DANGNT@SGU at SemEval-2026 Task 1: A Two-Stage Mistral Generator with DistilBERT Reranking for English Humor Generation

Tan Loc Nguyen[†], **Dang Tuan Nguyen**[‡*]
[†]Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam
[‡]Faculty of Information Technology, Saigon University, Ho Chi Minh City, Vietnam
nguyenloctan.0409@gmail.com dangnt@sgu.edu.vn

Abstract

We describe DANGNT@SGU's system for the English track of SemEval-2026 Task 1 (MWAHAHA), Subtask A (text-based humor generation). Our pipeline combines a two-stage QLoRA-adapted generator based on `mistralai/Mistral-7B-Instruct-v0.2` with a DistilBERT reranker trained to distinguish jokes from non-jokes. The generator is first adapted on a raw joke corpus for general humor style, then further tuned on synthetic task-format instruction–response pairs for Word Inclusion and News Headline prompts. At inference time, we generate five candidates per input, optionally enforce lexical constraints for Word Inclusion prompts, and rerank candidates with the classifier. In the official English Subtask A results, our team DANGNT@SGU obtained Elo 962 (95% CI: 926–986), ranking 13th. The system is practical, reproducible, and based entirely on open models and public data.

1 Introduction

SemEval-2026 Task 1 (MWAHAHA) studies constrained humor generation; in Subtask A, systems generate jokes under explicit constraints such as mandatory word inclusion or a news-headline prompt (Castro et al., 2026). The task is evaluated with human pairwise judgments aggregated into Elo-style rankings, making output quality and perceived funniness the main optimization target.

We describe our **English-only** submission, which follows a simple engineering-oriented design: a QLoRA-adapted Mistral-7B-Instruct generator plus a lightweight DistilBERT reranker. Because the task does not provide gold joke targets for the official prompts, we create synthetic supervision from a public joke corpus and the official English input file. Our team DANGNT@SGU ranked **13th** in the English leaderboard with Elo **962** (95% CI: 926–986) (Chiruzzo, 2026). The main contributions are: (1) a two-stage generator adaptation strategy (general humor → task constraints), (2) a simple joke-vs.-non-joke reranker for candidate selection, and (3) a best-of-5 inference pipeline with optional lexical constraint enforcement for Word Inclusion prompts.

*Corresponding author: Dang Tuan Nguyen (dangnt@sgu.edu.vn).

2 Task Setup

We participate in SemEval-2026 Task 1 (MWAHAHA), Subtask A, English track (Castro et al., 2026). The task includes two prompt types:

- **Word Inclusion**: generate a short joke that must include two specified words.
- **News Headline**: generate a short joke related to a given news headline.

The official English TSV is used at inference time, but no gold joke targets are provided for direct supervised training on the official prompts; this motivates our synthetic-data pipeline. Official evaluation is human-centered and pairwise, and leaderboard scores are reported as Elo ratings with confidence intervals, which motivates our best-of-5 generation and reranking strategy.

3 Method

The overall system is illustrated in Figure 1, and can be summarized into four major stages: (1) synthetic data preparation, (2) Stage 1 generator training (general humor adaptation), (3) Stage 2 generator training (constraint adaptation), and (4) best-of-5 inference with reranking.

3.1 Data sources

We use three public data sources:

