HumorGen: Cognitive Synergy for Humor Generation in Large Language Models via Persona-Based Distillation

Anonymous ACL submission

Abstract

Humor generation poses a significant challenge for Large Language Models (LLMs), because their standard training objective (next-token prediction) inherently conflicts with the surprise and incongruity required for comedy. To bridge this gap, we introduce the **Cognitive Synergy Framework**, a methodology for generating high-quality humor data inspired by psychological theories of humor. Utilizing a *Mixture-of-Thought (MoT)* approach, we deploy six cognitive personas (e.g., *The Absurdist*, *The Cynic*) to synthesize diverse comedic perspectives for a given prompt. This framework produces a theory-grounded dataset, which we use to fine-tune a 7B-parameter student model. We further evaluate two alignment strategies, **Direct Preference Optimization (DPO)** and an offline group-relative variant **O-GRPO**, finding that neither improves over SFT. However, our 7B HumorGen model variants significantly outperform larger instruction-tuned baselines and achieve top-tier open-weight performance while remaining competitive with frontier proprietary systems. These results suggest that cognitively driven data curation is more critical than alignment algorithms or model scale for humor generation.

1 Introduction

Humor generation is a sophisticated creative task requiring mastery of context, nuance, and linguistic ambiguity (Khurana et al., 2024; Robison, 2024). While Large Language Models (LLMs) excel at logical reasoning, reliable humor generation remains an open problem because standard training objectives, which minimize perplexity, conflict with the incongruity and surprise required for comedy. This "alignment tax" often results in models that are safe and helpful but produce predictable, boring jokes or tedious explanations of humor.

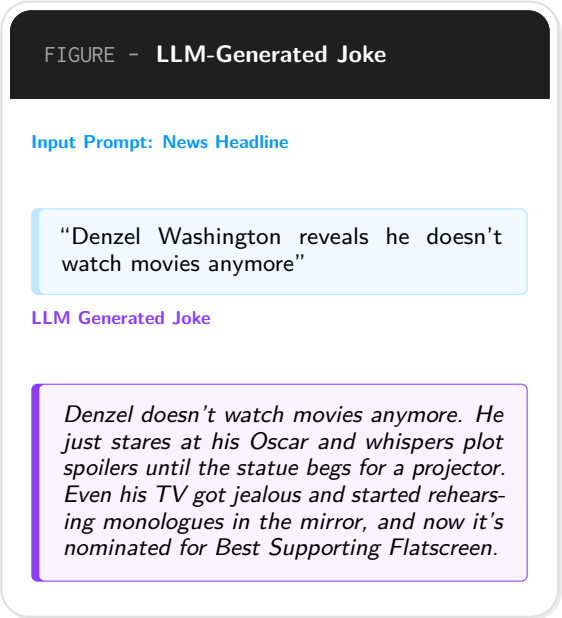

Figure 1: Example of an LLM-generated joke based on a news headline prompt, synthesized using the Cognitive Synergy Framework.

Recent efforts to improve LLM humor generation have focused on logical "thought leaps" (Zhong et al., 2024) or multistep reasoning (Wang et al., 2025). While these approaches improve performance on specific humor generation tasks, they often miss the diverse cognitive styles underlying human humor, and primarily rely on instruction tuning, limiting their ability to capture the variety of ways humans construct jokes.

To bridge this gap, we introduce the **Cognitive Synergy Framework** which advance beyond generic instruction tuning by operationalizing psychological humor theories within a *Mixture-of-Thought (MoT)* framework designed to encourage creative divergence. Traditional language modeling is highly susceptible

to mode collapse in creative generation, converging toward the most probable (and therefore most generic) continuations. By instantiating six distinct "cognitive personas" (e.g., *The Absurdist*, *The Cynic*) as latent experts within the MoT framework, we consistently route the generation process into the low-probability, high-variance regions of the semantic space where humor naturally occurs. This ensemble approach mitigates mode collapse and yields a diverse, theoretically grounded corpus of synthetic data, enabling us to distill multi-faceted humor generation capabilities from a frontier teacher model into a highly efficient 7B-parameter student model.

Due to the highly subjective nature of humor, we investigate whether preference alignment (e.g., **Direct Preference Optimization (DPO)** and **Offline Group Relative Policy Optimization (O-GRPO)**) improves over supervised fine-tuning. Our experiments show that neither alignment method visibly improves the models over the SFT baseline as DPO achieves similar performance to SFT, while O-GRPO is less impressive. Thus, under our setup, the alignment exercises did not improve the models, and the quality of the underlying cognitive data (Cognitive Synergy Framework) is the primary driver of generation performance.

Our contributions are:

- We introduce the **Cognitive Synergy Framework (CSF)**, a methodology that operationalizes psychological humor theories into six cognitive personas to generate diverse, theory-grounded humor data via Mixture-of-Thought.
- We investigate whether preference alignment (DPO, O-GRPO) improves over SFT for humor generation, finding that neither method yields consistent gains beyond a well-curated SFT baseline in this subjective humor generation domain.
- We introduce **Humor Transfer Bench (HTB)**, a 400-prompt benchmark spanning eight input domains to test generalization in textual humor generation.
- We show that our 7B student models, **HumorGen** variants, achieves top-tier open-weight performance on HTB and SemEval MWAHAHA humor-generation benchmark and remains competitive with much larger proprietary systems, showing that high-quality data can outweigh model size for humor.

2 Related Work

2.1 Computational Humor Generation

Computational humor research has largely focused on detection and recognition tasks (Jentzsch and Kersting, 2023; Dsilva, 2024), while humor generation remains comparatively underexplored. As a result, existing work on humor generation is often fragmented, targeting specific forms such as puns (Chen et al., 2024), domain-based humor (Shafiei and Saffari, 2025), punchline generation (Zhang et al., 2020), or language-based humor generation (Chen et al., 2023; Zhong et al., 2024). However, while classical humor theories (Lintott, 2016; Scheel, 2025; McGraw and Warren, 2010; Ajayi and Mitra, 2026a) offer valuable linguistic and semantic interpretations of humor, they do not directly translate into reliable generative mechanisms capable of consistently producing effective or high-quality humor.

2.2 Reasoning-Enhanced Humor Generation

With the rise of Large Language Models (LLMs), recent work has explored prompting and reasoning strategies for humor generation. Prior studies show that Chain-of-Thought (CoT) reasoning (Wei et al., 2022), while effective for logical tasks, is often misaligned with humor generation, which depends on divergence, incongruity, and non-linear associations (Zhong et al., 2024; Wang et al., 2025; Tikhonov and Shtykovskiy, 2024). As a result, CoT-based prompting often produces coherent but non-comedic outputs.

To address this, *Creative Leap of Thought (CLoT)* (Zhong et al., 2024) introduces leap-of-thought reasoning to encourage non-obvious associations during humor generation. The *LoL* framework (Wang et al., 2025) builds on *CLoT* by incorporation of external knowledge to support multi-hop creative reasoning, while related work explores structured multistep rea-

soning for humor generation (Tikhonov and Shtykovskiy, 2024). In contrast, naive prompting with ChatGPT (Jentzsch and Kersting, 2023) shows strong repetition patterns, with 90% of 1,008 generated jokes repeating only 25 templates. These findings suggest that standard reasoning paradigms remain limited for humor generation.

2.3 Preference Optimization in Creative Generation

Preference optimization methods have been widely studied for aligning LLMs with human preferences in subjective generation tasks (Yasuda and Toda, 2025; Lou et al., 2025; Vikhorev et al., 2024). With methods such as Direct Preference Optimization (DPO) and Group Relative Policy Optimization (GRPO) having demonstrated effectiveness across domains such as code and image generation (Govande et al., 2025; Tong et al., 2025).

In humor generation, prior work integrates Supervised Fine-Tuning (SFT) with DPO-based alignment (Wang et al., 2025). However, many preference optimization methods rely on online sampling or pairwise comparisons, which can be unstable and computationally expensive in highly subjective settings. We build on these ideas by additionally exploring an offline group-relative formulation (O-GRPO) using fixed preference datasets without online sampling.

3 The Cognitive Synergy Framework

Humor generation needs both coherence and surprise: a joke must first establish a plausible setup, then introduce an unexpected shift. Standard LLM decoding tends to favor high-probability next-token predictions, which often reduces surprise and yields generic outputs.

