Team hugang11 at SemEval-2026 Task 1: A CoT-SFT, Teacher-Constructed DPO, and Deterministic Post-processing Pipeline for Chinese Humor Generation

Gang Hu, Liu Yang and Jing Li
School of Information Science and Engineering,
Yunnan University, Kunming 650091, China
{12025215204, 12025215185, lijing_fyjo}@stu.ynu.edu.cn

Abstract

We describe the **hugang11** system for **SemEval-2026 Task 1 (MWAHAHA)**, participating in **Subtask A (Chinese)** (Castro et al., 2026). Our system addresses a practical trade-off in creative text generation: models that produce sharper and more stylized jokes often become less stable in output format. We build a three-stage pipeline that combines chain-of-thought-augmented supervised fine-tuning (CoT-SFT), teacher-constructed direct preference optimization (DPO), and deterministic post-processing. Built on `unsloth/Qwen2.5-7B-Instruct-bnb-4bit`, our final submission used the **CoT-DPO** configuration and achieved a live leaderboard rating of 991 with a 95% confidence interval of [958, 1036] based on 209 votes, placing it in the second rank group. Our results suggest that for constrained humor generation, alignment-oriented training improves style, while robust inference-time control remains essential to prevent rationale leakage and malformed outputs.

1 Introduction

SemEval-2026 Task 1 (MWAHAHA) focuses on automatic humor generation from short prompts across multiple languages (Castro et al., 2026). We participated in Subtask A for Chinese, where the system must generate a humorous response under lexical or topical constraints. Humor generation is difficult because fluent text alone is not enough: successful jokes often depend on a setup–punchline structure and on an incongruity that violates an expected continuation (Xie et al., 2021; Amin and Burghardt, 2020). Recent surveys further emphasize that humor generation remains underexplored beyond puns, and that modern language models still lag behind human authors on context-dependent humor (Loakman et al., 2025).

This challenge is particularly relevant for Chinese. Prior work on Chinese comical crosstalk shows that large language models substantially improve humor generation quality, but the gap to human-written humor remains large (Li et al., 2023). Recent preference-learning work on Chinese and English pun generation also suggests that humor quality benefits from multi-stage supervision rather than plain next-token training alone (Chen et al., 2024). These findings motivate a pipeline that separates identifying a humorous angle from realizing it in a competition-compliant final answer.

Our strategy therefore divides the task into three stages: reasoning, stylistic alignment, and output control. First, we apply CoT-augmented supervised fine-tuning (CoT-SFT), encouraging the model to analyze the latent absurdity or contrast in the prompt before producing the joke. Second, we use teacher-constructed DPO so that the model prefers sharper and more scene-based responses over weaker baseline outputs. Third, because explicit reasoning can leak into the final answer and break the required TSV format, we adopt constrained decoding and deterministic post-processing to extract and sanitize the final punchline.

Our main finding is that better humor style does not automatically imply better submission robustness. In our experiments, CoT-style supervision and DPO made outputs more structured and stylistically sharper, but also increased the risk of rationale leakage, repetition artifacts, and malformed text. On the official live leaderboard, our final submission **hugang11** achieved a rating of 991 with a 95% confidence interval of [958, 1036] based on 209 votes, placing it in the second rank group. Our code is publicly available.[1]

[1] https://github.com/hugang1114-debug/hugang11-SemEval-2026-Task-1

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1724–1729

2 Background and Related Work

Computational humor generation has evolved from templates and rule-based systems to neural and large-language-model approaches (Amin and Burghardt, 2020). Across this literature, a recurring view is that many jokes can be understood as a setup that builds an expectation and a punchline that violates or reframes that expectation (Xie et al., 2021). Recent overviews of computational humour generation and explanation likewise argue that humor is creative, context-dependent, and still challenging for contemporary LLMs (Loakman et al., 2025). For Chinese, the crosstalk benchmark of Li et al. (2023) and the Chinese pun work of Chen et al. (2024) show both the promise and the limitations of current LLM-based humor generation.

