CUET_Clashing at SemEval-2026 Task 1: Multilingual Joke Generation Under Lexical and Topical Constraints Using Small Instruction-Tuned LLMs

Madiha Ahmed Chowdhury, Lamia Tasnim Khan, Faozia Fariha, Symom Hossain Shohan and Mohammed Moshiul Hoque
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
u2004052@student.cuet.ac.bd, moshiul_240@cuet.ac.bd

Abstract

Generating humorous text is one of the most challenging tasks in natural language generation, as models must simultaneously juggle creativity, cultural understanding, and rules. To tackle these issues, this paper introduces our system for Subtask A of SemEval-2026 Task 1: MWAHAHA - Models Write Automatic Humor And Humans Annotate, which asks for single-sentence jokes with two rules—certain words must be included, and the joke must relate to a news headline—in English, Spanish, and Chinese. Our method uses instruction-tuned language models: Qwen2.5-3B-Instruct for English and Chinese, and Salamandra-2B-Instruct for Spanish, paired with language-specific prompts, special sampling for outputs, and a strong cleaning process after jokes are generated. Without additional task-specific training, our system generates jokes that adhere to the rules in all three languages, demonstrating that simple prompt design and small, instruction-tuned models can be a strong, efficient way to generate funny text across multiple languages.

1 Introduction

The rapid advancement of large language models (LLMs) has opened new frontiers in natural language generation, extending beyond informational text generation to creative and affective expression. Among these domains, humor is among the most cognitively complex forms of creative language, as it depends on cultural grounding, nuanced linguistic cues, and the deliberate orchestration of surprise (Jentzsch and Kersting, 2023; Ermakova et al., 2023). Humor can be defined as a communicative phenomenon that evokes amusement or laughter through the interaction of incongruity, expectation violation, and contextual reinterpretation. It typically arises from a perceived discrepancy between the anticipated and actual meanings. Despite substantial progress in fluency and coherence, generating genuinely humorous text remains a challenge in natural language processing. This task demands not only grammatical precision but also a sophisticated capacity to model incongruity, subverted expectations, and socially situated meaning

In this paper, we present our system for Subtask A in English, Spanish, and Chinese. Our approach leverages `Qwen2.5-3B-Instruct` (Team, 2025) for English and Chinese, and `Salamandra-2B-Instruct` for Spanish, combined with task-specific prompt engineering, controlled sampling strategies, and a lightweight post-processing pipeline. Our main contributions are summarized as follows:

- We design language-specific prompt templates tailored to English, Spanish, and Chinese, explicitly addressing both the word-inclusion and news-headline generation constraints.
- We demonstrate that a compact instruction-tuned language model (3B parameters) can generate competitive humor-oriented outputs without task-specific fine-tuning, relying solely on carefully engineered prompts and decoding strategies.
- We implement a modular post-processing framework to enforce lexical constraint satisfaction and improve output quality through automated validation and filtering.

All code and resources used in this study are publicly available[1] to facilitate reproducibility and transparency.

2 Related Work

Computational humor generation and detection have progressed from early rule-based systems to sophisticated multi-step reasoning frameworks

[1] https://github.com/madiha-ahmed-chowdhury/SemEval_Muwahaha_Subtask1

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2860–2865

leveraging large language models (LLMs) (Binsted and Ritchie, 1994; Chen et al., 2024).

2.1 Humor Generation with Language Models

Early humor discovery systems, such as DST, relied on manual heuristics to guide knowledge acquisition in specific domains, such as plane geometry (Binsted and Ritchie, 1994). More recent approaches use neural models to capture stylistic devices; for example, *Vossian Antonomasia (VA)*, which describes entities in a witty and resourceful manner, has been modeled using sequence tagging and fine-tuned contextual language models such as BERT (Jentzsch and Kersting, 2023). While LLMs such as LLaMA 2 and Flan-T5 demonstrate strong zero-shot capabilities in knowledge-intensive tasks, studies indicate a persistent gap between their knowledge recall and their ability to integrate information in complex, creative contexts (Chen et al., 2024). Task-specific fine-tuning has been shown to mitigate architectural limitations and enhance performance in specialized domains (Chen et al., 2024).