To address this, we introduce the **Cognitive Synergy Framework (CSF)**, which leverages Mixture-of-Thought (MoT) (Fein-Ashley et al., 2025) for humor generation by running multiple persona-guided reasoning paths in parallel, where each *Cognitive Persona* (see Section 3.2) encodes a different humor mechanism (e.g., absurdity, social critique, wordplay), enabling the model to explore diverse comedic directions instead of converging to a single safe continuation. The resulting candidate set is therefore both coherent and diverse, improving downstream selection and distillation for humor-focused training.

3.1 Divergent Reasoning via MoT

Unlike standard Chain-of-Thought (CoT) prompting, which optimizes for a single logical path, our framework generates K distinct reasoning traces in parallel. Given an input premise x, we sample a set of diverse reasoning paths $\{z_1, z_2, \ldots, z_K\}$ seeded by different cognitive priors, yielding a *pool* of K candidates that is retained in full. All candidates are subsequently ranked and used for training, so the model learns from the full distribution of generated joke candidates rather than a single top-ranked joke.

3.2 Cognitive Personas

To guide this diversity, we define six **Cognitive Personas**, each grounded in a specific psychological theory of humor (Table 1; see Appendix E for a full overview of the underlying theories). These personas act as soft constraints on the reasoning process, ensuring that our candidate pool covers a wide spectrum of comedic mechanisms.

By using these personas, we created a "synergy" between different styles of thought. This structural diversity proved critical for our subsequent alignment stage, as it provided a rich variety of distinct candidates for the model to learn from during preference optimization.

4 Methodology

We frame humor generation as a conditional language modeling task, where the goal is to generate a humorous response y given a context x. We first build a strong initialization model via supervised fine-tuning (SFT) using MoT-generated candidates and silver LLM supervision. Starting from this model, we explore two preference alignment strategies based on LLM-judged pairwise evaluations: Direct Preference Optimization (DPO) (Rafailov et al., 2023) and an offline group-relative variant, **O-GRPO**, both of which aim to better align outputs with learned preference signals.

4.1 Supervised Fine-Tuning (SFT)

This initial stage establishes baseline humor capabilities and internalizes the various cogni-

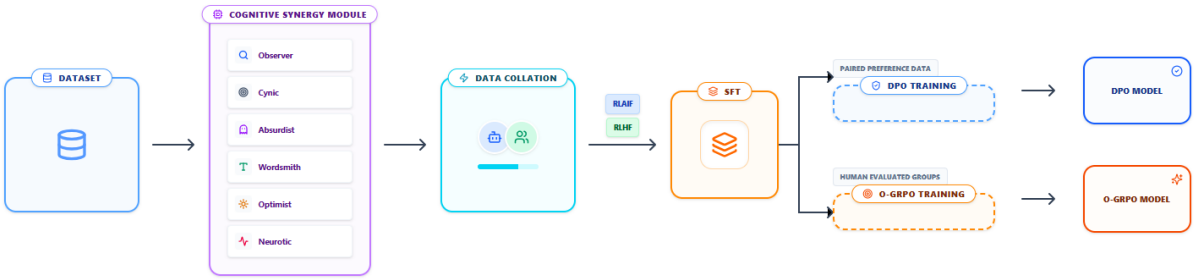

Figure 2: The HumorGen training pipeline. **(A) Generation:** Input headlines are processed by the Cognitive Synergy module (MoT), generating diverse candidates from 6 distinct personas. **(B) Collation:** Candidates are ranked via pairwise LLM judging (Llama 3.3-70B) to compute Elo ratings; the label "Human evaluated groups" in the diagram refers to this automated pairwise judging step, not literal human annotation. Human evaluation is conducted separately as a post-hoc validation of the judge's reliability (Section 6.5). **(C) SFT:** The base policy is fine-tuned on the top-ranked candidates. **(D) Alignment:** The model is further optimized via two parallel experimental branches: Pairwise DPO (top) or Group-Relative O-GRPO (bottom) driven by Elo-based preference data.

Persona	Humor Theory	Mechanism	Cognitive Focus
Neurotic	Relief Theory	Tension Release	Internal anxiety, overthinking, and social insecurity.
Cynic	Superiority Theory	Social Critique	Hypocrisy, biting sarcasm, and moral contradictions.
Observer	Incongruity	Social Mapping	Mundane minutiae and unwritten awkward social norms.
Wordsmith	Linguistic	Ambiguity	Puns, double entendres, and phonological play.
Optimist	Benign Violation	Recontextualization	Wholesome misinterpretations of potentially negative traits.
Absurdist	Incongruity	Surrealism	Non-sequiturs, dream logic, and fractured causality.

Table 1: The six Cognitive Personas used in our framework. We map each persona to a foundational humor theory and a specific cognitive focus to ensure divergent candidate generation.

tive personas. We construct a dataset $\mathcal{D}_{SFT}$ using a "Silver Teacher" protocol. Given the candidate pool $\mathcal{C}_{total}$ generated by our Mixture-of-Thought (MoT) ensemble, we employ a pairwise LLM evaluation system to compute Elo ratings for all candidates. We select the top-ranked candidates for each prompt based on these Elo ratings:

$$y^* = \text{argmax}_{y \in \mathcal{C}_{total}} \text{Score}_{LLM}(y|x)$$

We fine-tune a base Qwen-7B model using standard cross-entropy loss to maximize the likelihood of these "winner" responses:

$$\mathcal{L}_{SFT}(\theta) = -E_{(x,y^*)\sim\mathcal{D}_{SFT}}[\log \pi_\theta(y^*|x)]$$

This stage effectively distills the creative diversity of the larger teacher model into the student model.

4.2 Direct Preference Optimization (DPO)

To further align the model with humor preferences, we employ DPO using a dataset $\mathcal{D}_{DPO}$ of high-quality pairwise preferences derived from the LLM-judged Elo rankings. Each pair (y_w, y_l) consists of a high-ranking joke y_w and a low-ranking candidate y_l for the same prompt, selected based on their Elo gap. We optimize the policy π_θ directly without a reward model:

$$\mathcal{L}_{DPO}(\pi_\theta; \pi_{ref}) = -E_{(x,y_w,y_l)\sim\mathcal{D}_{DPO}}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)}\right)\right] \quad (1)$$

4.3 Offline Group Relative Policy Optimization (O-GRPO)

Beyond pairwise preference alignment, we experiment with an offline variant of Group Relative Policy Optimization, **O-GRPO**, which adapts GRPO (Shao et al., 2024) to a fixed preference dataset. Rather than sampling candidates and computing rewards online during training, O-GRPO operates on the pre-constructed candidate pool: all 24 candidates

per prompt are ranked once via a complete Bradley-Terry tournament, and group-relative advantages are derived from those static scores. Specifically, for each joke y_i, we compute its advantage as

$$A_i = \frac{r_i - \mu_{\text{group}}}{\sigma_{\text{group}} + \epsilon} \quad (2)$$

where r_i is its pre-computed ranking score. We then apply exponentially weighted SFT, where each candidate contributes proportionally to the weights defined in Appendix F. This up-weights relatively stronger responses and down-weights weaker ones using fixed offline rankings, without live reward computation.

4.4 Cognitive Synergy Distillation (CSD)

In the SFT stage of the training pipeline, the teacher's persona-specific reasoning traces are not used, and the student model observes only the final jokes. In contrast, CSD trains the student on both the reasoning traces and the corresponding joke:
`<think>` *persona-specific brainstorming* `</think>` *joke*

This is process distillation: the student learns not just *what* to generate but *how* the teacher planned it. For DPO, both chosen and rejected responses include reasoning traces (symmetric format), so the model cannot shortcut by learning that the mere presence of reasoning correlates with winning; it must learn which *content* leads to better jokes.

At inference, the model generates reasoning followed by the joke. The reasoning is stripped for evaluation (ensuring fair comparison with non-CSD models) but retained for interpretability. Unlike generic CoT, ineffective for humor (Zhong et al., 2024; Tikhonov and Shtykovskiy, 2024), CSD's reasoning is grounded in specific humor theories through the cognitive personas, making it a form of *theory-grounded creative distillation*.

5 Experimental Setup

5.1 Datasets and Data Synthesis

We utilize the official SemEval 2026 Task 1 (MWAHAHA) experimental set (Castro et al., 2026), comprising 1,200 news headlines and word-pair prompts as inputs to our generation pipeline. Using the Cognitive Synergy Framework, we generate 24 candidates per prompt (4 per persona × 6 personas) from a teacher ensemble of *Kimi-K2* and *Qwen 2.5-32B-Instruct*, yielding a raw pool of ∼28,800 candidates. These candidates are scored and ranked via a pairwise LLM evaluation system using Llama 3.3-70B-Instruct as the judge, producing per-prompt Elo ratings for all 24 candidates. We construct three training subsets from these rankings:

- SFT Data ($\mathcal{D}_{SFT}$, $N = 12,000$): For each of the 1,200 prompts, we select the top 10 Elo-ranked candidates (rather than only the single best). Using multiple top-ranked candidates per prompt avoids mode collapse: the student learns a diverse range of humor styles (e.g., wordplay, absurdity, sarcasm) instead of collapsing toward one dominant style.

