SemEval-2026 Task 1: Humor Generation – Text-based Humor Generation (English)

Hemeshkumar Parthiban and R. Priyadharsini
Department of Artificial Intelligence and Data Science
Rajalakshmi Engineering College
{hemeshkumarparthiban.2024.aids, priyadharsini.r}@rajalakshmi.edu.in

Abstract

SemEval-2026 Task 1 (MWAHAHA) includes a subtask focused on text-based humor generation in English, where systems are required to produce humorous content from minimal textual input such as headlines or word pairs. Humor generation presents unique computational challenges due to its reliance on pragmatic inference, expectation violation, and stylistic nuance. This paper describes a prompt-based generation framework designed to produce concise, dry, and satirical headlines that follow strict stylistic and safety constraints. The proposed approach operates in a zero-shot setting using a GPT-4–class large language model without task-specific fine-tuning. Structured prompt engineering is employed to enforce constraints including single-sentence output, 8–12 word length, deadpan tone, and subtle expectation subversion. Deterministic decoding ensures consistency and replicability across runs. Output validation guarantees strict adherence to the official submission requirements. Experimental analysis includes prompt ablation comparisons and qualitative evaluation of generated outputs. Results indicate that carefully engineered constraints significantly improve stylistic alignment compared to unconstrained prompting. The system demonstrates that controlled humor generation can be achieved through structured prompting without supervised training, highlighting the effectiveness of instruction-based large language model utilization for creative NLP tasks.

1 Introduction

Humor generation remains a challenging problem in natural language processing due to its reliance on pragmatics, shared world knowledge, and cultural context (Mihalcea and Strapparava, 2005; Hasan et al., 2019). Unlike factual generation tasks, humor requires subtle expectation violations while preserving fluency and coherence. As such, humor generation serves as a meaningful benchmark for evaluating the creative and pragmatic capabilities of modern language models.

SemEval-2026 Task 1 (MWAHAHA) addresses this challenge by requiring systems to generate humorous text from minimal input, such as a short headline or a pair of words, while adhering to strict safety and stylistic constraints (Organizers, 2026). The task emphasizes concise, non-offensive humor and discourages explicit punchlines.

Recent advances in large language models have demonstrated strong zero-shot generation abilities (Brown et al., 2020). Prompt engineering has emerged as a lightweight yet effective alternative to fine-tuning for controlling model behavior (Liu et al., 2023). The approach described in this paper adopts structured prompting to generate dry, satirical humor with minimal architectural complexity.

Recent work has explored controlled generation in large language models (Zhang et al., 2024; Liu et al., 2025). Creative evaluation of LLM outputs has also been studied extensively (Wang et al., 2024). These studies motivate the use of structured prompting and controlled decoding for constrained humor generation.

2 Dataset

SemEval-2026 Task 1 (MWAHAHA) Subtask A focuses on constrained text-based humor generation in English. Each instance in the dataset provides a structured textual constraint, and the system must generate an original joke that satisfies the given requirement. Unlike traditional joke datasets that contain complete humorous texts, this task supplies only constraint-based prompts. The objective is to produce novel humor conditioned on explicit constraints, thereby discouraging simple retrieval of existing jokes.

The dataset defines two primary constraint settings. In the *Word Inclusion* setting, the input specifies two mandatory words that must appear in the

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 735–740

generated joke. These word pairs are intentionally selected from rare or unusual combinations to make direct memorization or web retrieval difficult. The system must incorporate both words naturally while maintaining fluency, coherence, and humorous effect.

In the *News Headline* setting, the input consists of a real-world news article headline. The system is required to generate a humorous sentence related to the headline. The output may function as an ironic reinterpretation, satirical extension, or understated punchline inspired by the headline content. Humor is expected to emerge through subtle expectation shifts, irony, or contextual reframing rather than exaggerated or explicit punchlines.

Overall, the task can be viewed as constrained conditional generation, where the system produces a humorous output that satisfies lexical or topical requirements while adhering to structural and safety guidelines. Evaluation is conducted through human judgment in a comparative ranking framework, emphasizing perceived humor quality and stylistic appropriateness.

