

Cartoon Caption Humor Quality Assessment and Generation with DPO and LoRA

Stanford CS224N {Custom}

Isaiah Flores

Department of Computer Science
Stanford University
isaiahf@stanford.edu

Katherine Wang

Department of Computer Science
Stanford University
khw@stanford.edu

Abstract

We investigate whether learning the relative humor quality of cartoon caption pairs through Direct Preference Optimization (DPO) can improve LLM alignment with human humor judgments, a longstanding NLP challenge requiring contextual inference and cultural knowledge. Using the New Yorker Caption Contest dataset — a corpus containing over 9,500 pairwise caption preferences derived from human contest judgments — we train a text language model to predict the human-preferred caption given a cartoon text description, using DPO with parameter-efficient fine-tuning via Low-Rank Adaptation (LoRA). Previous work on zero-shot/few-shot models often diverge from human references, suggesting that likelihood alignment is insufficient for humor assessment and directly encoding human preferences may improve upon baselines. We evaluated multiple base models including GPT-2 (117M parameters), Qwen2.5-0.6B, and Qwen2.5-1.5B, and find that stronger base models combined with DPO yield substantial improvements in pairwise preference accuracy, with our largest difference gain to be 12.5 percentage points from 49.6% to 62.1% with Qwen2.5-1.5B. We also examine whether our preference-optimized model generates higher-quality, more humorous captions for unseen cartoons.

1 Key Information to include

- Mentor: Mirac Suzgun
- External Collaborators (if you have any): No
- Sharing project: No
- Late Days (if you are using any):

Team member	Late days remaining
Isaiah Flores	5
Katherine Wang	6
Total pooled	11
Team size	2
Late days applied for the report	1

2 Introduction

Humor has been positioned as a long-standing challenge in NLP as it depends on contextual inference, cultural knowledge, and subjective human judgment. Our project shifts from evaluation-only LLM benchmarks toward preference-based training for humor understanding, leveraging human judgment signals from the New Yorker Caption Contest to train models that align with human assessments of joke quality.

Specifically, we investigate whether Direct Preference Optimization (DPO) improves a language model’s ability to understand humor by ranking cartoon captions in alignment with human judgments. Our work builds on the benchmark introduced in *Do Androids Laugh at Electric Sheep?* (Hessel et al., 2023) [1], which shows that zero-shot, few-shot, and fine-tuned models often diverge from human rankings, suggesting that likelihood and semantic alignment alone are insufficient for humor assessment. Though the dataset encodes human preferences (e.g., winning vs. non-winning captions), the paper does not explicitly optimize models for these preferences.

We address this gap by applying DPO on caption preference pairs and use LoRA to ensure feasibility under limited computational resources. Through this, we seek to learn whether explicit optimization on human preference pairs enables a smaller language model to better approximate human humor judgments than likelihood-based baselines.

3 Related Work

Computational humor has been studied through both benchmark-style evaluations and explicit modeling of humor mechanisms, including humor detection, classification, and generation. Prior research has shown that humor can be broken down into interpretable categories including irony, sarcasm, exaggeration, incongruity, self-deprecation, and wit.

Recent publications demonstrate that modern neural models, especially transformer-based models, are capable of learning these categories and differentiating between them. Li et al. [2] analyze multiple deep learning approaches for humor detection and show that transformer models outperform baselines and earlier neural models, highlighting the importance of contextual representation of humor. More recently, humor classification has been formulated as a six-way genre and technique prediction task, finding that fine-tuned LLMs achieve strong performance when explicit humor supervision is provided [4].

These previous findings complement the results of *Do Androids Laugh at Electric Sheep?* [1] as they also demonstrate that state-of-the-art models struggle with deeper humor understanding. Overall, these works suggest that explicit humor-type modeling would improve model understanding and generation of jokes. Our work extends this line of research by applying preference-based optimization via DPO, directly training models to align with human humor judgments rather than relying on classification or generation objectives alone.

