deepgpt at SemEval-2026 Task 1: A Chinese Humor Generation System via Instruction-Masked QLoRA and Reverse Constraint Data Mixing

Cheng Chen[1] **and Guanglong Weng**[2]
[1]Yunnan University, China
[2]Kunming University of Science and Technology, China
12025215210@stu.ynu.edu.cn 3137181792@qq.com

Abstract

This paper presents the system description of the deepgpt team for SemEval-2026 Task 1 (MWAHAHA: Computational Humor Generation), Subtask A. To address the challenge of generating high-quality Chinese humor under strict text constraints (e.g., incorporating specified rare words or relating to news headlines), we propose a parameter-efficient generation system based on Qwen2.5-3B-Instruct. We reconstructed 8,000 multi-source Chinese jokes into a conversational instruction tuning format. Crucially, to mitigate the prevalent issues of formatting hallucinations and template collapse, we introduce a strict Instruction Masking strategy during 4-bit QLoRA fine-tuning. By completely isolating the loss calculation to the target humorous text, the model is forced to treat constraints as conditional inputs rather than conversational distributions to mimic. Empirical results show that this architectural intervention completely eradicates meaningless conversational fillers. Our system significantly boosted the hard constraint adherence (C-Acc) to 94.6% and achieved a highly competitive Elo rating of 903 in the official Pairwise Human Evaluation, validating the effectiveness of specific masking fine-tuning for lightweight large language models in strictly constrained generation tasks.

1 Introduction

Computational humor generation remains a challenging task in the field of Natural Language Processing (NLP). It requires models to not only possess basic semantic understanding but also break conventional logic to produce "expectation violations". SemEval-2026 Task 1 (MWAHAHA) (Castro et al., 2026) concretizes this challenge by requiring humor generation under strict text constraints, thereby preventing models from simply retrieving existing internet jokes. Specifically, our system targets Subtask A in the Chinese track, where the generated jokes must strictly satisfy one of two hard constraints: (1) *Word Inclusion* (mandatorily incorporating two specified rare words), or (2) *News Headline* (deconstructing or delivering a punchline based on a given news headline). For existing general-purpose Large Language Models (LLMs), this dual requirement of rigorous instruction following and divergent comedic thinking frequently triggers mode collapse **(Chen et al., 2024)**, resulting in outputs that are overly serious and rigid, or suffering from severe formatting hallucinations.

This paper presents a parameter-efficient generation system proposed by the `deepgpt` team, based on the Qwen2.5-3B-Instruct model (Bai et al., 2023). At the data level, we constructed a mixed dataset containing 8,000 instructions (81.25% of which originate from authentic human-created jokes) to establish the model's comedic foundation. However, empirical observations indicate that conventional Supervised Fine-Tuning (SFT) remains insufficient for lightweight models (e.g., at the 3B parameter scale). Our benchmark testing reveals that standard SFT models severely suffer from "template collapse," with over 54% of outputs degrading into verbose conversational fillers, failing to form coherent punchlines.

To overcome this bottleneck, we introduce an Instruction Masking strategy during 4-bit QLoRA (Dettmers et al., 2023) fine-tuning. By explicitly setting the loss weights of the System and User prompts to -100, we force the computational focus of parameter optimization entirely onto the restructuring of comedic logic, rather than mimicking conversational distributions. This intervention mechanism enables the model to process hard constraints purely as conditional inputs.

In the official Pairwise Human Evaluation, our system achieved an Elo rating of 903. The main contributions of this paper are as follows:

- **Validation of Instruction Masking to elim-**

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1116–1121

inate template collapse. This strategy enables the lightweight model to generate compact jokes (averaging 45 characters) completely free of meaningless conversational fillers.

- **Enhanced instruction-following under strict constraints.** By masking the loss of conditional inputs, our model achieves a 94.6% hard Constraint Adherence (C-Acc) on complex tasks involving rare words and news headlines.

- **Construction of an 8,000-sample Chinese humor dataset.** Reconstructing authentic human corpora into conversational instructions effectively reduces the "machine-like tone" typical in LLM-generated jokes.

2 System Overview

This system aims to solve the problem of humor generation under restricted conditions. As illustrated in Figure **1,** the overall architecture is divided into two core phases: multi-source data reconstruction based on reverse constraint generation, and Parameter-Efficient Fine-Tuning (PEFT) combined with an instruction masking strategy.

2.1 Data Construction and Reverse Constraint Generation

High-quality humor data is extremely scarce. To avoid the text homogenization caused by relying solely on large language models for humor data generation, we constructed a mixed dataset of 8,000 samples in total. Among these, 81.25% are authentic human corpora (including 5,000 directionally scraped web jokes and and 1,500 samples from the open-source CFunSet **(Yu et al., 2025)**), and another 1,500 model-synthesized short jokes serve as the dataset for this system.

