MINDS at SemEval-2026-Task 1: Enhancing Humor Generation through RAG and Synthetic DPO Alignment

Sina Eskandari* **Seyed Amirreza Mousavi*** **Amirreza Rahimi***
Mona Pouresmaeil* **Marcello Vitaggio*** **Claudio Savelli†**
Riccardo Coppola† **Flavio Giobergia†**
Politecnico di Torino
* {firstname.lastname}@studenti.polito.it
† {firstname.lastname}@polito.it

Abstract

Humor generation presents significant challenges due to subjectivity and the limitations of automatic metrics. In this work, we address Task 1 of SemEval 2026 (Subtask A) by evaluating three instruction-tuned models (Llama 3.1, Gemma 2, and Qwen 2.5) via a round-robin LLM judging framework. We investigate the impact of Retrieval-Augmented Generation and Direct Preference Optimization (DPO) on performance. Our results identify Llama 3.1 as the strongest baseline and demonstrate that DPO consistently improves humor quality across configurations. These findings confirm the efficacy of LLM-based judging as a practical training signal for optimizing subjective generation tasks.

1 Introduction

Humor is a fundamental aspect of human communication, enabling social bonding and creativity. Generating humor, however, remains challenging for Natural Language Processing systems because comedic success depends on subtle semantic incongruity, cultural knowledge, and highly subjective human preferences. Unlike tasks such as translation or summarization, humor lacks a clear notion of correctness, making both modeling and evaluation difficult.

Recent advances in Large Language Models (LLMs) have improved fluency and coherence in text generation, and modern instruction-tuned models can produce jokes and wordplay. Nevertheless, humor generation still faces two key limitations: automatic metrics correlate poorly with perceived funniness, and human evaluation is expensive and hard to scale. The Humor Generation shared task (Castro et al., 2026) of SemEval 2026 addresses this challenge by evaluating systems through human pairwise preferences under explicit generation constraints, but this setting also removes access to labeled data or explicit reward signals.

In this work, we investigate whether small instruction-tuned LLMs can be further improved in this unsupervised setting by using LLMs themselves as scalable judges and training signals. We compare three open models using a round-robin self-judging framework, study the effect of retrieval-augmented generation, and apply Direct Preference Optimization (DPO) using synthetic preferences derived from LLM judgments.

Our contributions are:

- A scalable LLM-based evaluation framework for constrained humor generation.
- An empirical comparison of three small instruction-tuned LLMs and the impact of retrieval augmentation.
- An application of DPO with synthetic preferences for improving humor generation in a subjective task.

2 Related Work

Computational humor has evolved from template-based systems to LLMs, which offer strong generative capabilities but often require adaptation to capture subtle forms of wit and wordplay. Parameter-efficient fine-tuning methods such as Low-Rank Adaptation (LoRA) (Hu et al., 2021) enable cost-effective specialization of large models. Beyond supervised fine-tuning, aligning models with subjective preferences is crucial for humor generation. While Reinforcement Learning from Human Feedback (RLHF) is commonly used for alignment, Direct Preference Optimization (Rafailov et al., 2024) provides a simpler and more stable alternative by directly optimizing on preference data.

The usage of using larger models to produce a useful signal for smaller LLMs has been explored thoroughly in literature. Indeed, the usage of LLM-assisted pseudo-labeling can help overcome the lack of annotated data in generation-related tasks

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2967–2971

(Tan et al., 2024; Borra et al., 2024), highlighting the potential of LLMs as scalable sources of supervision.

Humor generation can also benefit from contextual grounding. Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) incorporates external knowledge during inference, helping models remain relevant to input prompts and reducing hallucinations. For evaluation, traditional n-gram metrics are poorly suited to subjective tasks like humor. As a result, LLM-as-a-Judge frameworks (Zheng et al., 2023; Verga et al., 2024) have emerged as scalable proxies for human evaluation, enabling pairwise comparison and preference-based assessment aligned with human perception.

3 Task Description

We address the English-only track of Subtask A (Text-based Humor Generation) of the SemEval shared task. The goal is to generate a short joke j from a prompt p under one of two constraints: (1) **Word Inclusion**, where two specified rare words must appear in a coherent joke, and (2) **News Headline**, where a given headline must be humorously reinterpreted. Examples of both settings are shown in Table 1. The official test set contains 300 prompts (275 headlines and 25 word-inclusion cases), and systems must produce exactly one constraint-compliant joke per prompt.