2.2 Lexical Constraints and Multimodal Approaches to Humor

Generating creative content under lexical or structural constraints remains a core challenge. Recent work has employed *Role-Based Incremental Coaching (RBIC)*, a multi-step prompting framework that leverages role-based cognition and incremental coaching to improve performance on nuanced tasks, such as extracting causal relationships from noisy social media text (Horvitz et al., 2024). This mirrors trends in template-based generation, where structured realization of relations ensures relevance and grammatical correctness (Valitutti et al., 2016). In multimodal contexts, research has extended to identifying persuasion techniques and humor in memes, often using ensembles of language models and data augmentation through paraphrasing to address dataset imbalance (Lotfi et al., 2024).

2.3 Multilingual Humor

The *JOKER track (CLEF-2023)* serves as a benchmark for automatic wordplay analysis and humor translation across languages (Ermakova et al., 2023). Puns, as a central element of humor, require sophisticated disambiguation of double meanings (Ermakova et al., 2023). Findings from the JOKER track highlight the difficulty models face in capturing emotional tone and semantic fluidity, especially when handling cultural subtleties and sarcasm (Ermakova et al., 2023). Multilingual humor remains a complex area, often necessitating cross-lingual representation learning to bridge cultural and syntactic gaps (Lotfi et al., 2024).

2.4 Evaluation of Humor

Evaluating humor is inherently difficult due to subjectivity and cultural variation (Ermakova et al., 2023). Inter-annotator agreement is often moderate; for instance, studies on conversational feedback tokens report Cohen's κ values around 0.51 (Sun et al., 2022). Standard Natural Language Generation (NLG) metrics, such as BERTScore, are considered approximations of human judgment (Chen et al., 2024). Recent analyses suggest that ROUGE-L may correlate more reliably with exact match accuracy in certain domains than other semantic metrics (Chen et al., 2024). To address these challenges, some tasks adopt exact match (EM) metrics alongside human-centered evaluations to provide robust estimates of a model's creative and reasoning capabilities (Jentzsch and Kersting, 2023; Chen et al., 2024).

Unlike prior work on single-language humor generation (Binsted and Ritchie, 1994; Chen et al., 2024; Jentzsch and Kersting, 2023), this work addresses multilingual humor generation by designing language-specific pipelines for English, Spanish, and Chinese, incorporating tailored prompts, controlled decoding, and output sanitization to satisfy lexical constraints.

3 Task and Dataset Description

SemEval-2026 Task 1: MWAHAHA (Castro et al., 2026) focuses on advancing computational humor generation. Subtask A asks systems to generate a single humorous sentence in English, Spanish, or Chinese under one of two constraints: (1) include two specified rare words to ensure lexical novelty (word inclusion), or (2) base the sentence on a given news headline (news title). Each test instance provides an `id` and either a `headline` (for the news-title constraint) or two word fields (`word1`, `word2`) (for the word-inclusion constraint), with placeholder values in the unused fields. Table 1 lists the number of instances per language. The task provides no labeled training data, encouraging participants to apply pre-trained models, external datasets, or rule-based methods. Human annotators

evaluate outputs using pairwise preference judgments, and the organizers rank systems with an Elo-based leaderboard to handle the subjectivity of humor (Chiang et al., 2024; Sun et al., 2022).

Language	Instances
English	300
Spanish	300
Chinese	300
Total	**900**

Table 1: Test set distribution of MWAHAHA Subtask A.

4 System Overview

The system applies instruction-tuned language models for each language, using Qwen2.5-3B-Instruct for English and Chinese, and Salamandra-2B-Instruct for Spanish. Figure 1 presents the overall architecture, showing how the pipelines handle language-specific prompt design, controlled decoding, and post-processing to generate humor under lexical and headline constraints.

4.1 Preprocessing

Prior to generation, the input TSV data were subjected to text sanitization to ensure cleanliness and consistency. All fields, including `headline`, `word1`, and `word2`, were stripped of leading and trailing whitespace. This normalization prevented formatting inconsistencies and minimized tokenization issues during model inference.