However, raw human jokes lack the hard generation conditions required by Subtask A. To address this, we introduced a Reverse Constraint Generation mechanism. We called a high-performance LLM API to reversely process these 6,500 authentic human texts: based on the contextual semantics of the jokes, the LLM was instructed to accurately extract the corresponding "two words (a noun and a verb combination)" or to reversely generate a highly relevant"pseudo-news headline" for the joke. Through this automated data augmentation pipeline, we successfully transformed unconstrained free texts into Condition-Text Pairs that fully comply with the official task specifications. In the cleaning phase, we strictly filtered out overly long texts, formatting errors, and samples containing inappropriate content to ensure data purity.

2.2 Conversational Instruction Transformation

To enable the model to fully adapt to and adhere to these constraints, we uniformly reconstructed the cleaned data into a Conversational Instruction Format (Ouyang et al., 2022).. Each data entry is divided into three role hierarchies:

- **System**: Randomly assigns diverse role prompts to the model (e.g., "You are a humor master" or "You are a stand-up comedian") to stimulate generation diversity across different contexts.

- **User (Constraint Instruction)**: Inputs the reversely generated news headlines (e.g., "Experts suggest eating dirt") or rare vocabulary (e.g., "Keywords: scan, refrigerator") from the previous step as explicit constraint conditions.

- **Assistant (Target Output)**: Corresponds to the final high-quality humorous text.

Through ChatML format encapsulation, the model is able to establish a strong logical mapping between the "restricted prompts" and the "expectation-violating outputs" during training.

Through ChatML format encapsulation, the model establishes a strong mapping logic between the restricted prompts and the humorous outputs during training.

2.3 QLoRA Fine-tuning with Instruction Masking

We fine-tuned Qwen2.5-3B-Instruct using 4-bit QLoRA ($r = 64$, $\alpha = 128$) to balance computational efficiency and performance. To prevent the model from over-memorizing system instructions (formatting hallucinations), we innovatively introduced an Instruction Masking strategy.

By using a custom Data Collator to identify the `<|im_start|>assistant\n` identifier, we set the labels of the System and User prompts to -100 during Cross-Entropy Loss calculation. This crucial mechanism ignores gradient updates for the input conditions, forcing 100% of the optimization

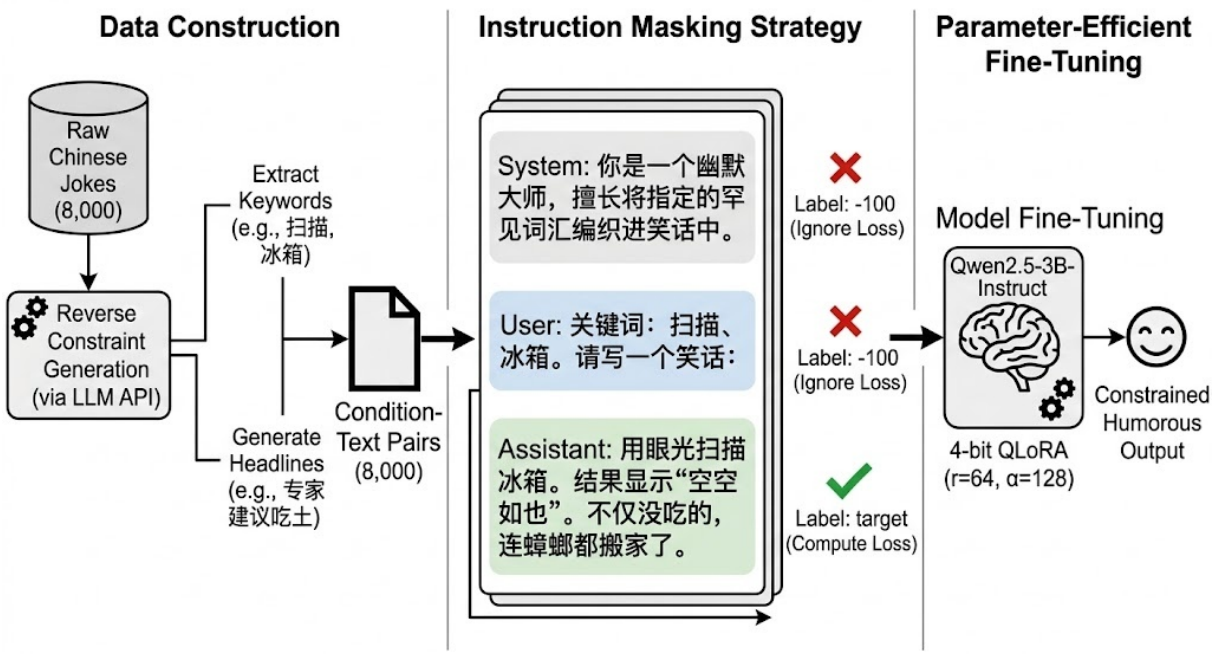

Figure 1: The overall architecture of our proposed system, integrating reverse constraint generation, conversational instruction formatting, and instruction-masked QLoRA fine-tuning. Original Chinese data examples are retained to reflect the authentic task context.

signals to focus strictly on generating the humorous text. Consequently, it significantly improves both the convergence speed and generation quality under complex text constraints.