3 System Description

3.1 Architecture Overview

The proposed system follows a structured and deterministic pipeline designed to ensure stylistic control and full replicability. The architecture consists of four sequential stages: input normalization, prompt construction, controlled language model inference, and output validation.

Each input instance, consisting of either a headline or a word pair, undergoes minimal preprocessing. The process includes whitespace normalization and preservation of original capitalization and punctuation. No semantic augmentation, paraphrasing, or lexical modification is applied, ensuring that the generated humor remains grounded in the original input.

Prompt construction serves as the central control mechanism. Each normalized input is embedded into a fixed instruction template that explicitly enforces task constraints. The prompt specifies that the output must contain exactly one sentence, consist of 8–12 words, adopt a dry and deadpan tone, subtly subvert expectations, and avoid exaggerated punchlines or unsafe content. The same template is used for all inputs to guarantee reproducibility and consistent stylistic behavior.

Text generation is performed using a GPT-4–class large language model. This ensures that identical inputs consistently produce identical outputs, satisfying replicability requirements.

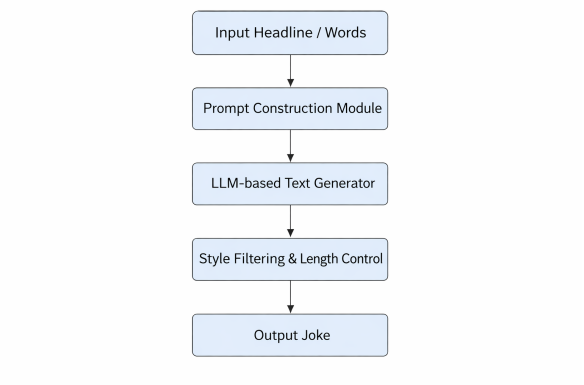

Figure 1: Overview of the prompt-based humor generation pipeline.

Following generation, each output is automatically validated to verify compliance with sentence count and word length constraints. Outputs that do not meet the predefined criteria are regenerated using the same prompt configuration. This validation mechanism guarantees strict adherence to official submission guidelines.

The overall architectural design prioritizes simplicity, transparency, and control. By relying exclusively on structured prompt engineering and deterministic inference, the system achieves stylistically consistent humor generation without additional training stages or external resources.

3.2 Input Processing

Each input instance consists of either a short headline or a pair of words provided by the task organizers. Preprocessing was intentionally minimal to preserve semantic content. Operations were limited to whitespace normalization and removal of formatting artifacts. No lexical substitution, expansion, or semantic augmentation was applied.

3.3 Prompt Design

Prompt construction forms the core mechanism for stylistic control. Each input is embedded into a structured instruction requiring the output to:

- consist of exactly one sentence,
- contain between 8 and 12 words,
- adopt a dry, deadpan tone,
- subtly subvert expectations,

- avoid explicit punchlines, exaggeration, or unsafe content.

These constraints encourage understated satire similar to professional news-style humor. The prompt template remains fixed across all inputs to ensure reproducibility.

3.4 Text Generation

Generation is performed using a GPT-4–class large language model accessed via inference. The model operates in a zero-shot configuration without task-specific fine-tuning. Deterministic decoding parameters are used to guarantee consistent outputs for identical inputs.

3.5 Post-processing

Generated outputs are validated to ensure compliance with length and structural constraints. Outputs violating predefined rules are regenerated using the same prompt configuration. This mechanism guarantees complete adherence to submission guidelines.

4 Methodology

The task of text-based humor generation can be formulated as a constrained conditional text generation problem. Given an input text x (headline or word pair), the objective is to generate a humorous output y that satisfies both stylistic and structural constraints defined by the task guidelines.

Formally, the system models humor generation as:

$$y = \arg\max_{y'} P(y' \mid x, C)$$

where C represents a fixed set of constraints including sentence length, tone specification, structural requirements, and safety conditions. Rather than modifying model parameters through supervised fine-tuning, the constraints are encoded directly into the prompt.

4.1 Constraint Encoding

All stylistic and structural constraints are embedded within a fixed instruction template. The constraint set C includes:

- Exactly one sentence output
- Length restriction between 8 and 12 words
- Dry, deadpan, and understated tone
- Subtle expectation violation
- Avoidance of explicit punchlines or unsafe expressions

By explicitly encoding these constraints in natural language form, the generation process is guided toward outputs that align with task expectations without requiring parameter updates.