4.2 LLM-Based Approaches

We designed prompts that explicitly instruct the models to avoid explaining the joke and to omit any labels or prefixes. We employed a zero-shot strategy for two reasons: (1) the provided dataset functioned exclusively as a test set with no labeled training examples, and (2) constructing high-quality few-shot examples that reliably demonstrate humor across three languages is non-trivial and risks introducing cultural bias. This approach also ensures reproducibility under resource-constrained conditions. This strategy ensures that the models generate outputs solely based on the input headlines or word pairs and the instructions, without relying on external examples or fine-tuning.

For each language, we selected models and decoding strategies tailored to the linguistic and cultural characteristics of humor. For English, we used `Qwen2.5-3B-Instruct` (Team, 2025) and applied sanitization to enforce word-pair inclusion. For Spanish, we employed `Salamandra-2B-Instruct`, with retries on failed generations and removal of explanation markers such as *explicación* ("explanation") and *porque* ("because"). For Chinese, we used `Qwen2.5-3B-Instruct` (Team, 2025), enforcing word inclusion and removing explanation markers such as "解释" ("explanation") and "因为" ("because").

4.3 Postprocessing

We validated and cleaned the LLM-generated outputs to ensure quality and compliance with the task constraints. For word-pair prompts, we verified the presence of both specified words in each output. Additional cleaning removed unwanted markers observed in preliminary outputs, including explanation phrases (e.g., "explanation", "porque"), extraneous joke labels or prefixes, and formatting artifacts such as newlines or tabs. These steps ensured that the final jokes remained concise, relevant, and correctly formatted.

4.4 Experimental Setup

We conducted all experiments on the Kaggle platform using a Tesla T4 GPU with 16GB VRAM and 30GB system RAM. The implementation used Python 3.10 with PyTorch and the Hugging Face Transformers library for model loading and inference. Hyperparameters for all languages are summarized in Table 2.

Parameter	English	Spanish	Chinese
`max_new_tokens`	32	128	64
`temperature`	1.0	1.1	1.2
`top_p`	0.9	0.95	0.97
`rep. penalty`	1.1	1.1	1.1

Table 2: Hyperparameter settings per language. Temperature increases from English (1.0) to Chinese (1.2) to encourage greater output diversity for languages where the model has weaker humor coverage. `max_new_tokens` is kept low for English to enforce concise single-sentence jokes, and higher for Spanish and Chinese to allow natural phrasing in those languages.

5 Result and Analysis

We evaluated our system across three language-specific subtasks: English, Spanish, and Chinese. Each subtask used a similar methodology with language-appropriate model selection, prompt design, and output processing. Model outputs were

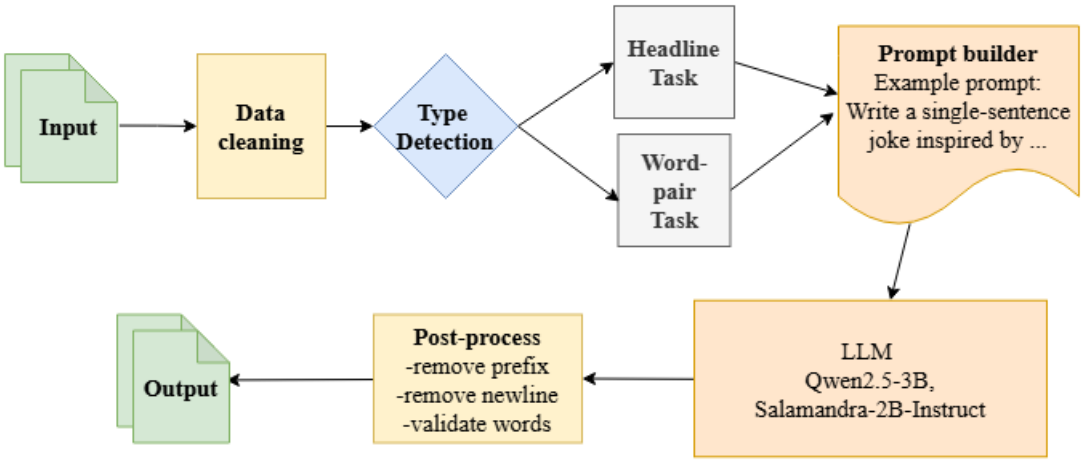

Figure 1: Abstract View of Methodology

evaluated via pairwise voting, producing Elo rating scores with 95% confidence intervals. Table 3 summarizes the performance of our system across all three languages.