- DPO Data ($\mathcal{D}_{DPO}$, $N = 6,000$): For each prompt, we construct 5 preference pairs by randomly pairing candidates from the top-5 Elo-ranked jokes (chosen, y_w) with candidates from the bottom-5 Elo-ranked jokes (rejected, y_l). This yields 5 pairs × 1,200 prompts = 6,000 preference pairs, with a sharp quality gap between chosen and rejected responses.

- O-GRPO Data ($\mathcal{D}_{GRPO}$): We use all 24 candidates per prompt across the 1,200 prompts, computing normalized group-relative Elo advantages per group ($G = 24$). This exposes the model to the full quality spectrum within each prompt group.

Appendix D reports the persona composition of each alignment corpus, confirming that all six cognitive personas contribute to SFT, DPO, and O-GRPO training data.

For evaluation, we use two benchmarks. (1) **SemEval-MWAHAHA**: from the official 300-prompt test set, we evaluate on the first 50 headlines for controlled leaderboard comparison. (2) **Humor Transfer Bench (HTB)**: our new benchmark with 400 prompts (8 domains × 50 prompts) designed to test transfer beyond headline-style inputs.

5.2 Baselines

We evaluate 15 total models: frontier systems (GPT-5, Kimi-K2, Gemini-2.5-Pro), open-weight general models (Qwen3-32B, GPT-OSS-120B, Base Qwen-7B), HumorGen variants (SFT, DPO, O-GRPO and their Think versions, plus HumorGen-Com-7B), and humor-specialized baselines (phi2-Humor (Abbas, 2026), JokeGPT (TzJ2006, 2026)).

5.3 Implementation Details

All models were trained on NVIDIA H100 (80GB) GPUs using LoRA (Hu et al., 2022) (r=16) with the Unsloth library. SFT ran for 3 epochs; DPO and O-GRPO for 5 epochs, both with early stopping (patience=2). Candidate ranking for the full pool consumed ∼132 H100 node-hours. For O-GRPO, groups of $G = 24$ candidates per prompt maximize the advantage-weighted learning signal.

5.4 Evaluation Protocols

Ranking methodology: For each pair of jokes (A, B), the LLM judge selects the funnier one (or tie), with presentation order randomized per match to mitigate position bias. We aggregate all outcomes into a contest matrix, fit a Bradley-Terry (BT) model (Gao et al., 2025; Ajayi and Mitra, 2026b) via the MM algorithm to estimate latent ratings, and report Elo-scale ratings with 95% bootstrap confidence intervals (100 samples). For cross-judge Kendall comparisons, exact BT ties are broken deterministically by model-name order before rank correlation is computed; key model comparisons have non-overlapping confidence intervals and are statistically significant. We do not report reference-based metrics (e.g., BLEU, ROUGE) because humor generation is inherently open-ended and subjective, motivating pairwise preference-based evaluation.

1. Automated Pairwise Evaluation: We report Llama-3.3-70B and Qwen 2.5 72B judged results for both benchmarks in the main paper: SemEval leaderboard (15 models, 5,250 pairwise comparisons) and HTB full benchmark (400 prompts, 15 models, 42,000 pairwise comparisons).

2. Human Validation: Human evaluators judge 60 curated pairwise comparisons across 12 ablation categories; they are blinded to model identity and presentation order is randomized to mitigate position bias. We report inter-annotator agreement, LLM–human consensus, and correlation with automated BT ratings.

5.5 Humor Transfer Bench (HTB)

A key challenge in evaluating textual humor generation models is the absence of diverse benchmark datasets. The only existing benchmark for this task is the SemEval 2026 MWAHAHA test set, which uses news headlines as prompts, a single, narrow input distribution that limits assessment of model generalizability. To address this gap, we introduce the **Humor Transfer Bench (HTB)**, a new benchmark of 400 prompts spanning eight structurally distinct input domains (50 each): *Neutral Facts*, *Everyday Life*, *Abstract Concepts*, *Dialogic Quotations*, *Scenario Inputs*, *Analogical Prompts*, *Direct Instructional*, and *News Headlines*. No model is fine-tuned on HTB data; any performance gain on out-of-distribution domains therefore reflects genuine humor transfer. Full domain descriptions and design rationale are in Appendix B.

6 Results

6.1 Model Performance

We evaluate all 15 models on the Humor Transfer Bench under the Llama 3.3-70B judge (42,000 pairwise comparisons). Figure 3 shows the Bradley-Terry leaderboard; Figure 4 shows the head-to-head win-rate heatmap across all model pairs. **HumorGen-SFT-7B** ranks 3rd (BT 1128.14) and **HumorGen-DPO-7B** ranks 4th (BT 1123.72), ahead of Gemini-2.5-Pro (BT 1059.07), GPT-OSS-120B (BT 1048.19), and Qwen3-32B (BT 990.44), all of which are 4–18× larger. Humor-finetuned baselines phi2-Humor and JokeGPT rank 12th and 15th, confirming that domain specialization alone, without cognitive-driven data curation, does not generalize. To verify rating stability across evaluators, we run Qwen 2.5-72B as a second independent judge; the cross-judge Kendall $\tau = 0.8667$ ($p = 1.54 \times 10^{-7}$) confirms that model ordering is robust across judges. We additionally evaluate both judges on the SemEval news-headline test set (15 models, 50 headlines), where cross-judge agreement

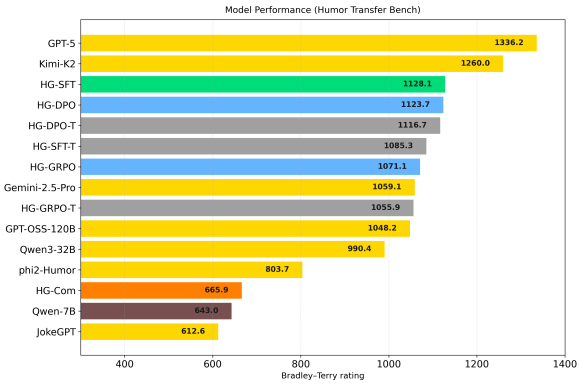

Figure 3: HTB BT leaderboard (Llama judge, 400 prompts, 15 models). HumorGen-SFT and DPO-7B rank 3rd/4th, outperforming models 4–18× their size.

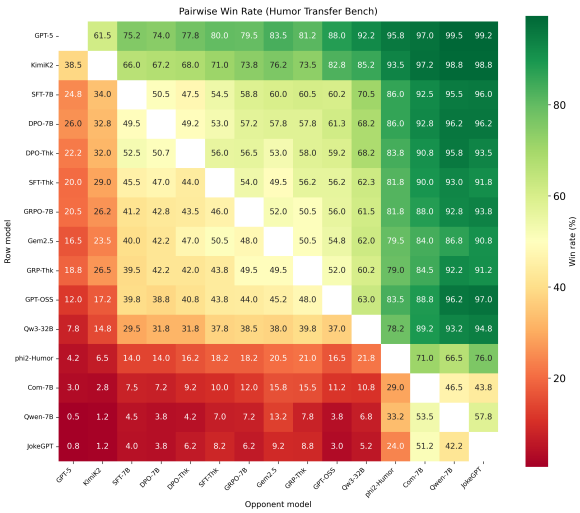

Figure 4: HTB pairwise win-rate heatmap (Llama judge). Each cell shows the row model's win rate against the column model.

reaches Kendall $\tau = 0.8286$ ($p = 9.95 \times 10^{-7}$). Full leaderboard tables with 95% CI and all Qwen-judge results are reported in Appendix A. Overall, HumorGen 7B variants consistently outperform significantly larger models across unseen prompt domains, validating the generalization capacity of the Cognitive Synergy Framework.