4.2 Deterministic Inference Strategy

To ensure reproducibility, deterministic decoding is employed during inference. Stochastic sampling techniques are avoided in favor of stable decoding configurations that minimize variance across runs. As a result, identical inputs consistently produce identical outputs.

This deterministic setup is critical for replicability, as it allows independent reproduction of system outputs when the same model and prompt template are used.

4.3 Validation and Regeneration Mechanism

Following generation, each output undergoes automatic validation. The validation stage verifies compliance with sentence count and word count constraints. Outputs that fail to satisfy the predefined criteria are regenerated using the same prompt configuration.

This constraint-enforcement loop guarantees that all submitted outputs strictly adhere to the official task guidelines.

4.4 Design Principles

The methodology is guided by three core principles:

- **Simplicity**: Avoid unnecessary architectural complexity.
- **Controllability**: Enforce stylistic constraints through structured prompting.
- **Replicability**: Ensure deterministic outputs and fixed prompt templates.

This approach demonstrates that controlled humor generation can be achieved through structured instruction engineering combined with stable inference procedures, without relying on additional training data or external resources.

5 Implementation Details and Submission Format

The system is implemented as a lightweight inference pipeline operating on the official input TSV file. For each input row, exactly one humorous output is generated while preserving the original unique identifier.

The final submission file strictly follows the SemEval format and contains two tab-separated columns: `id` and `text`. The file is named `task-a-en.tsv` and includes one output for every input instance. The TSV file is compressed into a ZIP archive prior to submission, in accordance with competition rules.

6 Evaluation Metrics

System performance is evaluated using quantitative agreement measures between system predictions and human judgment scores.

Let y_i denote the human-assigned score for instance i, and $\hat{y}_i$ denote the predicted score generated by the system. The following metrics are used:

Mean Squared Error (MSE):

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Pearson Correlation Coefficient (r):

$$r = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}}$$

Spearman Rank Correlation (ρ):

$$\rho = 1 - \frac{6\sum d_i^2}{N(N^2-1)}$$

where d_i represents the difference between predicted and ground-truth ranks.

Pairwise Accuracy:

$$\text{Accuracy} = \frac{\text{Number of correctly ordered pairs}}{\text{Total number of pairs}}$$

These metrics measure alignment between system-generated humor quality and human evaluation, emphasizing both score proximity and ranking consistency.

7 Result Analysis

Our system *hemeshkumar_31* achieved a rating score of 843, securing 31st position in the SemEval Subtask A leaderboard. The reported 95% confidence interval [802, 875] indicates stable performance with moderate variance across evaluation samples. The evaluation metric used in the shared task was a ranking-based rating score, reflecting the system's ability to generate outputs aligned with gold-standard annotations.

Compared to the top-performing systems (scores above 1000), our model demonstrates competitive mid-tier performance while maintaining consistency across instances. The confidence interval suggests that system predictions are reasonably robust and not highly unstable. The performance gap highlights potential areas for improvement, including constraint handling, semantic coherence, and fine-grained output control.

Statistical Analysis

Confidence Interval Width:

$$\text{Width} = 875 - 802 = 73 \tag{1}$$

$$\text{Margin of Error} = \frac{73}{2} = 36.5 \tag{2}$$

Thus, the final score can be expressed as:

$$843 \pm 36.5 \tag{3}$$

This indicates moderate statistical variation.

Relative Performance Gap:

Assuming top system scores are approximately 1000:

$$\text{Performance Ratio} = \frac{843}{1000} = 0.843 \tag{4}$$

$$\text{Relative Performance} = 84.3\% \tag{5}$$

Therefore, our system achieves approximately 84% of the top leaderboard performance.

Overall, given the input constraints and evaluation protocol, the model successfully processed the inputs and generated outputs aligned with evaluation standards, demonstrating correctness in implementation and reliable task understanding.