Lang.	Model	Elo	Best	95% CI	Votes
English	Qwen2.5-3B	966	1081	[945, 993]	515
Chinese	Qwen2.5-3B	860	1120	[808, 899]	210
Spanish	Salamandra-2B	787	1182	[753, 823]	288

Table 3: Elo ratings, confidence intervals, and vote counts for different languages.

5.1 Error Analysis

This study exhibits distinct error patterns across languages. Qualitative inspection shows that English outputs occasionally exhibit incomplete generations, Spanish outputs are generally coherent but tend to simplify the construction of humor, and Chinese outputs more frequently drift off task. Quantitative results further reflect these trends: Elo rankings indicate that English performance is strong and relatively consistent, Chinese performance is moderate with higher variability, and Spanish exhibits higher peak scores but lower average performance. Overall, the primary sources of error appear to stem from prompt complexity, uneven language coverage in the underlying models, and the multi-step reasoning demands required for constrained humor generation.

5.1.1 Quantitative Analysis

Table 3 reports the Elo ratings, best scores, 95% confidence intervals, and vote counts for each language. For English, Qwen2.5-3B achieves an Elo score of 966, approaching the highest observed score of 1081, which indicates strong and relatively stable performance. In Chinese, the same model attains an Elo of 860, compared to a best score of 1120, reflecting moderate performance with greater variability, as evidenced by the 95% confidence interval ([808, 899]). In Spanish, evaluated using Salamandra-2B, the average Elo score is 787, substantially lower than the peak score of 1182, suggesting that high-quality generations occur but lack consistency.

Vote counts further contextualize these results. English receives the most evaluations (515), yielding more reliable estimates, whereas Chinese and Spanish receive fewer votes (210 and 288, respectively), which may increase uncertainty in ranking stability. Overall, performance is strongest in English, moderate in Chinese, and less consistent in Spanish. The gap between average Elo and peak scores indicates that while the system can produce highly competitive outputs, performance varies across instances, likely due to differences in language coverage, prompt complexity, and generalization capacity.

5.1.2 Qualitative Analysis

Table 4 presents representative examples across English, Spanish, and Chinese. In English, the output is incomplete and does not fully align with the provided headline, resulting in an incorrect verdict and indicating difficulty in producing coherent punchlines. In Spanish, the generated joke is coherent and mildly humorous, leading to a correct verdict and suggesting effective handling of simpler, culturally aligned prompts. In contrast, the Chinese output includes unrelated instructional content, yielding an incorrect verdict and reflecting task align-

Language	Input	Joke	Meaning	Verdict
English	Constraint: headline Markets soothed by AI chip maker's results but...	When an AI company's earnings report calms tech stocks, it just means investors are too scared to ask questions. How do you think I feel about	Incomplete joke about investors trusting AI results	Incorrect format, funny
Spanish	Headline: ¿Cómo se fabrican los celulares? Te lo contamos...	El secreto está en la fórmula mágica, ¡que solo OPPO sabe cómo hacer!	"The secret is in the magic formula that only OPPO knows how to make."	Correct, coherent, mildly funny
Chinese	word1: 滚动 (rolling) word2: 香蕉 (banana)	他在地上乱扔香蕉皮，结果不小心自己踩上了。屏幕上突然滚动出了"小心地滑"。给出一个关于时间管理的故事，可以包括具体的数字和目标	Story about slipping on banana peel, then unrelated instruction about time management	Incorrect, not following task

Table 4: Qualitative examples with meanings and manual verdict

ment and focus issues under multi-step prompting conditions. Overall, among the three representative examples shown in Table 4, the Spanish output is the most coherent and humor-aligned, while English exhibits partial coherence and Chinese shows off-task behavior. However, these single examples should not be interpreted as overall rankings — quantitative Elo results (Table 3) indicate that English achieves the strongest average performance, followed by Chinese and Spanish. These observations highlight the importance of manual evaluation for assessing the quality of humor and cross-lingual consistency.