Our first stage is related to chain-of-thought reasoning. CoT prompting can improve multi-step reasoning by encouraging models to generate intermediate steps before the final answer (Wei et al., 2022). For smaller or more deployable models, teacher-generated rationales can also be used as supervision, allowing a stronger model to act as a reasoning teacher for a smaller one (Ho et al., 2023). We adapt this idea to constrained humor generation: the rationale is not meant to be shown to the evaluator, but to help the model identify a better humorous target during training and inference.

Our second stage is related to preference optimization. DPO provides a simpler alternative to RLHF by directly optimizing pairwise preferences without separately fitting a reward model (Rafailov et al., 2023). In humor generation, preference-based supervision is especially attractive because absolute humor scores are noisy, whereas pairwise judgments between a stronger and a weaker joke are often easier to construct. This is consistent with recent humor-specific alignment work such as Chen et al. (2024). Unlike prior Chinese humor work that focuses on crosstalk scripts or pun generation, we target short-form constrained humor on SemEval prompts and combine CoT supervision, teacher-built preferences, and deterministic cleaning in one pipeline.

3 System Overview

Our system consists of three stages: (1) setup-and-punchline CoT-SFT, (2) teacher-constructed DPO, and (3) constrained decoding with deterministic post-processing. Figure 1 summarizes the full workflow, from official data transformation to training, inference, and final submission cleaning.

3.1 Base Model

We use `Qwen2.5-7B-Instruct (4-bit)` as the base model. To fit training into limited GPU memory, we adopt parameter-efficient fine-tuning with LoRA (Hu et al., 2022) and 4-bit quantization for memory efficiency (Dettmers et al., 2023), using rank 16, alpha 16, and dropout 0.05. This setup preserves the instruction-following ability of the base model while allowing efficient adaptation to the humor generation task.

3.2 Setup-and-Punchline CoT-SFT

To reduce shallow or literal generations, we convert each training target into a two-part format consisting of a **setup/reasoning block** followed by a **punchline block**. We deliberately use the terms *setup* and *punchline* in the paper because they better match humor terminology than our earlier shorthand labels. Operationally, the setup block explains the absurdity, target, or contrast behind the joke, and the punchline block realizes the final humorous response.

For the CoT-SFT stage, we use 1,000 teacher-rewritten examples, one for each official Chinese training instance. In the released code, these rewrites are produced by a stronger teacher model accessed through an OpenAI-compatible API, with `deepseek-chat` as the default teacher model in the data-construction script. We use separate teacher prompts for headline inputs and word-pair inputs, but both prompts request the same two-part output: first explain the humorous angle, then provide the joke. The supervision objective is standard next-token prediction over the full formatted output.

3.3 Teacher-Constructed Preference Pairs

After CoT-SFT, we further align the model using DPO (Rafailov et al., 2023). We construct preference pairs automatically from the same task inputs. For each prompt, the **chosen** response is the teacher-generated rewrite and the **rejected** response is a weaker baseline generation for the same prompt. In our repository, the rejected side comes from a baseline TSV file, while the chosen side is loaded from the teacher-rewritten JSONL. This yields 1,000 preference pairs.

We explored two variants during development. In **plain DPO**, the chosen item is a teacher-written humorous answer and the rejected item is a weaker

