the keyword association process in 3.2, there are 8 words available as the source of ideas for joke creation for each news item or each pair of keywords. We allow the LLM to autonomously determine the word combination of extracted keywords and associative concepts, enabling flexible semantic blending.

LLM-driven Humor Style Selection. The model is instructed to dynamically select an appropriate provided humor style as the designated humorous style (e.g., pun, exaggeration, analogy) during generation. The humorous style is referenced from existing studies (Amin and Burghardt, 2020), and the complete list of humor styles is provided in the appendix B.

Joke Generation Based on Multi-source Semantic Constraints. For the news headline task, with news headlines and summaries serving as the semantic context and thematic boundaries, one word is selected from each of the two pre-extracted keywords and their respective associated word sets as core creative materials to generate a joke consistent with the news theme and the designated humorous style. For the keyword pair task, taking two directly assigned keywords as the core theme, one word is chosen from each keyword and its corresponding associated word set as core creative materials, so as to similarly generate a joke that incorporates both keywords and conforms to the specified joke style. The prompt templates are shown in the appendix A

3.4 Multi-Model Internal-External Voting-based Optimal Joke Selection

Furthermore, we employ multiple LLMs to separately perform word combination decisions and humor type selection for joke generation, resulting in multiple candidate jokes for subsequent internal-external voting.

Each large language model participating in the anonymous voting assesses every candidate joke across seven dimensions—comprehensibility, offensiveness, joke detectability, funniness, creativity, relevance, and fluency—to identify the optimal joke as determined by that model. Anonymous voting means each evaluator model receives only the candidate jokes in a randomized order, without access to the generating model identity or any model-origin metadata. The LLMs involved in the generation stage are employed for the hard voting, as an internal voting, and vote for their favorite joke. The candidate joke with the highest number of votes is designated as the final output.

Given the limited number of large language models used for evaluation, the vote counts of the optimal jokes in some candidate joke sets may be lower than a predefined threshold, which is set to 50% in this paper. Directly taking these jokes as the final output would obscure the truly best joke in the candidate set. Therefore, we adopt a third-party model to perform secondary arbitration, as the external voting. This external model is one that has not been used for joke generation. For such candidate joke sets, the voting result from the external model is used as the final result. Through the above humor generation pipeline, we obtain jokes corresponding to news headlines or keywords.

4 Experimental Setup

Models. We use four models for joke generation, namely Qwen-Max(Yang et al., 2025), DeepSeek-V3.2(DeepSeek-AI, 2025), GLM-4.7(Team et al., 2025a), and Kimi-K2-Thinking(Team et al., 2025b). Meanwhile, these models are also used to anonymously vote on the jokes they generate. For external voting, the model employed is GPT-5.2 (Singh et al., 2025).

Implementation. All our experiments use the large model APIs provided by the Alibaba Cloud Bailian Platform[1] and the official OpenAI APIs[2]. Our code is available at `https://github.com/wutaoshen/ICT-NLP-SemEval2026-Task1`

5 Results

5.1 Subtask A: Joke in Chinese

In the test set of the verification stage, there are 275 data instances of news headline type and 25 data instances of keyword type, giving a total of 300 inputs. Following the humor generation pipeline with the four LLMs described in the experimental setup, we obtained 1200 jokes across 300 groups. These four models were used to conduct anonymous evaluation of the jokes, and the results are shown in the left subplot of Figure 2. Among them, jokes generated by Kimi-K2-Thinking and DeepSeek-V3.2 performed the best, achieving 154 and 111 best-joke selections respectively, while Qwen-Max performed the worst with only 2 best-joke selections.

Furthermore, as shown in Figure 3, 71 optimal jokes were selected as the best due to receiving 4 votes, and 120 optimal jokes were selected as the best with 3 votes. The remaining optimal jokes were subject to secondary arbitration by the third-party model, with the results shown in the right subplot of Figure 2. Kimi-K2-Thinking remained the best-performing model, being selected as optimal 163 times in total; DeepSeek-V3.2 ranked second with 98 selections; followed by GLM-4.7 with 39 selections. After arbitration, the outputs generated by Qwen-Max were never selected as the best joke.

