DUTH at SemEval-2026 Task 1: Prompt-Based Zero-Shot Large Language Models for Constrained Multilingual Humor Generation

Georgios Arampatzis **Avi Arampatzis**

Department of Electrical & Computer Engineering,
Democritus University of Thrace
{geoaramp,avi}@ee.duth.gr

Abstract

Humor generation is a challenging problem for natural language processing systems due to its subjectivity, cultural dependence, and reliance on creative language use. These challenges are further amplified in constrained multilingual settings, where models must satisfy explicit lexical or topical requirements while producing short and humorous outputs.

In this paper, we present DUTH's system for SemEval-2026 Task A on constrained multilingual joke generation in English, Spanish, and Chinese. Our approach leverages instruction-tuned large language models in a zero-shot setting, combining prompt engineering, controlled decoding, and lightweight post-generation validation to enforce constraint satisfaction and language consistency. We evaluate multiple model families and parameter scales, including Qwen and Mistral models. Human evaluation demonstrates that larger models consistently outperform smaller ones, with Qwen2.5-14B-Instruct achieving the strongest overall performance. Error analysis highlights remaining challenges such as lexical constraint violations and cross-lingual interference.

1 Introduction

Humor generation remains a challenging problem in natural language processing, as it requires not only linguistic fluency but also cultural awareness, creativity, and pragmatic reasoning. Unlike many conventional text generation tasks, humor is inherently subjective and context-dependent, making it difficult to model, evaluate, and generalize across languages and domains. Consequently, humor-related benchmarks have often served as stress tests for the expressive and reasoning capabilities of modern language models.

Recent advances in large language models (LLMs) have substantially improved performance in a wide range of generative tasks, including creative writing, dialogue generation, and controlled text production (Brown et al., 2020; Chowdhery et al., 2022). Nevertheless, producing short, coherent, and genuinely humorous content under explicit constraints—such as mandatory lexical inclusion or references to real-world events—remains non-trivial. Prior studies indicate that models frequently struggle to balance creativity with constraint satisfaction, leading either to safe but unengaging outputs or to creative but off-topic generations (Weller and Seppi, 2019; Hossain et al., 2019).

The task addressed in this work focuses on constrained multilingual joke generation, where systems must generate humorous text in multiple languages while adhering to explicit lexical or topical constraints. Such constraints reflect realistic content generation scenarios, in which models are required to incorporate specific entities, keywords, or news headlines while maintaining stylistic coherence and humorous intent. The multilingual setting further amplifies the challenge, as humor does not transfer uniformly across languages and cultures, and even minor language control errors can undermine comedic effect (Mihalcea and Strapparava, 2006, 2005).

Evaluating humor generation presents additional difficulties. Automatic metrics often correlate poorly with human judgments of funniness, originality, and contextual appropriateness (He et al., 2019). As a result, shared tasks relying on human preference judgments provide a more reliable framework for benchmarking progress in computational humor generation.

Our work builds on prior research in multilingual and creative NLP tasks, where hybrid and ensemble approaches have proven effective in handling linguistic ambiguity, cultural variation, and constrained generation settings. Previous contributions by our team have explored controlled creativity in scientific text simplification and hallucination detection, emphasizing the strengths and limitations of instruction-tuned language models under strict

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 579–584

constraints (Arampatzis and Arampatzis, 2025a). We have also investigated computational humor in multilingual settings, particularly in wordplay translation and creative adaptation tasks (Arampatzis and Arampatzis, 2025b). Moreover, our earlier work on multilingual affective and subjective language phenomena highlights the benefits of combining robust representations with disciplined modeling strategies (Arampatzis et al., 2025, 2023).

In this paper, we explore the application of instruction-tuned large language models to SemEval-2026 Task 1: MWAHAHA (Castro et al., 2026), which focuses on constrained multilingual joke generation. We evaluate multiple model families and inference configurations, analyzing the trade-offs between model scale, constraint adherence, and perceived humorous quality.

Our team (DUTH) achieved consistently strong performance in the official evaluation of Subtask A, ranking third among 31 participating systems in English, third among 16 systems in Spanish, and attaining first place among 17 systems in Chinese. These results place our approach among the top-performing submissions across all evaluated languages, demonstrating robust multilingual constraint handling and competitive humor quality.

The remainder of this paper is organized as follows. Section 2 reviews related work on computational humor and constrained text generation. Section 3 presents the dataset and modeling approach. Section 4 describes the experimental setup. Section 5 reports and analyzes the experimental results, including error analysis. Finally, Section 6 concludes the paper and discusses future directions.

2 Related Work

Early research on computational humor primarily focused on humor recognition and classification, leveraging linguistic and semantic features such as incongruity, ambiguity, and wordplay (Mihalcea and Strapparava, 2005). While these approaches provided important theoretical insights, they were limited in generating humorous content, particularly in open-ended settings.