4 Approach

Building on the “quality ranking” benchmark task introduced by Hessel et al. [1], we model cartoon caption humor quality assessment as a preference learning task. Given a text description x of a New Yorker Caption Contest cartoon, y^+ and y^- represent a pair of captions, where y^+ is the chosen caption winner and y^- is a rejected non-winner. We apply DPO to a fine-tuned, text-based language model to optimize relative preference:

$$\mathcal{L}_{\text{DPO}} = -\log \sigma \left(\beta \left[\log \pi_{\theta}(y^+ | x) - \log \pi_{\theta}(y^- | x) \right] \right) \quad (1)$$

where π_{θ} is the fine-tuned policy and β controls the strength of the preference signal [3]. For efficient training, we apply LoRA to freeze the base model weights while updating only low-rank adapter weights.

To reduce spurious preference cues (e.g., verbosity effects), we constrain captions to 25 words for captions exceeding 80 tokens. We standardize policy input using a fixed prompt template and caption normalization. Metadata fields (e.g., contest identifiers) are retained only for analysis and not provided to the model.

The prompt template used during training is:

```
You are writing captions for The New Yorker Caption Contest.
Write a single New Yorker-style cartoon caption in ≤ 25
words.
Avoid explaining the joke. Output only the caption.
```

Cartoon description: {description}
Caption:

Our primary contribution is the application of DPO with LoRA to humor preference ranking as this approach is not explored in prior work on this dataset. Hessel et al. [1] evaluate models such as GPT-3 and GPT-4 in zero-shot and few-shot settings but do not explicitly optimize for human preference pairs. Li et al. [2] and Wu et al. [4] focus on humor detection and classification rather than preference-based alignment. By directly optimizing on pairwise human judgments via DPO, we move beyond likelihood-based scoring and classification toward a training objective that mirrors the structure of the task itself. We additionally evaluate across model scales (i.e. GPT-2 (117M parameters), Qwen2.5-0.6B, and Qwen2.5-1.5B) to introduce a systematic comparison of how model capacity interacts with DPO training on this benchmark.

5 Experiments

5.1 Data

We use the New Yorker Caption Contest corpus released by Hessel et al. [1], available at https://github.com/jmhessel/caption_contest_corpus. This dataset includes cartoon text descriptions, submitted human-written captions, and human caption preference rankings. As a pre-processing step, we create preference pairs from the 'ranking' subset by pairing winning captions against non-winning captions for the same cartoon. The 'ranking' dataset is split into training (9,576 pairs), validation (507 pairs), and test (513 pairs) sets.

After applying our length-based filtering (max 80 caption tokens, max 30-token length difference between caption pairs), the effective train split used for DPO is approximately 9,200 pairs.

Below is a small table showing sample cartoon descriptions alongside chosen and rejected captions.

Cartoon Description	Chosen Caption	Rejected Caption
A man is giving a hot dog to a woman outside a pyramid.	Of course it's kosher — who do you think built this place?	I'll admit, the grill marks really could pass for hieroglyphs.
Some musicians are on stage. A monster is among them.	This season, we welcome a new corporate sponsor.	Well, he was tired of playing second fiddle anyway.
Cave people are having a get together in a cave.	So how much of the story is autobiographical?	I tell you, I just didn't see this coming.

Table 1: Sample preference pairs from the New Yorker Caption Contest dataset. Each row shows a cartoon description alongside the human-preferred (chosen) and human-rejected caption.

5.2 Evaluation method

We evaluate models using pairwise preference accuracy, following the benchmark. Given a cartoon description and caption pair, the model scores each caption via its conditional log-likelihood and predicts the preferred caption as the higher-scoring one. Accuracy is defined as the fraction of pairs for which the model correctly assigns a higher score to the human-preferred caption:

$$\text{Acc} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(s_{\theta}(y_i^+ | x_i) > s_{\theta}(y_i^- | x_i)), \quad (2)$$

We also report the average log-likelihood gap between preferred and dispreferred captions to capture preference strength:

$$\Delta\ell = \frac{1}{N} \sum_{i=1}^N [\log p_{\theta}(y_i^+ | x_i) - \log p_{\theta}(y_i^- | x_i)] \quad (3)$$

A larger positive gap indicates stronger preference separation. Random performance corresponds to 50% accuracy and a gap near zero.

Pairwise preference accuracy and log-likelihood gap are standard metrics for evaluating DPO-trained models [3] and preference learning benchmarks [1].