Constraint	Input Prompt	Generated Joke
Word Inclusion	**Words:** *Spatula, Nebula*	I tried to flip a pancake with a **spatula** so hard it reached the Orion **Nebula**.
News Headline	**Headline:** "NASA finds water on Moon"	NASA found water on the Moon. Finally, a place where the Wi-Fi is bad but the "tide" is high.

Table 1: Examples of the two humor generation constraints.

The task is fully unsupervised: no gold labels or humor scores are provided, and automatic metrics such as BLEU are not suitable for evaluating funniness. Instead, systems are evaluated via pairwise human comparisons. For each prompt, two anonymous outputs are shown to annotators, who select the funnier joke. Rankings are computed using an Elo rating system based on accumulated pairwise judgments, and final leaderboard positions are determined by statistical significance at the 95% confidence level rather than raw Elo scores alone.

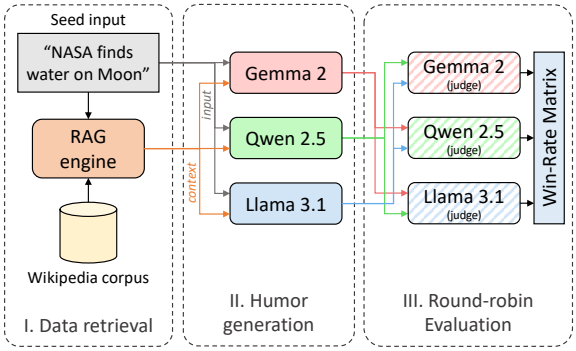

Figure 1: Overview of the proposed RAG-based humor generation pipeline. The framework consists of three phases: retrieval of contextual information, multi-model humor generation, and round-robin evaluation producing a final win-rate matrix.

4 Methodology

Our approach combines retrieval, parameter-efficient fine-tuning, and preference alignment to optimize instruction-tuned LLMs for constrained humor generation in a fully unsupervised setting. As illustrated in Figure 1, our pipeline integrates contextual retrieval, multi-model joke generation, and LLM-based pairwise evaluation. All systems are evaluated using a standardized judging framework mirroring the SemEval protocol: for each prompt, two outputs are compared by a judge model, and performance is measured in terms of win rates across pairwise competitions. We additionally explored a Supervised Fine-Tuning strategy and DPO to improve model performance.

4.1 Retrieval-Augmented Generation

To incorporate external knowledge, we implement a Retrieval-Augmented Generation pipeline over a 25,000-document subset of Wikipedia (not-lain, 2023), indexed using the `mxbai-embed-large-v1` embedding model (Mixedbread AI, 2024). For each input prompt, the top $k = 4$ documents are retrieved via cosine similarity, and up to 1,200 characters of context are appended to the prompt before generation.

To encourage grounding in retrieved content, we reduce the sampling temperature to 0.7 (vs. 0.9 in non-RAG baselines), maintain top-$p = 0.9$, and apply a repetition penalty of 1.15. For word-inclusion prompts, we enforce constraint satisfaction through

a two-attempt retry mechanism when mandatory keywords are missing. The resulting outputs are evaluated within our round-robin pairwise comparison framework.

4.2 Supervised Fine-Tuning

For Llama-3.1-8B-Instruct, we explore parameter-efficient supervised fine-tuning using LoRA (Rank 16, $\alpha = 16$) with 4-bit quantization, training for 2 epochs with an effective batch size of 8.

We experiment with two datasets. The first is a **Synthetic Short Jokes Dataset** consisting of approximately 3,500 entries derived from the Short Jokes (amoudgl, 2026) corpus and enriched via an LLM into instruction-reasoning-joke triplets. The second dataset is a **Distilled "Best Jokes" Dataset**, containing 1,200 top-performing outputs selected from an initial multi-model tournament (Llama, Qwen, Gemma). This dataset represents high-quality model-generated humor identified through pairwise preference. In this setting, we use a learning rate of 2×10^{-5} with linear scheduling and Paged AdamW (8-bit) optimization.

4.3 Direct Preference Optimization

We additionally explore the option of aligning a model with humor preferences, with DPO. Instead of human annotations, we construct a synthetic preference dataset

$$\mathcal{D} = \{(x^{(i)}, y_w^{(i)}, y_l^{(i)})\}_{i=1}^{N}$$

using LLM-based pairwise judgments, where y_w and y_l denote the preferred and rejected completions for prompt x.