Subsequent work explored structured humor generation tasks, including pun creation and constrained creative rewriting (He et al., 2019; Hossain et al., 2019). Although effective under controlled conditions, these methods often relied on task-specific architectures and handcrafted constraints, limiting their scalability and generalization.

The emergence of large language models has substantially advanced creative text generation. Models such as GPT-3 and PaLM demonstrate strong few-shot and instruction-following capabilities, enabling humor generation without explicit task training (Brown et al., 2020; Chowdhery et al., 2022). However, prior studies indicate that maintaining strict constraints while preserving humorous intent remains challenging, especially when lexical or contextual requirements are imposed (Hossain et al., 2019).

Multilingual humor generation remains comparatively underexplored. While multilingual models can produce fluent text across languages, ensuring humor quality and language consistency is difficult due to the cultural and linguistic specificity of humor. In contrast to previous work, this study investigates instruction-tuned large language models for constrained multilingual joke generation within a shared-task framework, systematically analyzing the impact of model scale and decoding strategies on creativity and constraint adherence.

3 System Description

Our system consists of four lightweight stages. First, each input instance is converted into a language-specific zero-shot prompt. Second, an instruction-tuned LLM generates one or more candidate jokes under a short-length constraint. Third, for lexical-constraint instances, we automatically verify whether the required words appear verbatim and trigger a minimal repair step only when necessary. Finally, when multiple candidates are available, we apply an LLM-based selection step and post-process the selected output into a single-line submission format.

3.1 Dataset

The dataset used in this work is released as part of the shared task on multilingual constrained joke generation. Each instance requires the generation of a short humorous text in one of three target languages: English, Spanish, or Chinese. Generation is subject to explicit constraints, which either enforce the verbatim inclusion of two predefined lexical items or require semantic alignment with a provided news headline. These constraints are designed to evaluate both creative flexibility and precise controllability of large language models under realistic content generation scenarios.

Languages	Samples	Input Type	Constraints	Output	Evaluation
EN, ES, ZH	300 each	Words / Headline	Inclusion, Relevance	Joke text	Human Elo ranking

Table 1: Overview of Subtask A (Text-based Humor Generation) in the MWAHAHA shared task.

The dataset is organized by language, with separate input files for each target language. Each entry includes a unique identifier, the language label, and the corresponding constraint specification. The requirement to produce concise jokes, typically limited to one or two sentences, emphasizes controlled creativity and mirrors real-world constrained text generation settings. This formulation extends prior work on creative rewriting and humor modeling by explicitly combining multilingual generation with strict lexical and semantic constraints (Mihalcea and Strapparava, 2006; Hossain et al., 2019). In addition to linguistic diversity, the dataset varies the type of constraint imposed on generation, enabling fine-grained analysis of different controllability challenges. Lexical inclusion constraints test the model's ability to preserve exact surface forms while maintaining narrative coherence and humor, whereas headline-based constraints evaluate semantic grounding and contextual creativity.

3.2 Prompt Construction, Validation, and Post-processing

Pre-processing was intentionally kept minimal in order to preserve the original structure and creative intent of the task. Input instances were lightly cleaned by removing extraneous whitespace and normalizing missing values using placeholder symbols. No token-level normalization, translation, stemming, or linguistic annotation was applied, ensuring that generation quality relied solely on model reasoning rather than handcrafted preprocessing.

For each instance, prompts explicitly specified the target language and constraint type, either requiring the inclusion of specific lexical items or referencing a given news headline. This uniform prompt structure enabled consistent behavior across languages and models while maintaining the naturalness of generated outputs.

For keyword-based constraints, we automatically checked whether both required words appeared verbatim in the generated joke. If either word was missing, we issued a single repair prompt to the same generation model, asking it to minimally revise the joke so that it included the exact required words while preserving humor and brevity. Concretely, the repair instruction followed the template: "Fix this joke so it includes EXACTLY these words: [word1] and [word2]. Keep it genuinely funny and short (1–2 sentences). Joke: [original joke]." We used this repair step only for lexical-constraint instances; headline-based instances were not subject to surface-form repair. In our implementation, the repair call used a maximum output budget of 120 tokens. This strategy follows prior findings that large language models benefit from explicit constraint verification and controllable decoding mechanisms in creative generation settings (Weller and Seppi, 2019; Hossain et al., 2019; Hokamp and Liu, 2017).

In contrast, headline-based constraints were evaluated semantically rather than through surface-form matching, allowing the model greater flexibility in creatively reinterpreting the news content. No lexical repair step was applied in this case. This distinction reflects the broader challenge of balancing form-level control with meaning-level alignment in constrained text generation (Holtzman et al., 2020).