Our milestone report considered NDCG (Normalized Discounted Cumulative Gain) as an evaluation metric, following its use in information retrieval settings. However, because our task uses only binary pairwise comparisons (one chosen caption vs. one rejected caption per instance), NDCG offers no additional signal beyond pairwise accuracy given its design for ranked lists of three or more items. We therefore report only pairwise preference accuracy and log-likelihood gap, which directly capture the structure of our task and are standard in the DPO literature [3].

5.3 Experimental details

We evaluate three model configurations: (1) a GPT-2 (117M parameters) baseline (no DPO training) and experimental run with DPO and LoRA trained on the full training set, (2) Qwen3-0.6B baseline and with DPO and LoRA trained on the full training set for 1 epoch, and (3) Qwen2.5-1.5B-Instruct baseline and with DPO and LoRA trained on the full training set for 1 epoch.

For DPO, we use a learning rate of 2×10^{-4} and preference scaling parameter $\beta = 0.1$. LoRA adapters use rank $r = 16$, $\alpha = 32$, and dropout 0.05. Maximum sequence length is 256 tokens. Effective batch size is 16 (per-device batch 4, gradient accumulation 4). Training uses the AdamW optimizer with a cosine learning rate schedule and 3% warm up.

5.4 Results

Table 2 summarizes pairwise preference accuracy and mean log-probability gap across all model configurations trained on the full datasets and tested on the validation and test sets.

Model	Val Pre	Val Post	Test Pre	Test Post	Test $\Delta\ell$
GPT-2 (117M)	47.9%	53.8%	49.6%	58.2%	+0.611
Qwen3-0.6B	49.7%	56.6%	49.8%	57.2%	+3.233
Qwen2.5-1.5B-Instruct	51.9%	61.7%	49.6%	62.1%	+10.774

Table 2: Pairwise preference accuracy and mean log-probability gap ($\Delta\ell$) on validation and test sets. Pre-DPO accuracy reflects the base model before training; Post-DPO reflects the model after DPO fine-tuning on the same evaluation split.

The untrained GPT-2 baseline achieves 47.9% on validation and 49.6% on the test set — near-random, with a negative log-probability gap (-0.071 on test), indicating the base model slightly prefers the rejected caption on average. After DPO training on the full dataset, GPT-2 improves substantially to 58.2% on the test set with a positive gap of $+0.611$, demonstrating that DPO successfully realigns the model’s likelihood estimates in the direction of human-preferred captions, but performance remains limited from significant improvements.

Why is the GPT-2 baseline below chance? The untrained GPT-2 model scores slightly below 50%, with a negative log-probability gap (-0.071), indicating a weak preference for rejected captions. We hypothesize two contributing factors. First, a *perplexity bias*: winning captions tend to rely on incongruity, unexpected register shifts, and cultural references, all of which are low-probability sequences for a model trained on generic web text. The rejected captions, by contrast, tend toward more literal observations using higher-frequency token patterns, which GPT-2 naturally scores higher. Second, a *length bias*: even after our filtering (max 80 tokens per caption, max 30-token difference), small systematic differences in token count can tip a near-random model below chance, since language models assign higher total log-probability to shorter sequences. DPO addresses both of these biases by training the model to assign higher likelihood to the human-preferred caption regardless of its surface-level probability from the base model.

Qwen3-0.6B, with similar approximately random baselines, trained for 1 full epoch, reaches 56.6% on validation and 57.2% on test, with a larger log-probability gap of $+3.233$. This larger gap relative

to GPT-2 likely reflects the fact that instruction-tuned models assign more confident probability scores to tokens, which increases the difference between chosen and rejected caption scores.

Qwen2.5-1.5B-Instruct with a similar near-random baseline, also trained for 1 full epoch, achieves the best result at 62.1% test accuracy with a log-probability gap of +10.774. This is consistent with our hypothesis that a larger model with more parameters is better able to capture the nuanced preference signals in the data set.

Training dynamics for the 1.5B model show clear convergence: the fraction of training pairs ranked correctly rises from 0.47 at step 25 to over 0.99 by step 1,700, while DPO loss drops from ~ 0.70 to below 0.03 by the end of training. The score gap between chosen and rejected captions also grows steadily throughout training, confirming that the model increasingly learns to prefer human-selected captions over time (See Appendix A for more details on training dynamics).