For the jokes generated by the models, we selected 15 groups of candidate jokes from the news headline type and 5 groups from the keyword type, then invited 7 human judges to evaluate these candidates. All annotators were native Chinese speakers and evaluated the jokes independently under blind conditions with randomized presentation order. The judges were asked to select the optimal joke in their opinion and determine whether each candidate joke could be regarded as a valid joke. Due to ties in the voting results, all jokes with the same highest number of votes were regarded as optimal. The final results are shown in Figure 4: Qwen-Max was selected as optimal 5 times, DeepSeek-V3.2 5 times, GLM-4.7 6 times, and Kimi-K2-Thinking 7 times. Furthermore, only one

[1] `https://bailian.console.aliyun.com/`
[2] `https://openai.com/`

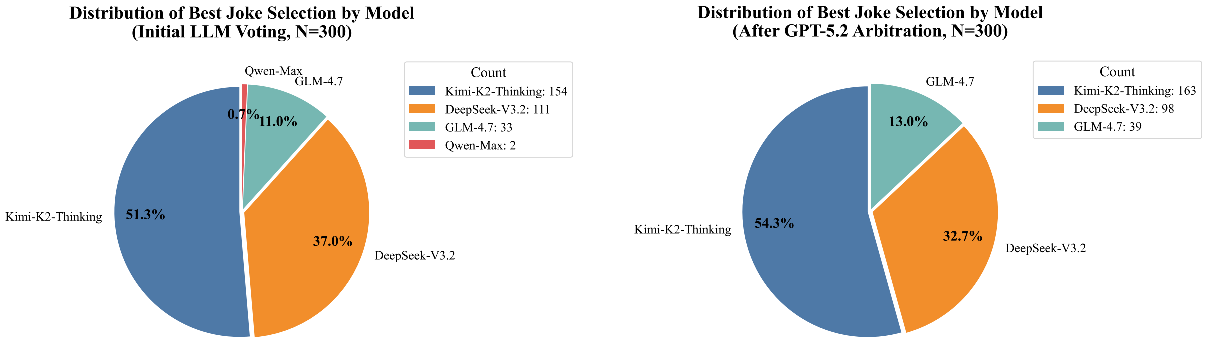

Figure 2: Comparison of the optimal joke distribution before and after secondary arbitration by the external model. The left figure shows the distribution before arbitration, and the right figure shows the distribution after arbitration.

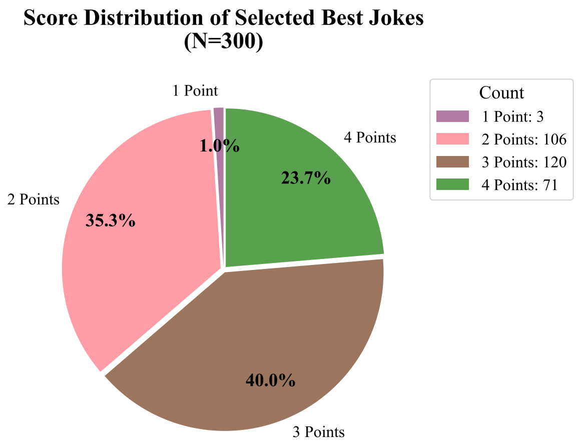

Figure 3: Score distribution of optimal jokes before arbitration: a total of 109 jokes require second-stage arbitration by the third-party model.

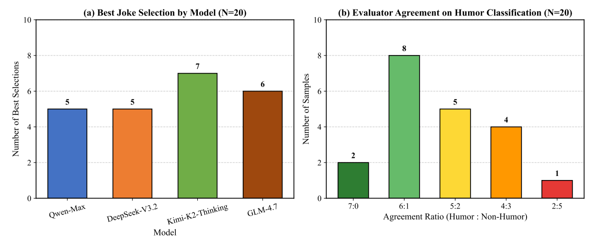

Figure 4: Result distribution obtained by selecting 20 groups of jokes from 300 candidate groups for human evaluation.The results show that the jokes generated by our method can generally be recognized as valid jokes, but also reflect the diversity of humor and inconsistent preferences among human judges.

group of candidate jokes was judged by human annotators as not constituting a valid joke: 5 judges deemed the content non-humorous, while 2 judges held the opposite view. For all other candidate joke groups, more than half of the judges agreed that the content could be recognized as a valid joke.

5.2 Competition Result

The final evaluation stage of the competition is judged based on human annotators' preferences using a pairwise comparison setup. Annotators select the funnier one from two generated texts produced under the same conditions, and an ELO-based leaderboard is used for ranking.

We only participated in the Chinese track of Subtask A, and the final joke outputs generated by our system achieved an overall score of 1052, with a 95% confidence interval of [1009, 1094]. According to the official ranking rule of the task, teams are ranked based on the dominance of their 95% confidence intervals: a team is ranked k−th if and only if there exist $k - 1$ teams whose 95% confidence intervals for scores strictly dominate that of this team. Under this ranking criterion, seven teams tied for first place, and our team was among them, meaning that no other team's results strictly outperformed ours. This is also consistent with the outcome of our aforementioned human evaluation stage, in which nearly all candidate outputs were judged as valid jokes by human annotators.