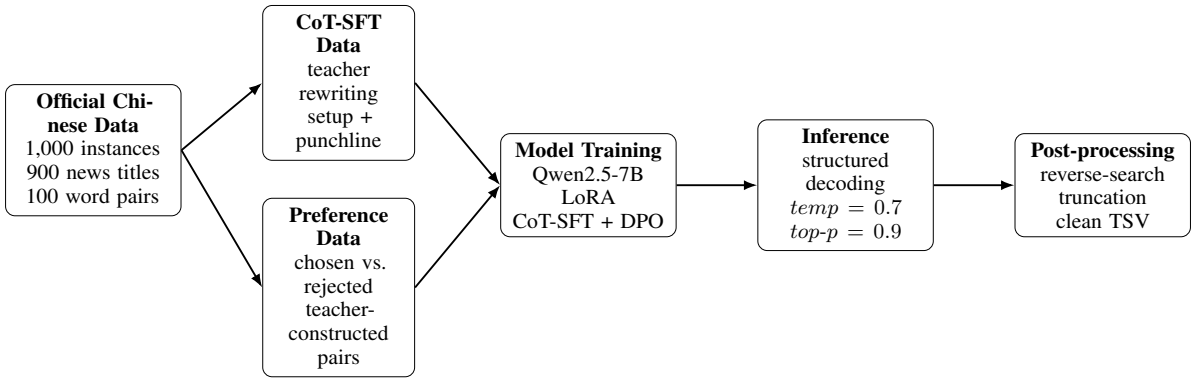

Figure 1: Pipeline of our system.

direct response. In **CoT-DPO**, the chosen item keeps the setup-and-punchline structure, while the rejected item is the shorter baseline response. The final official submission used the **CoT-DPO** configuration, which is also the one shown in the repository commands. In this paper, "sharper and better structured" therefore refers to an operational criterion rather than to a separate scalar score: the preferred output is the teacher rewrite because it is usually more concrete, more scene-building, and better organized than the baseline response.

We optimize the model with the standard DPO objective following Rafailov et al. (2023). Our implementation does not train a separate reward model and does not rely on human-annotated score gaps. Instead, it uses teacher-constructed pairwise supervision as a practical approximation for stylistic alignment.

3.4 Inference Strategy

At inference time, we retain the structured reasoning-first generation pattern because it improves consistency. However, this also creates a format-control problem, since intermediate reasoning traces are not allowed in the final submission. To balance creativity and stability, we use the following decoding parameters: temperature $= 0.7$, top-$p = 0.9$, and repetition penalty $= 1.2$. These settings reduce degenerate repetition while preserving enough diversity for humor generation.

3.5 Deterministic Post-processing

Post-processing is a core part of our system. Raw outputs may contain leaked setup blocks, repeated substrings, formatting residue, or malformed quotation marks. We therefore apply a deterministic sanitization pipeline with three main steps.

First, in **reverse-search extraction**, we search backward in the raw generated string to locate the final punchline marker and extract only the content that follows it. This helps discard earlier rationale text or malformed content.

Second, in **repetition truncation**, we apply heuristic rules to detect abnormal repeated substrings or gibberish-like loops. If such a loop is found, the output is truncated before the repetition expands.

Third, in **format sanitization**, we remove markdown markers, prompt residue, malformed quotation artifacts, and other characters that can reduce readability or break TSV formatting.

A concrete example is helpful. Suppose the raw model output contains a setup block such as "Now that registration is available nationwide, why are people still rushing to queue on a symbolic festival day?" followed by the punchline "Civil-affairs offices should offer a distributed-registration subsidy." Our cleaner discards the setup block and keeps only the final punchline for submission. Likewise, a malformed raw string such as "Intel executives promised the two sides would be 'deeply coupled' " "" is normalized to a single readable sentence without duplicated quotation marks. These examples match the explicit extraction and quote-cleaning rules implemented in our released cleaning script.

3.6 Example

Table 1 shows a concrete example of how one training instance is transformed into CoT-SFT and CoT-DPO resources. We use a prompt about the first Qixi Festival after nationwide marriage registration became available, where many cities saw a small peak in registrations. This example is representative of our pipeline because the joke relies on a clear social contrast: registration becomes adminis-

tratively easier, yet people still rush to complete it on a symbolic day.

4 Experimental Setup

We use only the official Chinese training data provided by the task organizers and do not rely on external corpora for data augmentation (Castro et al., 2026). The official Chinese training set contains 1,000 instances: 900 news-title prompts and 100 word-pair prompts. For local development, we split these 1,000 instances into 95% training data and 5% validation data, yielding 950 training instances and 50 validation instances. The final competition inference is run on 300 evaluation instances.