- **Humor corpus**: the *Short Jokes* dataset on Kaggle by Abhinav Moudgil

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2986–2990

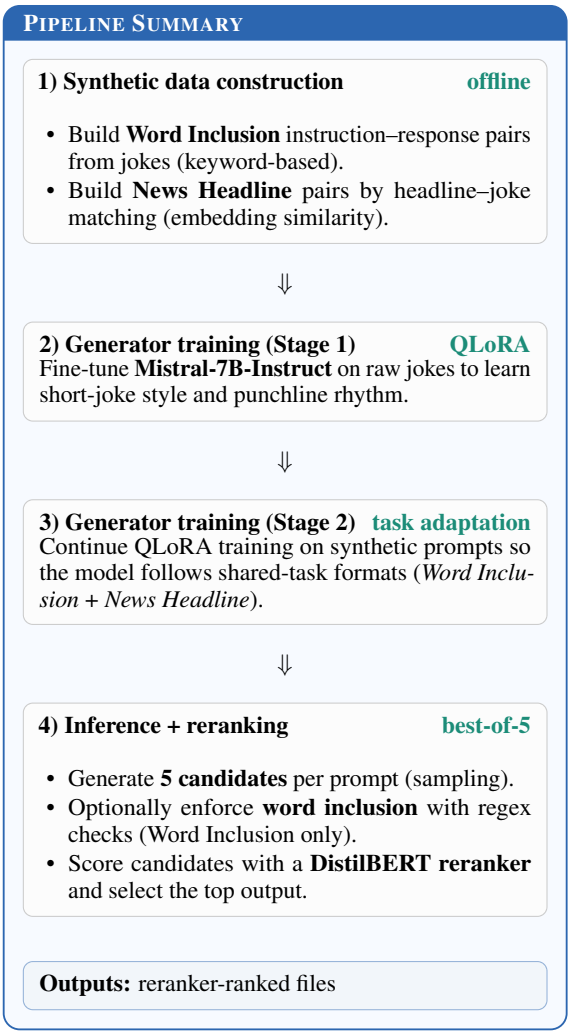

Figure 1: Compact overview of our two-stage generator + DistilBERT reranker pipeline for English Subtask A.

(`shortjokes.csv`) (Moudgil, n.d.), used for Stage 1 humor-style adaptation and for synthetic pair generation. This dataset has also been used in prior humor research (Chen and Soo, 2018).

- **Task inputs**: the official English input file for Subtask A (TSV), which contains both Word Inclusion and News Headline prompts.
- **Non-humor corpus for reranker negatives**: English Wikipedia from Hugging Face (`wikimedia/wikipedia`, snapshot `20231101.en`) (Hugging Face and Wikimedia Foundation, n.d.), used as negative examples for DistilBERT reranker training.

Because the task does not provide gold target jokes for the official prompts, we generate synthetic supervision by constructing task-format instruction–response examples using a public joke corpus (and official headlines for NH), with responses sampled or matched from the joke corpus.

3.2 Synthetic data preparation

Word Inclusion synthetic pairs. For each joke in the joke corpus, we extract keyword candidates using POS tagging and dependency parsing (spaCy). We prioritize content words (NOUN, PROPN, ADJ, VERB) and filter out stopwords and punctuation. We then select two keywords and build an instruction of the form:

> Create a short joke that includes the words "`w1`" and "`w2`".

The original joke text is used as the response. This yields a large set of synthetic word-inclusion training pairs.

News Headline synthetic pairs. To construct headline-style supervision, we embed official news headlines and joke texts using a sentence-transformer model. For each headline, we retrieve jokes whose cosine similarity exceeds a threshold (0.4), and create instruction–response pairs:

> Create a short joke related to this headline: "..."

This procedure introduces topical supervision, although it is noisy because the joke corpus is not natively aligned to real news.

Data balancing. The Word Inclusion synthetic pairs are much more numerous than the News Headline pairs. To avoid overfitting to one prompt type, we cap the ratio used in Stage 2 training (production setting: approximately 10:1 for Word Inclusion:News Headline).

3.3 Two-stage generator training

Backbone model and adaptation. We use `mistralai/Mistral-7B-Instruct-v0.2` as the generator. Training is performed with 4-bit quantization (QLoRA-style setup) and LoRA adapters inserted into attention projection layers (q, k, v, o).

Stage 1: general humor style adaptation. In Stage 1, we train the generator directly on the raw joke corpus as an instruction-following generation task, so the model learns short-joke style, rhythm, and punchline structure.

Stage 2: task-format adaptation. In Stage 2, we continue fine-tuning from the Stage 1 adapters using the synthetic task-format prompts (Word Inclusion + News Headline). This stage teaches the model to follow the exact prompt structures expected by the shared task while preserving the humor style learned in Stage 1.

3.4 DistilBERT reranker

We train a lightweight reranker to score humor quality proxies:

- **Positive class**: jokes from the humor corpus
- **Negative class**: non-humorous text sampled from English Wikipedia (Hugging Face `wikimedia/wikipedia`, snapshot `20231101.en`)

The reranker is a binary DistilBERT classifier. At inference time, we use its output score to choose the better candidate among multiple samples from the generator.

3.5 Inference and submission generation

At inference time, the pipeline reads the official English file and formats prompts automatically based on the row type (Word Inclusion vs. News Headline). We generate **five candidates** per input using sampling (best-of-5), then score all five candidates with the DistilBERT reranker and keep the top-ranked output for the main submission.