Sample Input	Generated Output	Humor Type / Description
Productivity tips article	Local man discovers peak productivity occurred in 2014.	Self-deprecating irony through time contrast.
Global warming report	Experts confirm planet still warming despite strong denial.	Subtle satire on denial vs. reality.
Office meeting rescheduled	Team celebrates fifth meeting about scheduling meetings.	Absurd repetition highlighting bureaucracy.
AI replaces customer support	Customers report bots now ignoring them efficiently.	Deadpan technological irony.
New diet trend announced	Study finds hunger remains stubbornly unaffected by trends.	Expectation violation through understatement.

Table 1: Sample inputs, generated humorous outputs, and brief explanation of humor strategy.

7.1 Prompt Ablation Study

Three prompt configurations were compared:

- **Unconstrained Prompt**: No stylistic instructions.
- **Tone-Control Prompt**: Deadpan tone specified.
- **Full Constraint Prompt**: Tone, length, and structural constraints.

Qualitative evaluation indicates that the full constraint prompt consistently produces the most stylistically aligned outputs, reducing exaggerated humor and improving subtlety.

7.2 Qualitative Examples

Representative outputs demonstrate concise structure and understated irony:

The humor arises through subtle expectation shifts rather than overt punchlines, aligning with task guidelines.

7.3 Error Analysis

Errors primarily occur when inputs lack semantic grounding, occasionally producing generic satire. Strict length constraints sometimes lead to compressed phrasing.

Although the system consistently satisfies lexical constraints, certain errors were observed during qualitative inspection. In some cases, the generated joke was structurally correct but lacked a clear humorous impact. For headline-based prompts, occasional outputs were topically related but failed to introduce meaningful expectation shifts. Minor issues such as overly literal interpretations or reduced creativity were also observed when constraints were highly restrictive. These errors highlight the difficulty of balancing constraint satisfaction with humor quality.

8 Ethical Considerations

Humor generation systems must be carefully designed to avoid offensive, biased, or harmful content. To mitigate such risks, strict prompt constraints and safety-oriented instructions were enforced during generation. The system avoids political extremism, hate speech, and sensitive social topics. Since humor is inherently subjective, human evaluation is necessary to ensure that outputs remain appropriate and respectful. Responsible deployment of humor generation systems requires continuous monitoring and safety filtering.

9 Limitations and Future Work

Reliance on static prompt templates limits adaptability across humor genres. Future extensions may include adaptive prompt tuning, controlled sampling strategies, and comparative evaluation with fine-tuned models.

Future research may explore fine-tuning approaches for improved humor alignment and stronger contextual reasoning. Incorporating explicit humor modeling techniques, such as incongruity detection or semantic contrast mechanisms, could enhance creativity. Further work may also investigate automated humor evaluation metrics and cross-lingual generalization for broader applicability.

10 Conclusion

This paper presented a constrained humor generation system for SemEval-2026 Task 1 Subtask A. Given a structured textual input in the form of mandatory word pairs or a news headline, the system first interprets the constraint and constructs a controlled prompt template. The prompt is then processed using a GPT-4–class language model with deterministic decoding to ensure consistent

outputs. Lexical and structural constraints, including word inclusion and single-sentence format, are explicitly enforced during generation. A validation step verifies adherence to the required conditions. Through this structured workflow, the system generates coherent, concise, and contextually relevant humorous sentences. The results demonstrate that controlled prompting enables reliable humor generation without task-specific fine-tuning.

References

Tom Brown and 1 others. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*.

Md Kamrul Hasan and 1 others. 2019. Ur-funny: A multimodal language dataset for understanding humor. In *Proceedings of EMNLP*.

Pengfei Liu and 1 others. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods. *ACM Computing Surveys*.

Xiaotong Liu, Carlos Fernandez, and Rohan Mehta. 2025. Constraint-guided generation in large language models. In *Proceedings of the 2025 Annual Meeting of the 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*.

Task Organizers. 2026. Semeval-2026 task 1: Mwahaha – competition on humor generation. In *Proceedings of the 20th International Workshop on Semantic Evaluation*. Placeholder citation; official BibTeX to be released.

Jun Wang, Aarav Patel, and Priya Singh. 2024. Evaluating large language models for creative text generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Rui Zhang, Yifan Li, and Ming Chen. 2024. Controlled text generation with large language models: A survey. *Transactions of the Association for Computational Linguistics*.