For CoT-SFT, we train on 1,000 examples with maximum sequence length 2048, 150 training steps, and an effective batch size of 8. For DPO, we train on 1,000 preference pairs for 1 epoch (125 steps) with learning rate 5×10^{-6}, $\beta = 0.1$, maximum prompt length 512, maximum total length 1024, and an effective batch size of 8.

Our implementation uses Unsloth and the Hugging Face ecosystem. Final large-scale inference is performed on an NVIDIA A100 40GB GPU. Generating outputs for all 300 evaluation instances took approximately 38 minutes.

The official task evaluation is based on the competition's live leaderboard, which reports a rating, a 95% confidence interval, and the number of votes (Castro et al., 2026). We use this as our main external evaluation signal. For internal analysis, we compare different model configurations and inspect representative examples qualitatively.

5 Results

5.1 Official Results

Table 3 summarizes the official live leaderboard results for SemEval-2026 Task 1, Subtask A (Chinese). Our final submission (**hugang11**) achieved a rating of 991, with a 95% confidence interval of [958, 1036], based on 209 votes, placing it in the second rank group.

We compare this result with the official baseline provided by the task organizers, which is based on Gemini 2.5 Flash with simple task-specific prompts and an explicit output-language instruction for Chinese. The official baseline obtained a rating of 1053, with a 95% confidence interval of [1003, 1090], based on 210 votes, and ranked in the first rank group.

These results show that our system was competitive, but did not outperform the organizer-provided baseline on the live leaderboard.

5.2 Internal Analysis

Internally, our pipeline suggests a trade-off between humorous structure and output stability. CoT-style supervision and DPO make outputs more structured and stylistically sharper, but they also increase the risk of rationale leakage and malformed text. In practice, deterministic post-processing was necessary to convert raw generations into a fully compliant final submission.

Table 2 summarizes the main training and inference settings of our system. Table 4 shows the direct effect of deterministic post-processing on submission validity. At the data-construction level, our official 1,000-instance Chinese training set was transformed into three aligned resources: plain DPO pairs, CoT-SFT targets, and CoT-DPO pairs. During development we tested both DPO variants, but the final submission used CoT-DPO because it produced sharper jokes, even though it also increased the need for cleanup.

As shown in Table 4, the raw output file contained 35 malformed lines even though 300 valid `id-text` rows could still be parsed from it. After cleaning, the file was converted into a fully compliant 300-instance submission with no malformed lines and no empty outputs. The cleaning pipeline also modified 107 generated texts, indicating that post-processing was not merely cosmetic but a necessary component of the final system.

5.3 Error Analysis

We identify three main error types. First, the system remains weak on highly temporal internet slang and prompts requiring rapidly changing cultural knowledge. Second, on longer or compositionally difficult prompts, the reasoning stage can lock onto the wrong humorous target, producing a fluent but misdirected joke. Third, the explicit setup block introduces leakage risk, including leaked tags and repetition artifacts, which must be handled downstream by post-processing.

Qualitatively, the system performs best when the prompt contains a clear semantic contrast that can be exaggerated or reframed into a concise satirical scene. It performs worse when humor depends on subtle cultural references that are not recoverable from the input alone.

Field	Content
Input prompt	“First Qixi Festival after nationwide marriage registration became available; many cities saw a small peak in marriage registrations.”
Plain target	“Marriage should also come with a cooling-off period.”
Teacher rewrite: setup	“If registration is available nationwide, why are people still rushing to queue on a symbolic festival day? The joke targets the contrast between administrative convenience and performative romantic urgency.”
Teacher rewrite: punchline	“Civil-affairs offices should offer a distributed-registration subsidy: anyone who marries on an ordinary weekday receives a ‘cooling-off-period certified’ certificate.”
CoT-DPO construction	Chosen response = the full teacher rewrite (setup + punchline); rejected response = the shorter direct target. This pair explicitly teaches the model to prefer a better motivated punchline over a flatter one-liner.