6.2 Preference Alignment

We investigate whether preference alignment (DPO, O-GRPO) improves over SFT. Across both benchmarks and judges, SFT and DPO perform similarly, with negligible and inconsistent differences: on HTB (Llama judge), SFT (BT 1128.14) slightly outperforms DPO (BT 1123.72), while under the Qwen judge, DPO (BT 1138.05) narrowly exceeds SFT (BT 1132.41). In both cases, the 95% confidence intervals overlap, indicating no statistically significant difference. O-GRPO consistently underperforms both SFT and DPO (e.g., BT 1071.13 on HTB under the Llama judge, rank 7th). This gap is attributable to a skewed advantage distribution in the O-GRPO training data (median advantage $= -0.479$), where a small number of high-quality samples dominate the group mean, causing most candidates to receive negative relative advantages. As a result, the optimization signal primarily penalizes weaker samples rather than reinforcing strong ones, leading to stable underperformance. These findings suggest a *data quality ceiling*: when SFT data is already well-curated via CSF, additional preference optimization yields limited gains in this subjective humor generation setting.

6.3 CSD and the Explainer Trap

The "explainer trap" emerges when we train the 7B HumorGen variants to *think* by applying Cognitive Synergy Distillation (CSD) so the student is trained on the teacher's `<think>` reasoning traces alongside the joke (see § Methodology). On automatic leaderboards, Think variants *usually* rank below their matched non-thinking counterparts (e.g., SFT-Think 6th vs. SFT 3rd on HTB under Llama; DPO-Think 5th vs. DPO 4th), but the effect is not uniform: DPO-Think ranks 3rd under the Qwen judge on HTB, ahead of DPO (4th). Distilling reasoning traces appears to bias outputs toward *explaining* the joke rather than delivering it (Appendix J), though we did not test whether teacher models themselves over-explain. This extends prior work showing limited benefit of CoT for humor (Zhong et al., 2024; Tikhonov and Shtykovskiy, 2024): reasoning-augmented training does not reliably improve judged funniness in our setup.

6.4 Comedian Adaptation

To further probe the data quality thesis, we fine-tuned HumorGen-SFT on 998 stand-up jokes (Shaun Eli (Eli, 2026)). Results show that performance regressed sharply (BT: 1083.9 → 653.1), as *performance-native* stand-up is optimized for live delivery rather than written punchlines, confirming that data origin matters as much as data quantity. Full analysis of the comedian adaption experiment is provided in

Appendix N.

6.5 Human Evaluation

Three evaluators gave 170 blind pairwise judgments on 60 curated pairs (12 categories, 5 each) over 50 held-out headlines. Inter-annotator agreement was 31.7% (one-third of pairs), reflecting humor's subjectivity. To formalize this, we computed Krippendorff's alpha for the multi-annotator subset, achieving $\alpha = 0.425$ across three evaluators and $\alpha = 0.412$ across two, which is consistent with the highly subjective nature of humor evaluation. The LLM judge matched human consensus on 58.3% of pairs (Gold) and individual votes at 52.4% (Micro-Avg). In this "Good vs. Good" regime (high-quality outputs, no objectively worse option), 58.3% indicates the judge captures shared preferences well above chance. More details of the human evaluation experiments are provided in Appendix K.

7 Analysis

7.1 What Makes Jokes Win?

For each HTB pairwise comparison, the Llama judge assigns structured tags to the winner (humor *mechanisms* and *delivery*) and to the loser (flaw tags). Table 2 aggregates all canonical tags across 41,976 non-tie comparisons from the full HTB Llama evaluation. Percentages denote the share of comparisons in which a tag was assigned (multiple tags per comparison are allowed).

Incongruity and absurdity dominate winning mechanisms (>84%); punchline positioning (96.1%) and conciseness (72.0%) dominate delivery. Among losers, weak punchlines (69.0%) and cliché (64.9%) are most frequent; overexplained (22.3%) and buried punchlines (12.9%) align with the explainer trap (see section 6.3). Two recurring failure patterns emerge: (1) *generic punchlines* defaulting to safe, high-probability completions, and (2) *overextended setups* burying the joke. See Appendix P for qualitative examples.

Feature	**%**
Humor Mechanisms *(winner)*	
Incongruity	85.2
Absurdity	84.9
Surprise	57.2
Wordplay	42.0
Irony	22.5
Sarcasm	17.3
Narrative	16.5
Observational	14.7
Delivery Features *(winner)*	
Punchline positioning	96.1
Conciseness	72.0
Framing commitment	57.7
Escalation	33.0
Deadpan	11.7
Timing	4.9
Loser Features *(loser)*	
Weak punchline	69.0
Cliché	64.9
Overexplained	22.3
Buried punchline	12.9
Confusing	8.8
Offensive	0.5

Table 2: Judge-assigned tags on HTB (Llama; 41,976 non-tie comparisons). % of comparisons tagged.

8 Conclusion

We introduce the Cognitive Synergy Framework, which operationalizes psychological humor theories into six cognitive personas to generate diverse, high-quality humor data via Mixture-of-Thought. HumorGen model variants achieve strong performance among open-weight models and are competitive with frontier systems, outperforming significantly larger models, suggesting that targeted CSF data curation matters more than model scale for humor generation. Our central finding is a *data quality ceiling*: when SFT data is diverse and well-curated, preference optimization (DPO, O-GRPO) yields no significant gains. We also observe that reasoning-augmented training can reduce judged funniness, while CSF-curated synthetic data outperforms human-written jokes when used in training. Human evaluation further supports that the LLM judge captures subtle preference differences even in subjective "Good vs. Good" pairwise comparisons among high-quality jokes. Overall, our results indicate that in modeling humor generation, performance gains are driven less by optimization complexity and more by the quality and structure of the underlying training data.

9 Limitations

This study defines a focused empirical scope that future work should broaden. Our analysis is limited to English, text-only humor and two

benchmarks (SemEval-MWAHAHA and Humor Transfer Bench), motivating evaluation in additional languages, domains, and multimodal settings. Human evaluation is constrained by the inherent subjectivity of humor and the presence of multiple valid outputs per prompt; we therefore use a small annotator pool ($N = 3$) over 60 sampled pairwise comparisons, and larger evaluations would provide higher statistical coverage. Model comparisons are restricted to 15 selected models (including models from Qwen, Llama, GPT-model families, and specialized humor models) due to compute constraints; however, this can be expanded to examine more models in the future. Finally, given the subjective nature of humor, there is a tendency for models to produce jokes that are offensive (see Appendix O) to some listeners or be culturally insensitive.

10 Ethics Statement

Humor generation risks producing offensive content. Our framework encourages creative mechanisms (e.g., wordplay, absurdity) over denigration or prejudice. Training data derives from the SEMEVAL 2026 MWAHAHA headlines and our synthetic candidates generated from those prompts; evaluation additionally uses HTB prompts for transfer assessment. Human evaluators were volunteers recruited by invitation; no payment was provided. All content in this work was written and verified by the authors. We used an LLM-based tool solely for grammar correction, language polishing, and LaTeX formatting assistance. The authors take full responsibility for the final content.

References

Insa Abbas. 2026. Humor generation model for SemEval 2026. https://huggingface.co/insaabbas/phi2-4-epoch-humor-model. SemEval 2026 humor-generation checkpoint (phi2-4-epoch-humor-model).

Edward Ajayi and Prasenjit Mitra. 2026a. Automatic humor detection: A comprehensive survey from theoretical foundations to large language models. *ResearchGate Preprint*.

Edward Ajayi and Prasenjit Mitra. 2026b. Humorrank: A tournament-based leaderboard for evaluating humor generation in large language models. *arXiv preprint arXiv:2604.19786*.

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.

Yuyan Chen, Zhixu Li, Jiaqing Liang, Yanghua Xiao, Bang Liu, and Yunwen Chen. 2023. Can pre-trained language models understand chinese humor? In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 465–480.

Ryan Rony Dsilva. 2024. Augmenting large language models with humor theory to understand puns. Master's thesis, Purdue University.

Shaun Eli. 2026. Expired comedy (topical humor). https://www.brainchampagne.com/writings/expired-comedy-topical-humor. Accessed: 2026-03-16.

Jacob Fein-Ashley, Dhruv Parikh, Rajgopal Kannan, and Viktor Prasanna. 2025. Mixture of thoughts: Learning to aggregate what experts think, not just what they say. *arXiv preprint arXiv:2509.21164*.

Mingqi Gao, Yixin Liu, Xinyu Hu, Xiaojun Wan, Jonathan Bragg, and Arman Cohan. 2025. Reevaluating automatic llm system ranking for alignment with human preference. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4605–4629.

Soham V Govande, Taeuk Kang, and Andrew Shi. 2025. Teaching models to reason about vision-based code generation using grpo.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Sophie Jentzsch and Kristian Kersting. 2023. Chatgpt is fun, but it is not funny! humor is still challenging large language models. *arXiv preprint arXiv:2306.04563*.