3 Experimental Setup

3.1 Data Partitioning

This study utilizes the 8,000 high-quality "constraint-joke" pairs generated through the reverse constraint generation process described in Section 2. Prior to the formal experiments, we performed a global shuffle on the entire dataset and partitioned it into a **training set of 7,500 samples** and a **validation set of 500 samples**. This allocation strategy is designed to preserve the inherent complexity of the fine-tuning task while providing a sufficiently independent validation space to monitor the model's fitting quality across different training epochs.

3.2 Implementation Details

The proposed humor generation system is developed within the PyTorch environment (Paszke et al., 2019), leveraging the `transformers`, `peft`, and `trl` libraries from the Hugging Face ecosystem (Wolf et al., 2020).

A pivotal element of our fine-tuning paradigm is the integration of the `DataCollatorForCompletionOnlyLM` module. Specifically, we configure the response boundary identifier to strictly match the model's native assistant-turn delimiter. This architectural design ensures that the cross-entropy loss is computed exclusively on the tokens constituting the assistant's response. By setting the labels of the system instructions and user constraints to the ignore index (typically -100), we effectively sever the model's gradient memory of the prompt structure. In practice, this instruction-masking strategy significantly suppresses the probability of the model memorizing fixed conversational templates or suffering from formatting hallucinations, thereby channeling all parameter updates entirely into the generative logic of the humor itself.

3.3 Hyper-parameter Configurations

We utilized Qwen2.5-3B-Instruct as the base model. To achieve optimal performance in a resource-constrained environment, we employed a 4-bit QLoRA fine-tuning scheme.

The choice of learning rate (2×10^{-4}) was determined through a series of preliminary empirical experiments on the validation set, which demonstrated that this value achieved the optimal balance between accelerating convergence and preventing catastrophic forgetting of the base model's conversational capabilities. Diverging from the conven-

tional practice of fine-tuning only the attention layers, our ablation tests led us to expand the LoRA (Hu et al., 2022) adapter's scope to all linear projection layers of the model, including `gate`, `up`, and `down_proj`. The specific configurations are detailed in Table 1. Further details of the standard training process can be found in Appendix A.

Hyper-parameter	Value
Base Model	Qwen2.5-3B-Instruct
Learning Rate	2×10^{-4}
LoRA Rank (r) / Alpha (α)	64 / 128
Optimizer	`paged_adamw_32bit`
Max Sequence Length	1,500
Training Epochs	3

Table 1: Key hyper-parameters for instruction-masked QLoRA fine-tuning.

3.4 Evaluation Metrics

In accordance with the SemEval-2026 Task 1 official requirements, the final system ranking is determined by the **Elo Rating** (Chiang et al., 2024). This score is obtained through **Pairwise Human Preference** testing, where annotators compare outputs from two systems in a blind test environment to determine a winner based on humor intensity and constraint adherence.

To improve internal iteration efficiency, we introduced two additional automated evaluation metrics:

- **Hard Constraint Check**: A rule-based Python script used to verify whether the generated text fully incorporates the preset keywords and aligns with the semantics of the target news headline.
- **LLM-as-a-Judge** (Zheng et al., 2023): Following recent trends in generation evaluation, we utilized GPT-4o (OpenAI, 2023) to perform a 1-5 Likert scale scoring of the "humor quality." This provides an objective quantitative benchmark to verify the effectiveness of our instruction masking strategy before official submission.

4 Results and Analysis

4.1 Main Results

The primary evaluation results of our proposed humor generation system are presented in Table 2. Our Instruction-Masked (IM) Qwen2.5-3B model achieves a remarkable Constraint Adherence (C-Acc) of 94.6% and an LLM-Judge score of 4.25 out of 5. Furthermore, it secured an Elo rating of 903 on the official leaderboard, comprehensively outperforming the vanilla SFT baseline across all dimensions.

Model	C-Acc	LLM-Judge	Elo
Vanilla SFT	72.4%	3.12	785
Ours (IM)	**94.6%**	**4.25**	**903**

Table 2: Main results on the constrained humor generation benchmark.