Table 1: Concrete example of training data transformation used in our pipeline.

Item	Value
Base model	`Qwen2.5-7B-Instruct-bnb-4bit`
Tuning method	LoRA + 4-bit quantization
LoRA config	rank 16, alpha 16, dropout 0.05
Train data	1,000 Chinese instances
Data split	95% train / 5% validation
CoT-SFT data	1,000 examples
CoT-SFT length	2048
CoT-SFT steps	150
CoT-SFT batch size	8
DPO data	1,000 pairs
Final DPO variant	CoT-DPO
DPO epochs / steps	1 / 125
DPO learning rate	5×10^{-6}
DPO beta	0.1
DPO max prompt len.	512
DPO max total len.	1024
Decoding	temp 0.7, top-p 0.9
Repetition penalty	1.2
Inference GPU	NVIDIA A100 40GB
Inference workload	300 instances
Inference time	about 38 minutes

Table 2: Main training and inference settings.

System	Rating	95% CI	Votes	Group
Official baseline	1053	[1003, 1090]	210	1
hugang11 (ours)	991	[958, 1036]	209	2

Table 3: Official leaderboard results on Subtask A (Chinese).

Measure	Before	After
Body lines in TSV	335	300
Parsed `id-text` rows	300	300
Malformed lines	35	0
Unique IDs	300	300
Missing IDs	0	0
Empty text rows	0	0
Texts modified by cleaning	–	107

Table 4: Effect of post-processing on submission validity.

6 Conclusion

We presented the **hugang11** system for **SemEval-2026 Task 1 (MWAHAHA)**, participating in the Chinese track of Subtask A. Our system combines CoT-augmented supervised fine-tuning, teacher-constructed DPO, and deterministic post-processing in a single humor-generation pipeline. Built on Qwen2.5-7B-Instruct-bnb-4bit, our final submission achieved a live leaderboard rating of 991 and placed in the second rank group. Our findings suggest that for creative generation tasks such as humor generation, alignment-oriented training is helpful, but robust inference-time control remains essential. In future work, we plan to explore stronger preference data construction, controlled human evaluation, and safer humor generation with improved cultural grounding.

Acknowledgments

We thank the organizers of SemEval-2026 Task 1 for creating the MWAHAHA shared task and evaluation platform. We also thank the developers of Qwen, Unsloth, and the open-source libraries used in our experiments.

A Additional Implementation Details

We provide additional low-level implementation details, including training scripts, decoding settings, post-processing rules, and data construction utilities, in our public code release.[2]

B Additional Post-processing Details

Our sanitization pipeline includes heuristics for removing rationale leakage, markdown artifacts,

[2] https://github.com/hugang1114-debug/hugang11-SemEval-2026-Task-1

malformed quotation marks, prompt residue, and abnormal repeated substrings. These rules are deterministic and are applied after generation but before final TSV export.

C Additional Examples

We include additional generation examples and failure cases in the supplementary material to illustrate both successful joke construction and common failure modes.

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, Online. International Committee on Computational Linguistics.

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, Bangkok, Thailand. Association for Computational Linguistics.

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

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14852–14882, Toronto, Canada. Association for Computational Linguistics.

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*.

Jianquan Li, XiangBo Wu, Xiaokang Liu, Qianqian Xie, Prayag Tiwari, and Benyou Wang. 2023. Can language models make fun? a case study in Chinese comical crosstalk. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7581–7596, Toronto, Canada. Association for Computational Linguistics.

Tyler Loakman, William Thorne, and Chenghua Lin. 2025. Who's laughing now? an overview of computational humour generation and explanation. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 780–794, Hanoi, Vietnam. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Yubo Xie, Junze Li, and Pearl Pu. 2021. Uncertainty and surprisal jointly deliver the punchline: Exploiting incongruity-based features for humor recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 33–39, Online. Association for Computational Linguistics.