T. Khurana, K. Pillalamarri, V. Pande, and M. Singh. 2024. Lolgorithm: Integrating semantic, syntactic and contextual elements for humor classification. *Preprint*, arXiv:2408.06335.

Sheila Lintott. 2016. Superiority in humor theory. *The Journal of Aesthetics and Art Criticism*, 74(4):347–358.

Xingzhou Lou, Junge Zhang, Jian Xie, Lifeng Liu, Dong Yan, and Kaiqi Huang. 2025. Sequential preference optimization: Multi-dimensional preference alignment with implicit reward modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27509–27517.

A Peter McGraw and Caleb Warren. 2010. Benign violations: Making immoral behavior funny. *Psychological science*, 21(8):1141–1149.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Greg Robison. 2024. The last laugh: Exploring the role of humor as a benchmark for large language models. Accessed: 2026-03-15.

Tabea Scheel. 2025. Definitions, theories, and measurement of humor. In *Humor at work in teams, leadership, negotiations, learning, and health*, pages 11–37. Springer.

Mohammadamin Shafiei and Hamidreza Saffari. 2025. Not All Jokes Land: Evaluating Large Language Models Understanding of Workplace Humor. *arXiv preprint*. ArXiv:2506.01819 [cs].

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Alexey Tikhonov and Pavel Shtykovskiy. 2024. Humor Mechanics: Advancing Humor Generation with Multistep Reasoning. *arXiv preprint*. ArXiv:2405.07280 [cs].

Chengzhuo Tong, Ziyu Guo, Renrui Zhang, Wenyu Shan, Xinyu Wei, Zhenghao Xing, Hongsheng Li, and Pheng-Ann Heng. 2025. Delving into rl for image generation with cot: A study on dpo vs. grpo. *arXiv preprint arXiv:2505.17017*.

TzJ2006. 2026. JokeGPT-Model. PPO-tuned LoRA adapter on Qwen3-8B.

Dmitry Vikhorev, Daria Galimzianova, Svetlana Gorovaia, Elizaveta Zhemchuzhina, and Ivan P Yamshchikov. 2024. Cleancomedy: Creating friendly humor through generative techniques. *arXiv preprint arXiv:2412.09203*.

Han Wang, Yilin Zhao, Dian Li, Xiaohan Wang, Gang Liu, Xuguang Lan, and Hui Wang. 2025. Innovative Thinking, Infinite Humor: Humor Research of Large Language Models through Structured Thought Leaps. *arXiv preprint*. ArXiv:2410.10370 [cs].

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yusuke Yasuda and Tomoki Toda. 2025. Automatic design optimization of preference-based subjective evaluation with online learning in crowdsourcing environment. *Computer Speech & Language*, page 101888.

Hang Zhang, Dayiheng Liu, Jiancheng Lv, and Cheng Luo. 2020. Let's be humorous: Knowledge enhanced humor generation. *arXiv preprint arXiv:2004.13317*.

Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. 2024. Let's Think Outside the Box: Exploring Leap-of-Thought in Large Language Models with Creative Humor Generation. pages 13246–13257.

HumorGen: Cognitive Synergy for Humor Generation in Large Language Models via Persona-Based Distillation

Appendix - Table of Contents

A Benchmark Leaderboards 13
A.2 HTB Full Results (Llama 3.3-70B Judge) 13
A.3 HTB Full Results (Qwen 2.5-72B Judge) 13
A.4 SemEval Full Results (Llama 3.3-70B Judge) 14
A.5 SemEval Full Results (Qwen 2.5-72B Judge) 14
B HTB Design and Domain Descriptions 16
C Per-Persona Analysis 17
D Training Dataset Persona Distribution 17
E Psychological Theories of Humor .. 18
G Training Details and Hyperparameters 19
F O-GRPO Objective 18
G.1 Hyperparameter Configurations 19
G.2 Training Dynamics and Results 19
G.3 Evaluation Loss Trajectories 19
G.4 Comedian Adaptation Hyperparameters 20
H Full Persona Prompts 20
I Immersive Persona Comparison 21
J Think vs. Non-Think 21
K Human Evaluation Details 22
L Evaluation UI 23
N Comedian Adaptation Analysis ... 24
O Qualitative Output Examples: Out-of-Distribution Headlines 25
P Failure Mode Examples 26

A Benchmark Leaderboards

Full leaderboard tables (BT ratings with 95% bootstrap confidence intervals) and plots for all four evaluation conditions: HTB and SemEval, under both the Llama 3.3-70B and Qwen 2.5-72B judges.

A.1 Bootstrap and CI Estimation Details

For each evaluation setting (HTB/SemEval × Llama/Qwen), we build a complete model–model outcome matrix from pairwise comparisons and fit Bradley–Terry ratings via the MM algorithm. We estimate uncertainty with non-parametric bootstrap resampling ($B = 100$): each bootstrap replicate re-samples the evaluation units with replacement, reconstructs the full contest matrix, and re-fits the BT model. We report 95% confidence intervals as percentile bounds (2.5th, 97.5th) over the bootstrap rating distribution for each model. This procedure is applied consistently to all leaderboard tables in this appendix.

A.2 HTB Full Results (Llama 3.3-70B Judge)

Table 3 reports full BT ratings with 95% bootstrap confidence intervals for all 15 models on HTB under the Llama 3.3-70B judge. Figure 5 visualises the leaderboard; Figure 6 shows the head-to-head win-rate matrix. HumorGen-SFT-7B and HumorGen-DPO-7B rank 3rd and 4th with non-overlapping CIs from all models below rank 6, confirming statistically significant separation from Gemini-2.5-Pro, GPT-OSS-120B, and all smaller baselines.

Rank	Model	BT Rating	95% CI	Wins / Losses
1	GPT-5	1336.18	[1323.3, 1348.3]	4738 / 862
2	Kimi-K2	1259.98	[1249.7, 1268.5]	4362 / 1237
3	**HumorGen SFT-7B**	**1128.14**	[1118.3, 1138.1]	3565 / 2035
4	**HumorGen DPO-7B**	**1123.72**	[1115.7, 1134.9]	3536 / 2064
5	HumorGen DPO-Think-7B	1116.65	[1107.9, 1127.1]	3489 / 2110
6	HumorGen SFT-Think-7B	1085.31	[1075.8, 1096.5]	3281 / 2319
7	HumorGen GRPO-7B	1071.13	[1060.8, 1080.1]	3186 / 2414
8	Gemini-2.5-Pro	1059.07	[1049.3, 1068.4]	3104 / 2494
9	HumorGen GRPO-Think-7B	1055.94	[1043.8, 1066.8]	3084 / 2516
10	GPT-OSS-120B	1048.19	[1039.7, 1057.1]	3032 / 2568
11	Qwen3-32B	990.44	[981.4, 999.4]	2648 / 2952
12	phi2-Humor	803.72	[794.5, 818.2]	1539 / 4060
13	HumorGen-Com-7B	665.93	[645.5, 680.0]	897 / 4682
14	Base Qwen-7B	643.01	[628.3, 658.0]	818 / 4781
15	JokeGPT	612.58	[597.6, 627.4]	697 / 4882

Table 3: HTB leaderboard, Llama judge (400 prompts, 42,000 comparisons).

A.3 HTB Full Results (Qwen 2.5-72B Judge)

Table 4 and Figures 7–8 report HTB results under the Qwen 2.5-72B judge. The rank ordering is highly consistent with the Llama judge (Kendall $\tau = 0.8667$, $p = 1.54 \times 10^{-7}$), confirming that the leaderboard is robust to judge choice. Notable difference: HumorGen-DPO-Think-7B rises to 3rd under Qwen, suggesting the Qwen judge is slightly more receptive to reasoning-augmented outputs.

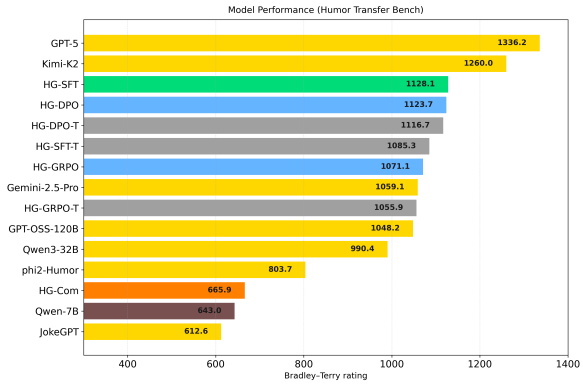

Figure 5: HTB BT leaderboard, Llama judge (400 prompts, 15 models).