5.3 Analysis

The challenge in humor generation lies in the diversity and subjectivity of humor. The inconsistency between human judges' evaluations and LLMs' evaluations of candidate jokes indicates that it is difficult to generate jokes that satisfy everyone. The subjective nature of humor implies that it is challenging for LLMs to fully simulate human assessment. In addition, it was found in the experiments that the generation results of Qwen-Max were relatively poor, mainly because it selected the Q&A

Reversal style 127 times when selecting joke styles, resulting in excessively short jokes, which put it at a disadvantage in the later internal voting. This also suggests a future research direction: how large language models can make better choices during the creative process.

6 Conclusion

In this paper, we present the humor generation method developed for SemEval-2026 Task 1, which achieves competitive performance and ranks 1st (tied) in the Chinese track of Subtask A. By adopting data augmentation, keyword extraction and association, we design a unified humor generation mode to fulfill the generation task under two distinct constraints. We then employ a multi-model internal voting to select the optimal candidates, and finally use an external model for secondary arbitration to mitigate biases from the earlier voting. Although our method can consistently generate valid humorous jokes, the results from human judges still indicate that the diversity of humor and using LLMs to model human evaluators remains challenging. This highlights the necessity of further research and architectural improvements in future work.

7 Ethical Considerations

Since the sources of humor creation may involve content such as hate speech, stereotypes or racism, our method explicitly mandates in the prompts that LLMs must not generate such inappropriate content. Furthermore, all final outputs were manually reviewed by the authors for offensive, hateful, racist, stereotypical, or discriminatory content before submission. In the human evaluation described in Section 5.1, native Chinese speakers independently reviewed sampled candidate jokes under blind conditions.

References

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

Santiago Castro, Luis Chiruzzo, Santiago Góngora, Salar Rahili, Naihao Deng, Ignacio Sastre, Victoria Amoroso, Guillermo Rey, Aiala Rosá, Guillermo Moncecchi, J. A. Meaney, Juan José Prada, 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)*.

Yuetian Chen, Bowen Shi, and Mei Si. 2023. Prompt to gpt-3: step-by-step thinking instructions for humor generation. *arXiv preprint arXiv:2306.13195*.

DeepSeek-AI. 2025. Deepseek-v3.2: Pushing the frontier of open large language models.

Shivam Dubey. 2025. Humorplansearch: Structured planning and hucot for contextual ai humor. *arXiv preprint arXiv:2508.11429*.

Evgenii Evstafev. 2025. Optimizing humor generation in large language models: Temperature configurations and architectural trade-offs. *arXiv preprint arXiv:2504.02858*.

Ruiqi He, Yushu He, Longju Bai, Jiarui Liu, Zhenjie Sun, Zenghao Tang, He Wang, Hanchen Xia, Rada Mihalcea, and Naihao Deng. 2025. Chumor 2.0: Towards better benchmarking chinese humor understanding from (ruo zhi ba). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21799–21818.

EunJeong Hwang, Peter West, and Vered Shwartz. 2025. Bottlehumor: Self-informed humor explanation using the information bottleneck principle. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22611–22632.

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*, pages 325–340.

Antonios Kalloniatis and Panagiotis Adamidis. 2024. Computational humor recognition: a systematic literature review. *Artificial Intelligence Review*, 58(2):43.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

GLM Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, and 152 others. 2025a. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *Preprint*, arXiv:2508.06471.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025b. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Alexey Tikhonov and Pavel Shtykovskiy. 2024. Humor mechanics: Advancing humor generation with multistep reasoning. *arXiv preprint arXiv:2405.07280*.

Joe Toplyn. 2014. Comedy writing for late-night tv: How to write monologue jokes. *Desk Pieces, Sketches, Parodies, Audience Pieces, Remotes, and Other Short-Form Comedy*.

Joe Toplyn. 2023. Witscript 3: A hybrid ai system for improvising jokes in a conversation. *Preprint*, arXiv:2301.02695.

Mor Turgeman, Chen Shani, and Dafna Shahaf. 2025. One joke to rule them all? on the (im) possibility of generalizing humor. *arXiv preprint arXiv:2508.19402*.