6 Analysis

Effect of DPO across model families. Across all three models, DPO consistently improves pairwise preference accuracy over their respective untrained baselines. The magnitude of improvement scales with model capacity: GPT-2 gains +8.6 percentage points on test, Qwen3-0.6B gains +7.4 percentage points, and, notably, Qwen2.5-1.5B reaches +12.5 percentage points. This is consistent with findings in preference learning literature that a stronger reference model leads to more effective DPO training [3], as the reference distribution π_{ref} is initialized from the base model’s weights.

Log-probability gap and over-optimization. The Qwen3-0.6B model shows a remarkably large log-probability gap (+3.233) relative to GPT-2 (+0.611), despite a lower test accuracy (57.2% vs. 58.2%). This pattern may reflect partial over-optimization, meaning that as the model sharpens its probability mass strongly on training-distribution examples, it generalizes less reliably to unseen pairs. DPO is known to be susceptible to reward over-optimization when the KL divergence from the reference grows large [3]. The Qwen2.5-1.5B model, trained for 1 full epoch with a stronger base, achieves the best accuracy (62.1%), suggesting that increased model capacity can absorb more training signal without losing generalizability.

Where DPO succeeds and fails. To understand what the model has learned, we examine representative success and failure cases from our best configuration (Qwen2.5-1.5B-Instruct, $\beta=0.01$, 63.1% test accuracy).

Successes. The model correctly ranks captions where the preferred caption deploys situational irony or register contrast. For a cartoon of a talking fish on a dinner plate, the model prefers "Just water for me, thanks." over "I flunked out of my school." The winner reframes the fish’s predicament through dry understatement, while the rejected caption is a loosely connected pun. Similarly, for a woman playing guitar while a man shovels snow, the model correctly picks "It’s like you’re not even trying to shovel on beat." over "Let me tell you a story ’bout a man named Joe. . .". The winner integrates both visual elements (music + shoveling) into a single incongruous complaint, while the rejected caption is a generic folk-song reference.

Failures. The model struggles with captions requiring multi-step cultural inference. For a pig at a complaint department window, the model prefers "I wish more people knew how to use the subjunctive." over the correct winner, "Whenever I eat an apple, everyone assumes that I’m dead." — the winning caption requires linking pigs, apples, and the phrase "pig with an apple in its mouth" (a roasted pig), a chain of reasoning the model misses. For a banana on a therapist’s couch, the model prefers "Tell me about your parents’ split." over "It’s normal to feel empty after a split." — both captions use the double meaning of "split" (banana split / emotional separation), but the winner adds a third layer ("empty" = hollow inside after peeling), which the model fails to value. These failure patterns suggest DPO improves surface-level preference signals (tone, register fit) but does not reliably capture cultural cues for deeper inferential humor.

Human agreement baseline. Hessel et al. [1] report human performance estimates on the quality ranking task, noting that three annotators each attempted 100 randomly sampled instances. On the quality ranking task, human annotators achieved approximately 68% accuracy against crowd-aggregated labels. Importantly, the authors emphasize that human performance is not an upper bound because "labels are not generated by a single human and tastes can vary" [1]. In theory, it’s possible for the model to reconstruct New Yorker editor preferences more reliably than an untrained

or specialized individual. The dataset’s preference labels reflect aggregate crowd judgments, and humor is inherently subjective. The paper even explicitly states that “disagreement about what is funny is expected and valid” [1]. This context suggests that our highest reported accuracy of 62.1% may be closer to human-level agreement than it first appears, and that the remaining error reflects genuine ambiguity in the preference labels rather than purely model limitations.

Generation quality and LLM-judge evaluation. We ran the DPO-trained Qwen2.5-1.5B-Instruct model through a generative evaluation on the test split (513 examples, temperature 0.7, max 64 new tokens). The generated captions were qualitatively degraded and produced fragmented or incoherent text in many cases. For example, given the pyramid hot-dog scene, the model generated “th? Sure! hot Dog! Here’sThe tourists looking hisc...”—a stark contrast to the human-preferred caption (“Of course it’s kosher—who do you think built this place?”). This pattern was consistent across examples: the model produces tokens that reflect partial caption-like structure but lack coherent humor or even grammatical form.