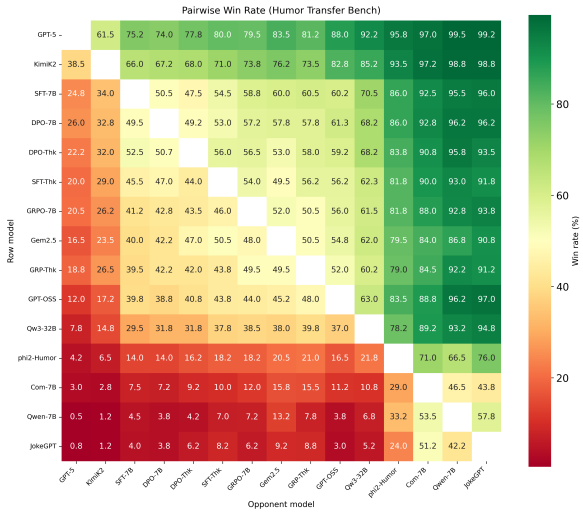

Figure 6: HTB win-rate heatmap, Llama judge (row beats column %).

Rank	Model	BT Rating	95% CI	Wins / Losses
1	GPT-5	1268.25	[1256.5, 1281.1]	4418 / 1181
2	Kimi-K2	1228.24	[1217.3, 1239.4]	4192 / 1408
3	HumorGen DPO-Think-7B	1147.52	[1136.8, 1158.8]	3686 / 1914
4	**HumorGen DPO-7B**	**1138.05**	[1130.3, 1147.2]	3623 / 1976
5	**HumorGen SFT-7B**	**1132.41**	[1121.0, 1141.2]	3586 / 2014
6	HumorGen SFT-Think-7B	1105.57	[1094.0, 1115.2]	3406 / 2194
7	HumorGen GRPO-Think-7B	1084.61	[1075.2, 1097.0]	3264 / 2336
8	HumorGen GRPO-7B	1080.19	[1070.3, 1090.1]	3234 / 2366
9	GPT-OSS-120B	1055.91	[1046.4, 1065.5]	3069 / 2531
10	Gemini-2.5-Pro	1035.10	[1025.1, 1046.1]	2926 / 2670
11	Qwen3-32B	1000.83	[991.1, 1011.4]	2698 / 2901
12	phi2-Humor	758.74	[746.6, 769.0]	1300 / 4300
13	Base Qwen-7B	702.60	[689.6, 715.1]	1046 / 4554
14	HumorGen-Com-7B	682.92	[666.9, 692.9]	951 / 4624
15	JokeGPT	579.04	[563.9, 596.0]	573 / 5003

Table 4: HTB leaderboard, Qwen judge (400 prompts, 42,000 comparisons). Kendall $\tau = 0.8667$ vs. Llama judge.

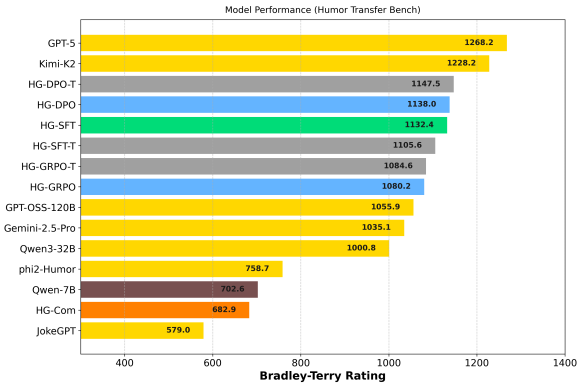

Figure 7: HTB BT leaderboard, Qwen judge (400 prompts, 15 models).

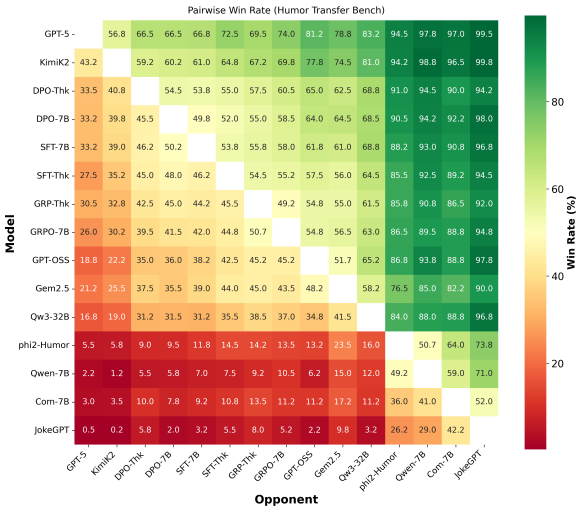

Figure 8: HTB win-rate heatmap, Qwen judge (row beats column %).

A.4 SemEval Full Results (Llama 3.3-70B Judge)

Table 5 and Figures 9–10 report results on the SemEval 2026 MWAHAHA test set (first 50 headlines, 15 models, 5,250 pairwise comparisons) under the Llama 3.3-70B judge. HumorGen-SFT-7B and HumorGen-DPO-7B rank 4th and 5th, behind the three frontier systems (GPT-5, Kimi-K2, Gemini-2.5-Pro) and ahead of all open-weight general models. The gap between SFT and DPO (4.88 BT points) is within overlapping confidence intervals, confirming no statistically meaningful difference.

A.5 SemEval Full Results (Qwen 2.5-72B Judge)

Table 6 and Figures 11–12 report SemEval results under the Qwen 2.5-72B judge. Rank ordering is consistent with the Llama judge at Kendall $\tau = 0.8286$ ($p = 9.95 \times 10^{-7}$). Under Qwen, HumorGen-DPO-7B rises to 3rd (BT 1202.76) and HumorGen-SFT-7B to 4th (BT 1193.81), both ahead of Gemini-2.5-Pro, consistent with the HTB Qwen pattern of the Qwen judge rating DPO-aligned outputs slightly higher.

Rank	Model	BT Rating	95% CI	Wins / Losses
1	GPT-5	1378.73	[1346.1, 1421.5]	605 / 94
2	Kimi-K2	1279.63	[1245.0, 1322.1]	549 / 151
3	Gemini-2.5-Pro	1247.80	[1212.4, 1279.7]	528 / 172
4	**HumorGen SFT-7B**	**1140.37**	[1107.9, 1173.2]	449 / 251
5	**HumorGen DPO-7B**	**1135.25**	[1101.9, 1160.2]	445 / 255
6	HumorGen GRPO-7B	1089.84	[1060.7, 1114.1]	409 / 291
7	GPT-OSS-120B	1049.99	[1019.8, 1081.5]	377 / 323
8	HumorGen SFT-Think-7B	1049.99	[1016.3, 1084.5]	377 / 323
9	HumorGen DPO-Think-7B	1031.30	[1002.1, 1058.0]	362 / 338
10	Qwen3-32B	1023.18	[997.0, 1046.0]	355 / 344
11	HumorGen GRPO-Think-7B	948.51	[914.7, 982.6]	297 / 403
12	phi2-Humor	791.32	[751.8, 826.1]	188 / 512
13	HumorGen-Com-7B	721.97	[682.3, 750.8]	148 / 552
14	Base Qwen-7B	673.16	[643.1, 718.0]	123 / 577
15	JokeGPT	438.97	[384.3, 500.2]	37 / 663

Table 5: SemEval leaderboard, Llama judge (50 headlines, 15 models, 5,250 comparisons).

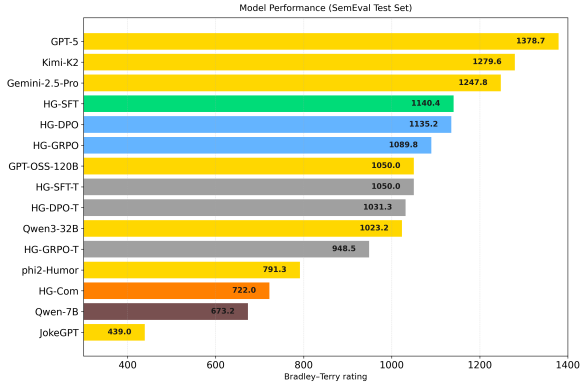

Figure 9: SemEval BT leaderboard, Llama judge (50 headlines, 15 models).