Han Wang, Yilin Zhao, Dian Li, Xiaohan Wang, sinbadliu, Xuguang Lan, and Hui Wang. 2025. Innovative thinking, infinite humor: Humor research of large language models through structured thought leaps. In *The Thirteenth International Conference on Learning Representations*.

Orion Weller and Kevin Seppi. 2019. Humor detection: A transformer gets the last laugh. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3621–3625.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhenghan Yu, Xinyu Hu, and Xiaojun Wan. 2025. Cfunmodel: A" funny" language model capable of chinese humor generation and processing. *arXiv preprint arXiv:2503.20417*.

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

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

A Prompt Templates

A.1 Prompt for News Headline

```
You are a Chinese humor generation assistant
    with expertise in diverse classical joke
    styles.
Input
News headline: {headline}
News summary: {summary}
Understand the core message, topic, and key
    entities of the news.
1. Candidate Words
- Set A: [{keyword1}, {association1}, {
    association2}, {association3}]
- Set B: [{keyword2}, {association4}, {
    association5}, {association6}]
2. Humor Types
Wordplay / Pun, Exaggeration / Hyperbole,
    Personification / Anthropomorphism,
    Metaphorical / Analogical Humor, Malapropism
     / Intentional Misinterpretation, Two-part
    Allegorical Saying / Xiehouyu, Setup and
    Punchline / Q&A Reversal, Brain Teaser /
    Riddle, Deadpan / Anti-humor / Nonsense,
    Self-deprecating Humor, Satire / Social
    Commentary, Traditional Satire, Absurdist /
    Surreal Humor, Observational Humor, Pop
    Culture Reference / Meme Humor, Occupational
     / Professional Humor.
3. Task Objective
Select one term from each word set and choose
    one humor type to maximize humor and
    relevance. If necessary, introduce
    additional associations and justify in "
    reason".
Using the chosen humor style to generate a
    Chinese joke that is relevant to the given
    news content and contains the word pair you
    selected in the word selection step. The
    joke should reflect the topic or situation
    described in the news.
4. Constraints
- Length: 140-160 Chinese characters
- Concise, coherent, original, humorous
- No sensitive or inappropriate content
5. Output Format
{"reason": "brief explanation", "joke": "
    generated Chinese joke"}
```

A.2 Prompt for Keyword Pair

```
You are a Chinese humor generation assistant
    with expertise in diverse classical joke
    styles.
1. Candidate Words
- Set A: [{association1}, {association2}, {
    association3}]
- Set B: [{association4}, {association5}, {
    association6}]
2. Humor Types
Wordplay / Pun, Exaggeration / Hyperbole,
    Personification / Anthropomorphism,
    Metaphorical / Analogical Humor, Malapropism
     / Intentional Misinterpretation, Two-part
    Allegorical Saying / Xiehouyu, Setup and
    Punchline / Q&A Reversal, Brain Teaser /
    Riddle, Deadpan / Anti-humor / Nonsense,
    Self-deprecating Humor, Satire / Social
    Commentary, Traditional Satire, Absurdist /
    Surreal Humor, Observational Humor, Pop
    Culture Reference / Meme Humor, Occupational
     / Professional Humor.
3. Task Objective
Select one term from each word set and choose
    one humor type to maximize humor and
    relevance. If necessary, introduce
```

```
    additional associations and justify in "
    reason".
Using the chosen humor style to generate a
    Chinese joke related to the two keywords {
    keyword1} and {keyword2} and the selected
    word pair. The joke must explicitly include
    {keyword1} and {keyword2}.
4. Constraints
- Must contain {keyword1} and {keyword2}
- Length: 140-160 Chinese characters
- Concise, coherent, humorous
- No sensitive or offensive content
5. Output Format
{"reason": "brief explanation", "joke": "
    generated joke"}
```

B Humorous Style and Joke Examples

ID	Humor Style
(1)	Wordplay / Pun
(2)	Exaggeration / Hyperbole
(3)	Personification / Anthropomorphism
(4)	Metaphorical / Analogical Humor
(5)	Malapropism / Intentional Misinterpretation
(6)	Two-part Allegorical Saying / Xiehouyu
(7)	Setup and Punchline / Q&A Reversal
(8)	Brain Teaser / Riddle
(9)	Deadpan / Anti-humor / Nonsense
(10)	Self-deprecating Humor
(11)	Satire / Social Commentary
(12)	Traditional Satire
(13)	Absurdist / Surreal Humor
(14)	Observational Humor
(15)	Pop Culture Reference / Meme Humor
(16)	Occupational / Professional Humor

Table 1: Taxonomy of humor styles.