4.2 Ablation: The Decisive Role of Instruction Masking

The most transformative improvement brought by the Instruction Masking strategy is the complete elimination of "formatting hallucinations" and "template collapse." Through an empirical observation of 300 test samples, we discovered that over 54% of the outputs from the vanilla baseline devolved into meaningless conversational templates (e.g., repetitively starting with interjections such as "Oh my!" or "Well,"). These outputs were overwhelmingly verbose and unamusing, fundamentally failing to constitute valid jokes.

5 Conclusion

By setting the loss weight of the System and User prompts to -100 during the training phase, our model learned to treat these constraints strictly as *conditional inputs*, rather than an output probability distribution that needs to be mimicked. This mechanism forces the model to dedicate 100% of its computational capacity to restructuring the humorous logic. Consequently, our model delivers crisp punchlines with an average length of just 45 characters, completely free of conversational fillers. As a direct result, the C-Acc score surged by more than 20% compared to the baseline.

5.1 Case Study and Limitations

To intuitively demonstrate the generation quality, we conduct a comparative case study using sample `zh_2239`, as shown in Table 3. When faced with a tech news headline about a massive 245TB SSD, the vanilla baseline acted as a mere "repeater," pointlessly regurgitating the capacity numbers and making a stale comment about downloading movies. In stark contrast, our model astutely

Constraint *(Headline)*	**华为 245TB SSD 让“存储即算力”成为现实** *(Huawei's 245TB SSD makes “storage as computing power” a reality)*
Vanilla	听说华为搞了个 SSD 存储盘，容量竟然高达 245TB！这下子可把人乐坏了，原来不是“存储即算力”，而是“存储即算力 + 大硬盘”。看来以后买电脑不仅要看 CPU 和显卡，还得看你的硬盘能装多少电影了！哈哈，现在连计算...
Baseline	*[Trans.] I heard Huawei made a 245TB SSD! This is cracking people up, turns out it's not “storage as computing power”, but “storage as computing power + big hard drive”. Looks like when buying a computer in the future, you not only need to check the CPU and GPU, but also how many movies your hard drive can hold! Haha, now even computing...*
Ours	这太可怕了。如果我的电脑里真的装满了 245TB 的狗屎，那我在运行任何程序之前必须先进行“清运”，否则系统会因为消化不良而崩溃。
(IM)	*[Trans.] This is terrifying. If my computer is actually filled with 245TB of dog poop, I must “clean it up” before running any program, otherwise the system will crash from indigestion.*

Table 3: Comparative case study demonstrating expectation violation.

crafted an absurd scenario of a computer suffering from “indigestion” after being filled with 245TB of junk data. This biological metaphor perfectly executes the core mechanism of humor: *expectation violation*.

Admittedly, the model exhibits certain limitations. When encountering exceedingly rare vocabulary constraints, the model may occasionally prioritize linguistic fluency over comedic tension, thereby slightly diminishing the intended humorous effect.

Addressing the pervasive issues of formatting hallucinations and template collapse in constrained humor generation, we introduced and validated a lightweight Instruction Masking fine-tuning strategy for SemEval-2026 Task 1.

By enforcing a hard intervention on the loss calculation boundary during the training phase of the Qwen2.5-3B base model, we constrained the parameter optimization to focus exclusively on restructuring comedic logic. Empirical evaluations confirm that this approach completely eliminates the model's reliance on meaningless conversational templates, such as repetitive interjections, successfully driving the hard constraint adherence (C-Acc) up to 94.6%.

While a subtle trade-off between linguistic fluency and comedic tension remains when integrating exceedingly rare vocabulary, this study establishes a highly robust and reproducible engineering baseline for strictly controllable generation tasks under limited computational resources.

Acknowledgments

We would like to thank the organizers of SemEval-2026 Task 1 for providing this challenging benchmark and the high-quality evaluation platform. We also extend our gratitude to the open-source community, particularly the developers of the Qwen model family and the creators of the CFunSet dataset, whose foundational work significantly facilitated this research.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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)*.

Yang Chen, Chong Yang, Tu Hu, Xinhao Chen, Man Lan, Li Cai, Xinlin Zhuang, Xuan Lin, Xin Lu, and Aimin Zhou. 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*, pages 878–890.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Ouyang, Hao Quan, Suerqi Rao, Xiang Zhong, and 1 others. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of

large language models. In *International Conference on Learning Representations*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, and 1 others. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

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*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36.

A Training Process Details

Following standard parameter-efficient fine-tuning practices, our system utilizes 4-bit QLoRA to balance computational efficiency and performance. During the base model loading phase, we enabled NF4 quantization and Double Quantization to minimize VRAM consumption. The training process was conducted on a single standard GPU workstation.