4 Experimental Setup

4.1 Resources and software

We use the following main components:

- **Generator**: Mistral-7B-Instruct-v0.2 (HF id: `mistralai/Mistral-7B-Instruct-v0.2`) (Jiang et al., 2023)
- **Reranker**: DistilBERT (`distilbert-base-uncased`) (Sanh et al., 2019)
- **Headline–joke matching**: Sentence-Transformers model `all-MiniLM-L6-v2` (Reimers and Gurevych, 2019)
- **Linguistic preprocessing**: spaCy (`en_core_web_sm`) (Honnibal et al., 2020)
- **Frameworks**: Transformers, Datasets, PEFT/LoRA, and TRL (Wolf et al., 2020; Lhoest et al., 2021; Hu et al., 2022; Dettmers et al., 2023)

4.2 Main hyperparameters

Table 1 summarizes the main hyperparameters and training/inference settings used in our final English Subtask A system. We include the generator, reranker, and decoding configurations to support reproducibility.

Component	Configuration
Stage 1 generator	Mistral-7B-Instruct-v0.2 + QLoRA (4-bit NF4, bf16 compute)
Stage 1 LoRA	$r = 64$, $\alpha = 16$, dropout $= 0.1$, target modules = q/k/v/o projections
Stage 1 training	3 epochs, batch size 16, grad accum 1, LR 2×10^{-4}, cosine schedule, warmup 0.03
Stage 2 generator	Continue training from Stage 1 adapters (same base model, 4-bit)
Stage 2 balancing	Word:News = 10:1 cap (enabled in production mode)
Stage 2 training	3 epochs, batch size 16, grad accum 1, LR 1×10^{-4}, cosine schedule, warmup 0.03
Reranker data	Positive = jokes, Negative = `wikimedia/wikipedia` (`20231101.en`)
Reranker	DistilBERT binary classifier, max length 128, train/test split = 90/10
Reranker training	1 epoch, batch size 64, LR 2×10^{-5}, eval/save steps = 2000
Inference	best-of-5 generation, reranker selection, optional regex word filter
Decoding defaults	temperature = 0.7, top-k = 50, max new tokens = 75, seed = 42

Table 1: Main system settings used in our English Subtask A pipeline.

4.3 Reproducibility

We release the full implementation and scripts for our submission pipeline.[1]. The script exports reranker-ranked outputs from top-1 to top-5 for each input. We submitted one official runs: `task-a-en_top1.tsv` (primary). Outputs `top2`–`top5` are provided for local inspection and reproducibility analysis. We also document the training schedule, key hyperparameters, and exact scripts used in the final pipeline. The system relies only on open pretrained models and public datasets, and our task-specific supervision is generated by rule-based preprocessing (keyword extraction, embedding-based headline pairing, and threshold filtering), which is deterministic once the seed and preprocessing configuration are fixed. For replication, we recommend using the library versions listed in the repository and the same decoding settings reported in this paper. Minor variation may still occur across hardware/software environments when sampled generation is used.

5 Results and Analysis

5.1 Official SemEval result

Table 2 reports the official English Subtask A result of our primary run (reranker top-1 selection) (Chiruzzo, 2026). The official evaluation is based on human pairwise judgments aggregated into Elo, so local automatic metrics should be interpreted only as proxy diagnostics.

[1] https://github.com/tanloc49/dangnt-sgu-mwahaha

Item	Value
Leaderboard name	DANGNT@SGU
Subtask / language	Subtask A / English
Official Elo	962
95% confidence interval	[926, 986]
Official rank	13

Table 2: Official result reported by the task organizers for our English submission.

5.2 Local proxy ablation protocol

Because the organizers do not release gold target jokes for the official prompts, we evaluate system variants on a local dev-style set for controlled comparison. Our local set contains 100 prompts sampled from the official English input format (50 Word Inclusion, 50 News Headline), and is used only for internal ablation analysis.

We compare three configurations to isolate the effect of stage-wise generator training under a fixed decoding/selection setting (top-1 output; best-of-1) and without reranking: (i) zero-shot Mistral-7B-Instruct, (ii) Stage 2-only is trained from the same base model as zero-shot (no Stage 1), using only the synthetic task-format data, and (iii) Stage 1+Stage 2. For Word Inclusion prompts, we measure raw lexical compliance without any post-hoc regeneration; the optional regex-based regeneration described in our inference procedure is used only in the final submission pipeline.

We report two lightweight diagnostics:

- **W.I. Success (%)**: percentage of Word Inclusion outputs that contain both required words (case-insensitive word-boundary regex).
- **Avg. DistilBERT**: mean reranker score of the final selected output (proxy for joke-likeness).

5.3 Quantitative ablation results

Table 3 highlights two main trends under a fixed decoding/selection setting (top-1 output). First, task-format adaptation (Stage 2) is the main driver of lexical constraint satisfaction, improving W.I. Success substantially over zero-shot prompting (68% $\rightarrow$ 96%, i.e., 34/50 $\rightarrow$ 48/50 on Word Inclusion). Second, Stage 1 humor-style adaptation improves the joke-likeness proxy strongly (0.58 $\rightarrow$ 0.74) while maintaining the same W.I. Success, suggesting that the two-stage design helps separate *format following* from *humor style learning*.