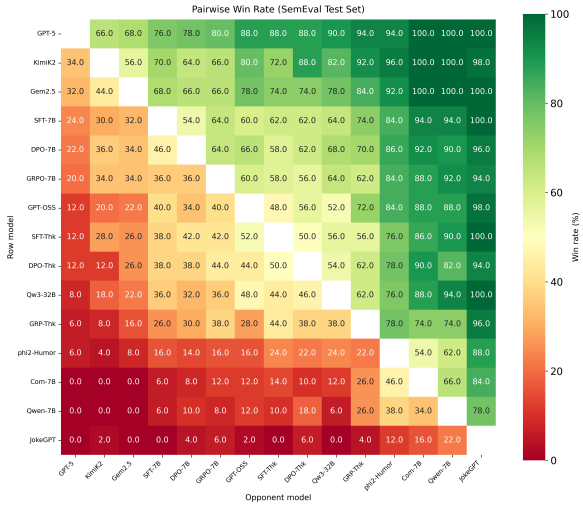

Figure 10: SemEval win-rate heatmap, Llama judge (row beats column %).

Cross-benchmark observations. Three consistent patterns emerge across all four evaluation conditions (HTB and SemEval,

Rank	Model	BT Rating	95% CI
1	GPT-5	1285.15	[1258.2, 1320.8]
2	Kimi-K2	1210.51	[1188.5, 1233.9]
3	**HumorGen DPO-7B**	**1202.76**	[1175.5, 1227.0]
4	**HumorGen SFT-7B**	**1193.81**	[1162.0, 1228.2]
5	Gemini-2.5-Pro	1144.23	[1114.3, 1171.8]
6	HumorGen DPO-Think-7B	1132.13	[1103.0, 1170.3]
7	HumorGen GRPO-7B	1128.51	[1106.3, 1159.4]
8	HumorGen SFT-Think-7B	1063.88	[1034.9, 1091.8]
9	GPT-OSS-120B	1038.60	[1009.1, 1065.9]
10	HumorGen GRPO-Think-7B	1028.90	[997.6, 1053.3]
11	Qwen3-32B	996.92	[968.0, 1030.5]
12	HumorGen-Com-7B	832.69	[802.3, 862.2]
13	Base Qwen-7B	713.39	[674.7, 745.2]
14	phi2-Humor	698.89	[657.3, 739.2]
15	JokeGPT	329.63	[237.7, 385.6]

Table 6: SemEval leaderboard, Qwen judge (50 headlines, 15 models). Kendall $\tau = 0.8286$ vs. Llama judge.

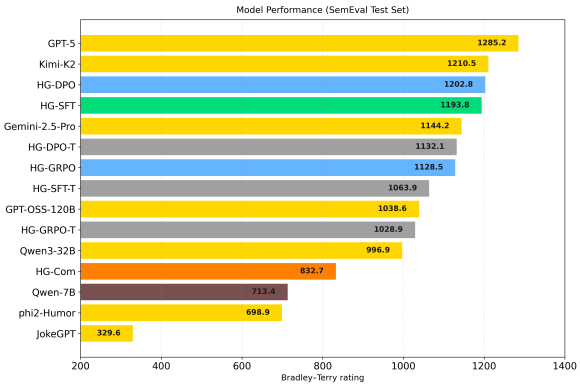

Figure 11: SemEval BT leaderboard, Qwen judge (50 headlines, 15 models).

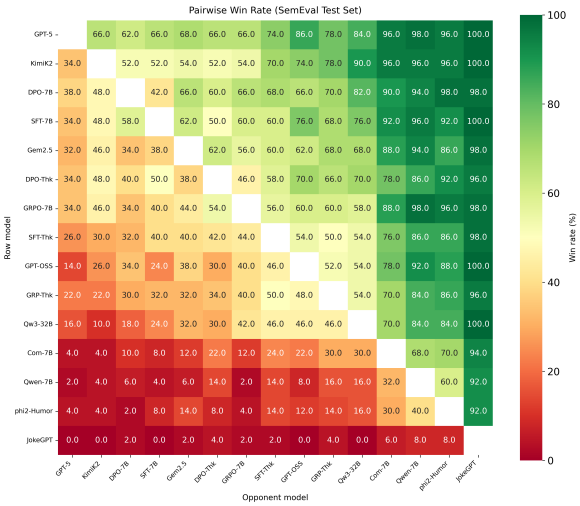

Figure 12: SemEval win-rate heatmap, Qwen judge (row beats column %).

Llama and Qwen judges). First, **HumorGen-SFT-7B and HumorGen-DPO-7B cluster tightly at ranks 3–5**, with BT differences between them within confidence interval overlap in every condition, consistent with the data quality ceiling finding in the main paper. Second, **Think variants most often trail their matched non-thinking counterparts** on automatic rankings (e.g., all three pairs on HTB–Llama and SemEval–Llama), though DPO-Think outranks DPO under the Qwen judge on HTB. Third, **humor-specialized baselines phi2-Humor and JokeGPT consistently rank 12th–15th** across both benchmarks and both judges, confirming that fine-tuning on humor data alone, without cognitive-driven curation, does not yield competitive performance. The high cross-judge agreement ($\tau \geq 0.83$ on both benchmarks) indicates these orderings are stable and not an artifact of judge choice.

B Humor Transfer Bench (HTB) Design and Domain Descriptions

The Humor Transfer Bench (HTB) tests whether humor generation models generalize beyond the single prompt distribution present in existing benchmarks. HTB holds the task constant (*Generate a funny joke given this prompt*) while varying the input domain. Each domain is defined by a distinct syntactic and pragmatic structure, making domain membership an observable property rather than a subjective label. The eight domains are described in Table 7.

Dom.	Name	Status	Description and Example
A	Neutral Facts	OOD	Surprising declarative facts. *"Bananas are technically berries."*
B	Everyday Life	OOD	First-person observational statements. *"My printer hates me."*
C	Abstract Concepts	OOD	Bare noun-phrase prompts. *"Bureaucratic nostalgia."*
D	Dialogic Quotations	OOD	Attributed utterances. *"The fridge said: I'm judging you."*
E	Scenario Inputs	OOD	Situational irony setups. *"A vampire is working in HR."*
F	Analogical Prompts	OOD	Cross-domain relational mappings. *"Life is a subscription service you never read the terms for."*
G	Direct Instructional	Near-distrib.	Directive format closest to training. *"Make a joke about X."*
H	News Headlines	Reference News	50 real-world BBC headlines.

Table 7: HTB domain descriptions used in evaluation.

B.1 Humor Transfer Bench Data Card

HTB is structured as eight equal-size prompt categories so that models can be compared consistently across varied humor-input styles.

Field	Value
Dataset name/version	Humor Transfer Bench (HTB), v1
Primary objective	Measure humor transfer under prompt-domain shift while holding the generation task fixed
Task/instruction	Text-to-text humor generation (*Generate a funny joke given this prompt*)
File format/encoding	TSV, UTF-8, Unix line endings
Schema (columns)	`id`, `domain`, `prompt`
Identifier format	`HTB_{DOMAIN}_{NNN}` (e.g., `HTB_A_001`)
Total size	400 prompts (8 domains × 50 prompts)
Domain inventory	A: Neutral Facts (50); B: Everyday Life (50); C: Abstract Concepts (50); D: Dialogic Quotations (50); E: Scenario Inputs (50); F: Analogical Prompts (50); G: Direct Instructional Prompts (50); H: News Headlines (50)
Domain definition	Each category is defined by observable prompt structure; detailed examples are listed in Table 7
Data source	Domains A–G are HTB-authored prompts; Domain H uses 50 real-world BBC headlines
License/provenance	HTB-authored prompts are released under CC BY 4.0; Domain H uses publicly available news headline text
Exclusions	No canonical joke templates (e.g., knock-knock / walks-into-a-bar / chicken-road) in prompts
Training usage	HTB is evaluation-only in this paper (no model is fine-tuned on HTB prompts)
Evaluation usage	15 models, pairwise judging under Llama-3.3-70B and Qwen-2.5-72B

Table 8: HTB data card aligned with the benchmark design specification.

C Per-Persona Analysis

A key design property of the Cognitive Synergy Framework is that *no single persona is hand-picked as the "best"*. All six personas generate candidates, the judge ranks them, and the Elo-based curation process decides which jokes enter training. Table 9 reports three independent measures of curation-stage performance, all derived from the HumorRank pairwise tournament over 28,789 candidates across 1,200 prompts.

- **Pairwise win rate:** proportion of individual 1v1 match-ups a persona's jokes won (source: full match history).
- **Mean rank (of 24):** average within-prompt rank across all 1,200 prompts (rank 1 = best in group; source: per-prompt rankings).
- **Top-3 rate:** proportion of prompts in which the persona produced a top-3 ranked joke.