All final outputs were post-processed into a single-line textual format by removing quotation marks, explanations, and other metadata that some models occasionally produced. This normalization ensured uniform evaluation by the shared task organizers and prevented unintended formatting artifacts from influencing human judgments.

Overall, the preprocessing pipeline prioritized simplicity and reproducibility while incorporating minimal validation mechanisms necessary to support reliable constrained multilingual generation.

3.3 Language Models

We employed instruction-tuned large language models (LLMs) in a zero-shot setting, without any task-specific fine-tuning. All generation behavior was guided exclusively through prompt engineering, explicitly specifying the target language, imposed constraints, and stylistic instructions encouraging concise and humorous outputs.

The models used in the official submission were:

- **Qwen2.5-14B-Instruct**, a large-scale decoder-only Transformer optimized for instruction following and multilingual generation.

- **Mistral-7B-Instruct-v0.3**, a compact decoder-only Transformer designed for efficient and fluent text generation.

Both models were executed locally on a single GPU using controlled decoding strategies, including low-temperature sampling, to reduce output variability while preserving creative expressiveness.

The use of instruction-tuned LLMs aligns with recent findings showing that large pretrained models can effectively perform complex creative and constrained generation tasks in zero-shot settings without specialized task-specific architectures (Brown et al., 2020; Chowdhery et al., 2022).

4 Experimental Setup

All experiments were conducted using instruction-tuned large language models in a zero-shot setting, without any task-specific fine-tuning or parameter updates. Model behavior was controlled exclusively through prompt design, decoding strategies, and lightweight post-generation validation.

For each instance, the model was prompted to generate a short joke in the specified language while satisfying the given constraint, either through mandatory lexical inclusion or semantic alignment with a news headline. To improve consistency, low-temperature sampling was applied, with temperature values between 0.8 and 1.1 depending on the model. We controlled joke length at two levels: prompt-level instructions requested exactly one short joke of 1–2 sentences, and decoding-level constraints capped generation at 120 tokens.

For instances where multiple candidate generations were produced, we applied an LLM-based reranking step to select the final output. We generated two candidates per instance before reranking. The reranker received the constraint description together with the numbered candidate jokes and was instructed to return only the index of the best candidate. The selection prompt explicitly favored originality, clarity, punchline twist, strong connection to the constraint, and conciseness, while penalizing generic, confusing, mean-spirited, or overly long outputs. The reranker selected among candidate jokes rather than among model families. In our implementation, reranking was performed with the same instruction-tuned model used for candidate generation. The reranking call used deterministic decoding (temperature 0.0) and returned only the index of the best candidate.

All experiments were executed locally on a single NVIDIA RTX A6000 GPU, using memory-efficient loading techniques such as low-bit quantization where applicable.

4.1 Evaluation Measures

Evaluating humor generation remains challenging due to the subjective nature of humor and the weak correlation between automatic metrics and human perception (He et al., 2019). Following the shared task protocol, system performance was assessed through human judgments of funniness, relevance to the imposed constraint, and overall coherence within an Elo-style ranking framework. During development, we used lightweight internal validation, including keyword verification, language consistency checks, length control, and reranking criteria aligned with the final generation setup (Hossain et al., 2019; Weller and Seppi, 2019).

5 Results

Official evaluation was conducted by the shared task organizers using human judgments, measuring humor quality, coherence, and constraint satisfaction through Elo-style comparative ranking. This evaluation framework enables fine-grained comparison between systems without relying on automatic metrics.

Table 2 reports the official results across the three target languages.

Our system achieved strong performance in all settings, ranking consistently within the top tier for English and Spanish and attaining first place for Chinese. The highest Elo-style rating was observed in the Chinese subset (1059), indicating particularly strong humorous quality and constraint adherence in this language.

In English and Spanish, the system ranked third with ratings of 1019 and 1048 respectively. Although slightly lower than the top-performing systems, the overlapping confidence intervals suggest competitive performance and stable behavior across languages.

The stronger results in Chinese may be attributed to the model's ability to leverage instruction-following behavior for concise and structured joke generation, which aligns well with the stylistic properties of short-form humor in Chinese. In contrast, English and Spanish jokes often rely more heavily on wordplay and cultural references, which can increase variability in humor perception.

Language	Rank	Rating	95% CI
English	3	1019	[984, 1045]
Spanish	3	1048	[1020, 1093]
Chinese	**1**	**1059**	[1018, 1091]

Table 2: Official human evaluation results for Subtask A across languages using Elo-style ratings. Higher ratings indicate stronger humorous quality and constraint satisfaction.

Across all languages, qualitative inspection confirmed that the model effectively balanced lexical constraint satisfaction with creative expressiveness, particularly for keyword-based prompts. Headline-based constraints remained more challenging, occasionally leading to weaker punchlines despite correct contextual grounding.