To quantify this, we conducted an LLM-as-judge evaluation on 100 generated captions using Claude Haiku (claude-haiku-4-5-20251001) as an external judge. The judge prompt provides the cartoon description and presents two captions: Caption A (generated) and Caption B (human winner). Then, we ask: “Which caption is funnier or more appropriate for this cartoon? Reply with exactly one of: A, B, or TIE.” The judge uses temperature 0 for deterministic outputs and ties are counted as 0.5 toward accuracy. We do not validate the judge against human raters on this specific task due to the constrained caption quality; future work on improved caption generation should measure judge-human agreement on a held-out sample to further confirm reliability.

Judge Preference	Count
Prefer generated (A)	0
Prefer human winner (B)	67
Tie	33

Table 3: LLM-judge evaluation of DPO-generated vs. human-preferred captions ($N = 100$).

The judge preferred the human caption in 67% of cases, declared a tie in 33%, and never preferred the generated caption—meaning generated captions were preferred or tied in only 33 of 100 cases (16.5% when counting ties as half). This aligns with the intent of DPO in optimizing the model’s *ranking* of existing captions via log-likelihood but not the *generation* distribution. The adapter was trained without an SFT warm-up stage on caption-style completions, leaving the generation head uninformed about the New Yorker stylistic register. This disconnect between preference ranking and generation quality is a known limitation of applying DPO directly on base models without SFT [3], and motivates a two-stage SFT→DPO pipeline for future work.

β sensitivity ablation. To understand how the DPO preference strength parameter β affects performance, we trained Qwen2.5-1.5B-Instruct with three values of β on the full training set for 1 epoch, holding all other hyperparameters constant.

β	Val Acc	Test Acc	Test Δ_ℓ
0.01	60.2%	63.1%	+55.387
0.1	61.7%	62.1%	+10.774
0.5	56.8%	53.9%	+0.662

Table 4: β ablation on Qwen2.5-1.5B-Instruct. All runs use the same LoRA configuration and full training set (1 epoch).

A small β (0.01) allows the policy to diverge further from the reference, producing the highest test accuracy (63.1%) but an extremely large log-probability gap (+55.4). This likely indicates that the model’s probability distribution has shifted dramatically from the base model. This aggressive divergence risks reward over-optimization as the model may be exploiting distributional shortcuts rather than actually learning generalizable preference signals. This can and is degrading performance on out-of-distribution examples. A large β (0.5) over-regularizes the policy, keeping it too close to the reference and limiting its ability to learn preferences on out-of-distribution samples. In this

ablation, test accuracy drops to 53.9% which is only marginally above the 51.9% pre-DPO baseline. The intermediate $\beta = 0.1$ balances preference learning against reference divergence, achieving strong accuracy (62.1%) with a moderate log-probability gap (+10.8).

Data scaling. We also ran GPT-2 + DPO across increasing training-set sizes (50, 100, 3,000, and all 9,576 pairs) to understand data scaling behavior. With 50 samples, the model achieves no significant improvement (49.6% test), with 3,000 samples it reaches 53.3% test, and with the full training set it reaches 58.2% test. This confirms that preference-ranking tasks require sufficient data coverage and that DPO is data-efficient but still requires scale.

Limitations. Several limitations constrain our conclusions. First, we use text descriptions rather than the actual cartoon images; the full task requires visual grounding, and text descriptions may omit visual humor cues. Second, we do not run a supervised fine-tuning (SFT) warm-up before DPO, which means the reference model has no prior exposure to New Yorker-style caption generation. Third, our caption generation evaluation with an LLM-judge was limited by the quality of generated captions. Future work on improved caption generation should validate whether log-likelihood ranking correlates with human judgments of generated captions.

7 Conclusion

We applied DPO with LoRA to the task of humor preference ranking on the New Yorker Caption Contest dataset, training models to assign higher conditional log-likelihood to human-preferred captions. Our results demonstrate that DPO is an effective method for aligning language models with human humor judgments. Most notably, we found that pairwise preference accuracy improves from 49.6% to 62.1% following DPO training on Qwen2.5-1.5B-Instruct, a gain of 12.5 percentage points. We also found that model capacity is a key factor; the larger instruction-tuned base models benefit more from DPO training to reach higher accuracy with a single epoch of training on the full dataset.