We optimize the model using the standard DPO objective:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log \sigma\Big(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\Big)\Big] \quad (1)$$

where σ is the logistic function and β controls deviation from the reference policy.

We use Llama-3.1-8B-Instruct as the base model and apply Rank-Stabilized LoRA (Rank 16, $\alpha = 16$) under 4-bit quantization. Training is performed for one epoch with learning rate 5×10^{-6}, sequence length 2048, 50 warmup steps, AdamW (8-bit), weight decay 0.01, and effective batch size 8. We set $\beta = 0.05$ to balance alignment strength and policy stability. Performance is assessed through win-rate comparisons against baseline and retrieval-augmented systems.

5 Results

All evaluations are conducted using the pairwise judging framework described in Section 4. For each prompt, two anonymized outputs are compared by a judge model, and performance is reported in terms of win rate (W.R.) across pairwise competitions.

5.1 Baseline Model Evaluation

We first conduct a triangular round-robin tournament among Llama 3.1 8B, Gemma 2, and Qwen 2.5. Results are summarized in Table 2. A clear hierarchy emerges: **Llama 3.1 8B** consistently outperforms the other models, achieving strong win margins when judged by both Gemma 2 (56.2%) and Qwen 2.5 (60.5%). In contrast, Gemma 2 and Qwen 2.5 exhibit near-parity when judged by Llama (49.7% vs. 49.1%).

Overall, the tournament establishes the baseline ranking: Llama 3.1 first, Gemma 2 second, and Qwen 2.5 third.

Judge Model	Model A	W.R.	Model B	W.R.	Ties
Llama 3.1	Gemma 2	**49.7%**	Qwen 2.5	49.1%	1.2%
Gemma 2	Llama 3.1	**56.2%**	Qwen 2.5	43.7%	0.2%
Qwen 2.5	Llama 3.1	**60.5%**	Gemma 2	39.5%	0%

Table 2: Tournament Results: Win Rates for the Pairwise Comparisons by Judge

5.2 Supervised Fine-Tuning

Experiment 1: Synthetic Short Jokes Dataset Fine-tuning Llama 3.1 on the Synthetic Short Jokes Dataset (labeled as "Llama FT (Best Jokes Subset)" in Table 3) results in severe performance degradation. As shown in Table 3, the base model overwhelmingly outperforms the fine-tuned variant (77.5-80% win rates), suggesting overfitting or catastrophic forgetting.

Judge Model	Model A	W.R.	Model B	W.R.	Ties
Qwen 2.5	Llama Base	**77.5%**	Llama FT (Best)	22.5%	0%
Gemma 2	Llama Base	**80.0%**	Llama FT (Best)	20.0%	0%

Table 3: Win Rate comparison: Llama 3.1 Base vs. Llama FT (Best Jokes Subset)

Experiment 2: Distilled "Best Jokes" Dataset Training on the Distilled "Best Jokes" Dataset (labeled as "Llama FT (SemEval Best)" in Table 4) reduces the performance gap but still fails to surpass the base model (Table 4). While margins shrink relative to Experiment 1, Llama Base remains consistently superior.

Judge Model	Model A	W.R.	Model B	W.R.	Ties
Qwen 2.5	Llama Base	**57.5%**	Llama FT	41.5%	1.0%
Gemma 2	Llama Base	**68.0%**	Llama FT	32.0%	0%

Table 4: Win Rate comparison: Llama 3.1 Base vs. Llama FT (SemEval Best)

5.3 Retrieval-Augmented Generation

We next evaluate retrieval-augmented variants of all three models. As shown in Table 5, RAG substantially alters the performance hierarchy. In contrast to the baseline setting, **Qwen 2.5 + RAG** emerges as the dominant model, outperforming both Gemma and Llama across judges.

Judge Model	Model A	W.R.	Model B	W.R.	Ties
Qwen 2.5	Gemma 2	**55.8%**	Llama 3.1	44.2%	0%
Llama 3.1	Gemma 2	20.8%	Qwen 2.5	**79.2%**	0%
Gemma 2	Llama 3.1	24.4%	Qwen 2.5	**75.5%**	0.1%

Table 5: RAG Tournament Results: Pairwise Comparisons by Judge, in terms of Win Rate.

5.4 Impact of Direct Preference Optimization

Applying DPO to Llama-3.1-8B-Instruct produces our aligned model, **DPO Llama v1**. When trained on synthetic preference pairs derived from round-robin judgments, v1 substantially outperforms both the base Llama and Qwen+RAG (Table 6). We show an example of generated jokes in Table 7.