Overall, the results demonstrate that instruction-tuned large language models can achieve competitive multilingual humor generation under explicit constraints, with especially strong performance in structurally concise language settings. These results place our system among the top-performing submissions across all languages.

5.1 Error Analysis

To better understand system limitations, we conducted a qualitative analysis of representative model outputs across languages and constraint types.

A primary source of error involved violations of explicit lexical constraints, where required keywords were omitted, paraphrased, or morphologically altered despite otherwise fluent and contextually relevant generations. This issue was substantially more frequent in **Mistral-7B-Instruct-v0.3**, helping to explain its comparatively weaker performance in human rankings. The findings highlight the persistent difficulty of exact surface-form control in constrained creative generation.

Another recurring issue concerned multilingual consistency, particularly in Spanish and Chinese outputs. Some generations contained isolated English words, mixed syntactic structures, or partial code-switching, which reduced perceived coherence and humor quality. While **Qwen2.5-14B-Instruct** demonstrated stronger language separation and fewer leakage cases, complete isolation across languages was not consistently guaranteed.

In terms of humor quality, several outputs across both models were grammatically correct and satisfied the imposed constraints but lacked a clear punchline or comedic escalation. This pattern reflects a broader tendency of large language models to favor fluency and semantic safety over humorous risk-taking, resulting in conservative or predictable joke structures.

Finally, headline-based semantic constraints proved more challenging than keyword inclusion. Although models often captured the topical relevance of a news headline, this grounding occasionally came at the expense of humor strength, leading to informative but weakly humorous outputs.

Overall, the observed errors indicate that instruction-tuned large language models are effective for constrained multilingual humor generation but remain sensitive to surface-level constraint enforcement, language control, and comedic intensity. These limitations directly align with the performance gaps observed across models and highlight key directions for improving controllable creative generation systems.

6 Conclusion

This paper investigated instruction-tuned large language models for constrained multilingual joke generation in a zero-shot setting. Through prompt engineering, controlled decoding, and lightweight post-generation validation, we showed that modern LLMs can generate coherent and humorous content while satisfying explicit lexical and contextual constraints across English, Spanish, and Chinese.

Experimental results demonstrated that the submitted models, **Qwen2.5-14B-Instruct** and **Mistral-7B-Instruct-v0.3**, are capable of high-quality constrained humor generation, with the larger model consistently achieving stronger multilingual fluency, improved constraint adherence, and more engaging punchlines.

Our error analysis identified persistent challenges, including occasional constraint violations, variability in humor strength, and residual multilingual inconsistencies, highlighting the limits of purely prompt-based control.

Overall, the findings confirm the strong potential of instruction-tuned LLMs for constrained creative generation, while motivating future work on tighter constraint integration during decoding, humor-aware reranking strategies, and more robust evaluation frameworks combining human and automatic measures.

References

Georgios Arampatzis and Avi Arampatzis. 2025a. DUTH at CLEF 2025 SimpleText track: Tackling scientific text simplification and hallucination detection. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025)*, volume 4038 of *CEUR Workshop Proceedings*, pages 4211–4224. CEUR-WS.org.

Georgios Arampatzis and Avi Arampatzis. 2025b. DUTH at CLEF JOKER 2025 tasks 2 and 3: Translating puns and proper names with neural approaches. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025)*, volume 4038 of *CEUR Workshop Proceedings*, pages 2791–2802. CEUR-WS.org.

Georgios Arampatzis, Vasileios Perifanis, Symeon Symeonidis, and Avi Arampatzis. 2025. DUTH at EXIST 2025: Multilingual sexism detection with soft labels and transformers. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025)*, volume 4038 of *CEUR Workshop Proceedings*, pages 1793–1800. CEUR-WS.org.

Giorgos Arampatzis, Vasileios Perifanis, Symeon Symeonidis, and Avi Arampatzis. 2023. DUTH at SemEval-2023 task 9: An ensemble approach for Twitter intimacy analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1225–1230, Toronto, Canada. Association for Computational Linguistics.

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, 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)*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, and 48 others. 2022. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

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*, pages 1734–1744, Minneapolis, Minnesota. Association for Computational Linguistics.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Nabil Hossain, John Krumm, and Michael Gamon. 2019. "President Vows to Cut <Taxes> Hair": Dataset and analysis of creative text editing for humorous headlines. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 133–142, Minneapolis, Minnesota. Association for Computational Linguistics.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 531–538, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Rada Mihalcea and Carlo Strapparava. 2006. Learning to laugh automatically: Computational models for humor recognition. *Computational Intelligence*, 22(2):126–142.

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, Hong Kong, China. Association for Computational Linguistics.