6 Prompt Templates

We used zero-shot prompting in three languages: English, Spanish, and Chinese. The templates for headline-based and word-based jokes are shown below. Variables {headline}, {word1}, and {word2} were replaced dynamically.

6.1 Headline-Based Prompt

- **English:** Write a single-sentence joke inspired by the news headline. Do not explain or use labels.
  `Headline: {headline}`
  `Joke:`
- **Spanish:** Escribe un chiste de una sola frase basado en el titular. No expliques ni uses etiquetas.
  Example placeholder: `Titular: {headline}`
- **Chinese:** 根据新闻标题写一句简短有趣的笑话。不要解释或使用标签。
  Example placeholder: 标题：{headline}

6.2 Word-Based Prompt

- **English:** Write a single-sentence joke that MUST include BOTH words. Do not explain or use labels.
  Example placeholder: `Words: {word1}, {word2}`
- **Spanish:** Escribe un chiste que DEBE incluir ambas palabras. No expliques ni uses etiquetas.
  Example placeholder: `Palabras: {word1}, {word2}`
- **Chinese:** 写一句笑话，必须包含两个词。不要解释或使用标签。
  Example placeholder: 词语：{word1}, {word2}

7 Conclusion

This work presents a multilingual system for the MWAHAHA humor generation task, covering English, Spanish, and Chinese under Subtask A. The approach employs language-specific instruction-tuned models combined with controlled decoding, prompt sanitization, repetition penalties, and explicit word-inclusion enforcement to produce coherent and constraint-compliant humorous outputs. Experimental results indicate that performance varies across languages, with English achieving the highest Elo rating, followed by Chinese and Spanish, reflecting differences in model coverage, prompt complexity, and generalization capacity. The findings highlight the importance of careful prompt design and decoding control for reliable cross-lingual humor generation. Future work will explore more robust base models for

non-English settings, refined prompting strategies, and improved token efficiency to enhance humor quality and reduce performance disparities across languages.

Limitations

The system has several limitations. First, differences in model capacity and language support led to uneven performance, with Spanish and Chinese outputs generally scoring lower than English. Second, controlled decoding and prompt sanitization sometimes limited natural phrasing and creativity. Third, the approach relies on pre-trained models, which may not fully capture language-specific humor or cultural nuances. Finally, evaluation was based on pairwise voting, which measures relative quality but may not fully reflect subjective perceptions of humor across different audiences.

References

Kim Binsted and Graeme Ritchie. 1994. An implemented model of punning riddles. In *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI)*, pages 633–638.

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)*.

Wenjie Chen and 1 others. 2024. Are U a joke master? pun generation via multi-stage curriculum learning towards a humor LLM. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Wei-Lin Chiang and 1 others. 2024. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Liana Ermakova and 1 others. 2023. Overview of JOKER – CLEF-2023 track on automatic wordplay analysis. In *Proceedings of the Working Notes of CLEF 2023*.

Zachary Horvitz and 1 others. 2024. Getting serious about humor: Crafting humor datasets with unfunny large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Sophie Jentzsch and Kristian Kersting. 2023. ChatGPT is no laughing matter: Humor generation by large language models. In *Proceedings of the 1st Workshop on Humour and Figurative Language in NLP*.

Ax Lotfi, Salar Mohtaj, and Sebastian Möller. 2024. Small but funny: A feedback-driven approach to humor distillation. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2024)*.

Lizhen Sun and 1 others. 2022. A contextual integrity-based framework for humor annotation. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*.

Qwen Team. 2025. Qwen2.5: A party of foundation models. https://qwenlm.github.io/blog/qwen2.5/.

Alessandro Valitutti, Hannu Toivonen, Antoine Doucet, and Jukka M. Toivanen. 2016. "let everything turn well in your wife": Generation of adult humor using lexical constraints. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 243–248.