Our analysis highlights a tension between preference ranking and generation quality: DPO improves ranking accuracy but does not directly improve the model’s ability to generate humorous captions. This motivates a two-stage pipeline — SFT followed by DPO — for future work, where the model is first trained to produce fluent, contextually appropriate captions before being fine-tuned to prefer funnier ones. Additional directions include: incorporating visual features using the `ranking_from_pixels` dataset split with a vision-language model and using an LLM judge to validate that preference-ranked models generate captions rated as funnier by external evaluators.

Team contributions

Katherine Wang implemented the data pre-processing and DPO for caption ranking. Isaiah Flores implemented LoRA and caption generation/LLM judge. Both parties worked on all reports and poster.

References

1. Hessel, Jack, Ana Marasović, Jena D. Hwang, Rowan Zellers, Robert Mankoff, Yejin Choi, Jeff Da, and Lillian Lee. *Do Androids Laugh at Electric Sheep? Humor “Understanding” Benchmarks from the New Yorker Caption Contest*. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), 2023. <https://aclanthology.org/2023.acl-long.41/>
2. Wu, Shih-Hung, Yu-Feng Huang, and Tsz-Yeung Lau. *Notebook for the CYUT Lab at CLEF 2024 JOKER Lab: Humour Classification According to Genre and Technique*. In Proceedings of the CLEF 2024 Conference and Labs of the Evaluation Forum, 2024. <https://ceur-ws.org/Vol-3740/paper-183.pdf>
3. Li, Zhuohang, Jiashuo Liu, and Yuci Wang. *Performance Analysis on Deep Learning Models in Humor Detection Task*. In Proceedings of the 2022 International Conference on Machine Learning and Knowledge Engineering (MLKE), pp. 93–97, 2022. IEEE. <https://doi.org/10.1109/MLKE55170.2022.00023>

4. Rafailov, Rafael, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. *Direct Preference Optimization: Your Language Model is Secretly a Reward Model*. NeurIPS 2023. <https://arxiv.org/abs/2305.18290>

Appendix

A Training Curves

The Qwen2.5-1.5B-Instruct DPO training run (2,394 steps, 1 epoch, full training set) showed the following progression of key metrics:

Step	Epoch	DPO Loss	Rewards/Acc (train batch)	Reward Margin
25	0.010	0.704	0.47	−0.012
125	0.052	0.616	0.76	+0.243
425	0.178	0.518	0.76	+0.701
825	0.345	0.463	0.83	+1.818
1225	0.512	0.222	0.92	+3.353
1625	0.679	0.154	0.95	+4.885
2025	0.846	0.028	0.99	+6.485
2375	0.992	0.029	0.99	+6.910

Table 5: Training dynamics for Qwen2.5-1.5B-Instruct + DPO.

Training rewards/accuracies converge near 1.0 by step $\sim 1,700$, while the reward margin ($\log p(\text{chosen}) - \log p(\text{rejected})$) grows from -0.012 to $+6.9$, confirming strong separation of chosen from rejected captions in the model’s internal scoring.

B GPT-2 Data Scaling Results

Training samples	Val Acc (post-DPO)	Test Acc	Test $\Delta\ell$
50	48.3%	49.6%	−0.079
100	47.1%	49.8%	−0.052
3,000	51.7%	53.3%	+0.203
Full (9,576)	53.8%	58.2%	+0.611

Table 6: GPT-2 + DPO performance across training set sizes.

More training data yields consistent improvements, with diminishing marginal gains and a sign change in $\Delta\ell$ occurring between 100 and 3,000 samples, indicating the model crosses from preferring rejected to preferring chosen captions at that scale.

C LoRA Adapter Configuration (Qwen2.5-1.5B-Instruct)

```
{
  "peft_type": "LORA",
  "base_model_name_or_path": "Qwen/Qwen2.5-1.5B-Instruct",
  "r": 16,
  "lora_alpha": 32,
  "lora_dropout": 0.05,
  "target_modules": ["q_proj", "v_proj"],
  "bias": "none",
  "task_type": "CAUSAL_LM"
}
```