Comparison	DPO W.R.	Other W.R.	Ties
DPO vs Base Llama	**66.6%**	33.4%	0%
DPO vs Qwen+RAG	**68.9%**	30.8%	0.3%

Table 6: Win Rate comparison of DPO Llama v1 ("DPO") performance against other methods. Gemma used as a judge.

5.5 Additional experiments

We present the results of the baseline models (Llama, Gemma, Qwen) on the test dataset, in Table 8. In this case, Qwen outperforms the other models consistently, with Llama being a close second, and Gemma behind.

Input Headline	Ryanair to cut 1 million more passenger seats in Spain
DPO Llama	Ryanair's latest move to slash a million Spanish seats means passengers will now get charged extra for arriving anywhere – including therapy sessions.
Qwen + RAG	Ryanair slashes a million seats in Spain, proving they can fly even when they're empty.

Table 7: Comparison of model generations for the same input headline.

Judge Model	Model A	W.R.	Model B	W.R.	Ties
Llama 3.1	Gemma 2	20.66%	Qwen 2.5	**77.67%**	1.67%
Gemma 2	Llama 3.1	42.67%	Qwen 2.5	**57.33%**	0%
Qwen 2.5	Llama 3.1	**90%**	Gemma 2	10%	0%

Table 8: Tournament Results: Pairwise Comparisons for Test Dataset in terms of Win Rate.

We constructed the following additional systems: DPO Llama v2 (trained on preferences derived from the test data), and DPO applied to Qwen 2.5 ("DPO Qwen"), in addition to the previously mentioned "Qwen+RAG". Their comparison against DPO Llama v1 is shown in Table 9.

Comparison	DPO Llama v1 W.R.	Other W.R.	Ties (%)
DPO Llama v2	**75.33%**	23%	1.67%
DPO Qwen	**77%**	22.33%	0.67%
Qwen+RAG	**80.33%**	19%	0.67%

Table 9: Comparison of DPO Llama v1 performance against other methods, in terms of Win Rate. Gemma used as a judge.

Iterative Refinement and Final Submission Further fine-tuning of v1 on aggregated preference pairs yields **DPO Llama v3**, which achieves a modest improvement over v1 (52.33% vs. 46.33%).

For final submission, we apply a best-of-two strategy between v1 and v3 using our internal judging framework. To ensure word-inclusion compliance, we verify mandatory keywords and replace non-compliant outputs when necessary.

5.6 Official Competition Results

Our final system achieved an Elo rating of 1022 (95% CI: [989, 1054]) and secured rank 2 in SemEval-2026 Task 1 (Subtask A).

6 Conclusion

We investigated humor generation in a fully unsupervised setting, where no gold labels or automatic metrics are available. We introduced a scalable LLM-as-a-judge framework to both evaluate systems and construct synthetic preference data for optimization.

Our experiments highlight three main findings. First, among the evaluated instruction-tuned models, Llama 3.1 and Qwen 2.5 emerge as strong baselines across different evaluation settings, in constrained humor generation. Second, retrieval-augmented generation substantially changes the performance hierarchy, with Qwen 2.5 benefiting the most from external context. Third, DPO with synthetic LLM judgments consistently improves humor quality, outperforming both the base model and the best RAG-based system.

These results confirm that LLM-based judging can serve as an effective training signal for subjective creative tasks. As future work, we plan to further analyze the role of retrieval by performing controlled ablations on the amount and structure of retrieved context, in order to better understand when and how external knowledge benefits humor generation.

References

amoudgl. 2026. short-jokes-dataset: Python scripts for building 'short jokes' dataset. https://github.com/amoudgl/short-jokes-dataset. Accessed: 2026-03-03.

Federico Borra, Claudio Savelli, Giacomo Rosso, Alkis Koudounas, and Flavio Giobergia. 2024. Malto at semeval-2024 task 6: Leveraging synthetic data for llm hallucination detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1678–1684.

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)*.

Edward J Hu, Yelong Shen, Phil Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Mixedbread AI. 2024. mxbai-embed-large-v1. Sentence embedding model.

not-lain. 2023. Wikipedia (embedded version). Pre-embedded Wikipedia corpus on Hugging Face.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957.

Pat Verga, David Cohen, and Ellie Pavlick. 2024. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*.

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. *arXiv preprint arXiv:2306.05685*.