Overall, this stage-wise ablation supports the final design choice: Stage 2 is necessary for prompt compliance, and Stage 1 improves stylistic quality. At the same time, these numbers remain proxy-based and do not replace human preference evaluation.

5.4 Error analysis

Our local error inspection suggests three common failure modes. **(1) Constraint–quality gap:** for Word Inclusion prompts, the model often includes both required words but produces a weak or literal punchline, indicating that lexical compliance alone does not guarantee humor quality. **(2) Headline relevance drift:** for News Headline prompts, outputs are often joke-like but only loosely related to the headline content, likely due to the noise in our synthetic headline–joke matching procedure. **(3) Reranker preference bias:** the DistilBERT reranker sometimes favors short, generic joke patterns over more creative candidates, because it is trained as a binary joke/non-joke classifier rather than on human pairwise preferences. These observations complement our stage-wise ablation and motivate future work on task-aware reranking and stronger prompt–response relevance modeling.

6 Discussion

Our results suggest that the system benefits from a modular design: Stage 1 learns short-joke style, Stage 2 adapts the model to the shared-task prompt formats, and reranking provides a practical selection rule under stochastic decoding (our reranker top-1 run outperforms the top-2 run in official Elo). This decomposition also makes the pipeline easier to debug because each component has a distinct role.

The ablation study also clarifies the limits of the current approach. The DistilBERT reranker is helpful as a *joke-likeness* proxy, but it is not trained on human pairwise preferences, so improvements in Avg. DistilBERT may not translate directly into Elo gains. Similarly, our synthetic News Headline supervision improves fluency and format compliance but does not fully solve headline relevance.

Two practical directions follow from these findings. First, a task-aware reranker trained on prompt–response relevance (and ideally human preference signals) would better match the official evaluation objective. Second, constrained decoding could replace or reduce regex-based post-checking for Word Inclusion prompts, improving robustness without relying on reactive filtering.

Pipeline configuration	W.I. Success (%) ↑	Avg. DistilBERT ↑
Zero-shot Mistral-7B-Instruct	68	0.44
Stage 2 only (task-format adaptation)	96	0.58
Stage 1 + Stage 2 (humor-style + format)	**96**	**0.74**

Table 3: Stage-wise local proxy comparison on a 100-prompt dev-style set (50 Word Inclusion, 50 News Headline), using the top-1 output (best-of-1, no reranker). W.I. Success is computed on the Word Inclusion subset; Avg. DistilBERT is the mean score of our DistilBERT joke-likeness classifier over all 100 final outputs. These proxy metrics are for internal system comparison only and are not directly comparable to the official human Elo evaluation.

7 Conclusion

We presented DANGNT@SGU's English system for SemEval-2026 Task 1 (MWAHAHA) Subtask A: a two-stage QLoRA-adapted Mistral generator with DistilBERT reranking for Word Inclusion and News Headline prompts. Our team DANGNT@SGU ranked 13th on the English leaderboard (Elo 962). The pipeline is lightweight, reproducible, and fully based on open models and public data, providing a practical baseline for controllable humor generation.

References

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)*.

Santiago Castro and Luis Chiruzzo. 2026. Submission list, annotation, and paper submission. Google Groups post to the *semeval-2026-task-1-humor-gen* mailing list, February 3, 2026.

Luis Chiruzzo. 2026. Final results. Google Groups post to the *semeval-2026-task-1-humor-gen* mailing list, February 21, 2026.

Albert Q. Jiang et al. 2023. Mistral 7B. arXiv preprint arXiv:2310.06825.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. Zenodo.

Thomas Wolf et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Quentin Lhoest et al. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Edward J. Hu et al. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Peng-Yu Chen and Von-Wun Soo. 2018. Humor Recognition Using Deep Learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 113–117, New Orleans, Louisiana. Association for Computational Linguistics. doi:10.18653/v1/N18-2018.

Abhinav Moudgil. n.d. *Short Jokes*. Kaggle dataset. https://www.kaggle.com/datasets/abhinavmoudgil95/short-jokes. Accessed: 2026-02-25.

Hugging Face and Wikimedia Foundation. n.d. *wikimedia/wikipedia*. Hugging Face Datasets. https://huggingface.co/datasets/wikimedia/wikipedia. Configuration used: 20231101.en. Accessed: 2026-02-25.