All three metrics tell the same story and reinforce each other: **Neurotic** leads across every measure, while **Wordsmith** consistently trails. Crucially, all six personas remain competitive across different prompts. Wordsmith, despite ranking lowest, still achieves a top-3 placement on 3.1% of prompts, and all personas contribute to the final training corpora (see Table 10).

Persona	Pairwise win rate	Mean rank (of 24)	Top-3 rate
Neurotic	63.4%	9.3	26.6%
Absurdist	55.8%	11.3	17.6%
Cynic	55.2%	11.4	15.9%
Observer	49.1%	12.5	6.2%
Optimist	41.1%	14.4	5.7%
Wordsmith	34.9%	16.1	3.1%

Table 9: Per-persona curation-stage performance across three independent metrics. *Pairwise win rate*: fraction of 1v1 judge comparisons won. *Mean rank*: average within-prompt rank out of 24 candidates (lower = better). *Top-3 rate*: fraction of prompts where the persona placed in the top 3. All three metrics converge on the same ordering.

The ordering reflects the humor mechanisms each persona draws on. Neurotic and Absurdist generate high incongruity and surprise through escalating anxiety and surreal logic respectively, consistent with the dominant winning tags in Table 2. Cynic relies on satire and superiority framing, which scores similarly to Absurdist. Observer occupies the boundary just below 50%, producing reliably solid jokes that rarely dominate. Optimist and Wordsmith trail because wholesome reframings lack edge, and forced wordplay is penalised by the judge. The consistency of the ordering across all three metrics confirms it is not an artefact of any single evaluation perspective.

D Training Dataset Persona Distribution

We trace each example in the SFT, DPO, and O-GRPO corpora back to its originating cognitive persona in the generation pool. Persona labels are taken from the raw CSF candidate pool and joined to the final alignment files by exact joke-text match (100% coverage on the v4 corpora). Table 10 summarizes the resulting composition.

The raw candidate pool is balanced across all six personas (∼16.7% each). Elo-based selection for SFT (top 10 per prompt) and DPO (top 5 vs. bottom 5) shifts mass toward **Neurotic**, **Absurdist**, and **Cynic** on the preferred side, which directly reflects their higher mean ranks and top-3 rates in Table 9, while **Wordsmith** and **Optimist** are more frequent among rejected DPO examples. O-GRPO retains all 24 candidates per prompt, preserving uniform persona coverage before advantage weighting. Taken together, Tables 9 and 10 show that CSF does *not* select a single dominant persona: the student model is trained on a distribution spanning all six cognitive styles, weighted by judge-assessed quality rather than by design.

Table 10: Persona composition of the alignment training corpora. Counts are followed by row percentages within each column (each column sums to 100%). All six personas appear in every corpus.

Persona	O-GRPO pool *N=28,789*	SFT (top-10) *N=12,000*	DPO chosen *N=6,000*	DPO rejected *N=6,000*
Neurotic	4,800 (16.7%)	2,899 (24.2%)	1,650 (27.5%)	328 (5.5%)
Absurdist	4,796 (16.7%)	2,364 (19.7%)	1,316 (21.9%)	783 (13.1%)
Cynic	4,796 (16.7%)	2,269 (18.9%)	1,257 (20.9%)	679 (11.3%)
Observer	4,798 (16.7%)	2,015 (16.8%)	934 (15.6%)	1,007 (16.8%)
Optimist	4,799 (16.7%)	1,375 (11.5%)	542 (9.0%)	1,334 (22.2%)
Wordsmith	4,800 (16.7%)	1,078 (9.0%)	301 (5.0%)	1,869 (31.1%)
Total	**28,789 (100%)**	**12,000 (100%)**	**6,000 (100%)**	**6,000 (100%)**

E Psychological Theories of Humor

Each cognitive persona in the Cognitive Synergy Framework is grounded in a classical psychological theory of humor. Table 11 summarizes the five theories and their corresponding personas. These theories do not provide a generative recipe for joke construction but characterize the semantic and linguistic structures that make language humorous, which informed our persona design.

Theory	Persona	Summary
Relief Theory	Neurotic	Humor functions as a release mechanism for suppressed tension, anxiety, or socially restricted thoughts. Jokes provide a culturally sanctioned outlet for uncomfortable truths, transforming nervous energy into laughter through cathartic release.
Superiority Theory	Cynic	Dating back to Plato and developed by Hobbes (1650), humor arises from feelings of superiority over others' failures, social hypocrisy, or misfortune. Laughter emerges through recognizing incompetence or pretension, reinforcing the observer's relative standing via critique or sarcasm.
Incongruity Theory	Observer, Absurdist	Kant (1790) and Schopenhauer (1819): humor arises from violated expectations through unexpected juxtapositions or logical impossibilities. The Observer persona reflects everyday social mismatches; the Absurdist extends this to surreal and illogical extremes.
Linguistic Theories	Wordsmith	Raskin's (1985) Semantic Script Theory of Humor and Attardo's extensions: humor arises from language itself via ambiguity, phonological similarity, double meanings, and wordplay. Humor emerges from the gap between literal meaning and intended interpretation.
Benign Violation Theory	Optimist	McGraw and Warren (2010): humor occurs when a norm violation is simultaneously perceived as benign. The Optimist persona operationalizes this through positive reinterpretation of otherwise negative or threatening situations.

Table 11: Psychological humor theories grounding each CSF cognitive persona (Lintott, 2016; Scheel, 2025; McGraw and Warren, 2010).

The empirical win rates in Table 9 are consistent with these theories: Relief (Neurotic, 63.4%) and Incongruity (Absurdist, 55.8%) generate the most consistently preferred jokes, echoing the dominance of incongruity and absurdity in the judge-assigned mechanism tags (Table 2). Linguistic humor (Wordsmith, 34.9%) scores lowest, suggesting that forced wordplay is penalized more than tension-release or expectation-violation mechanisms.

F O-GRPO Objective

O-GRPO is an offline variant of Group Relative Policy Optimization (GRPO) (Shao et al., 2024) adapted to our humor distillation pipeline. Standard GRPO samples fresh completions at each training step and scores them on the fly; for humor, each score requires pairwise LLM judging, which is too expensive to embed in the training loop. O-GRPO instead uses the fixed CSF candidate pool: all G=24 jokes per prompt are ranked once before alignment begins, and training reads from pre-computed scores only.

Training data. For each of the 1,200 SemEval prompts, the MoT ensemble produces 24 candidates (4 per persona $\times$ 6 personas). Llama-3.3-70B judges a complete pairwise tournament per prompt (Figure 2, step B), yielding an Elo rating r_i for each candidate y_i. This gives $\mathcal{D}_{GRPO}$ with 28,800 examples. By contrast, DPO uses five binary pairs per prompt (top-5 vs. bottom-5 Elo jokes; N=6,000). O-GRPO retains the full group so that every candidate is weighted relative to its peers on the same prompt, not only the extremes.

Advantages and loss. Raw Elo ratings are comparable only within a prompt (typical spread $\sim$107 points), so we normalize per group before training. For each prompt x, the advantage of candidate y_i is the group z-score (Equation (2)):

$$A_i = \frac{r_i - \mu_{\text{group}}}{\sigma_{\text{group}} + \epsilon}, \quad \epsilon = 10^{-6}. \tag{3}$$

Values are computed once at data preparation and stored with each example. At training time, advantages map to softmax weights over the group,

$$w_i = \frac{\exp(A_i/T)}{\sum_{j=1}^{G} \exp(A_j/T)}, \tag{4}$$

with temperature T=1.0 (Table 12). The objective is advantage-weighted SFT: each candidate contributes cross-entropy loss $\ell_i(\theta) = -\log \pi_\theta(y_i \mid x)$, computed on assistant-response tokens only,

$$\mathcal{L}_{\text{O-GRPO}}(\theta) = E_x\left[\sum_{i=1}^{G} w_i\, \ell_i(\theta)\right]. \tag{5}$$

Higher-ranked jokes receive larger gradient weight; lower-ranked jokes are down-weighted but still present in the loss. There is no PPO-style clipping and no live reward model during training.

Training and observed dynamics. O-GRPO initializes from the HumorGen-SFT checkpoint and runs for up to 5 epochs with early stopping (patience = 2); full hyperparameters are in Table 12. In practice, the advantage distribution across $\mathcal{D}_{GRPO}$ is strongly right-skewed (median $A = -0.479$): a few high-Elo candidates per prompt pull up μ_{group}, leaving most jokes with negative advantages. The optimizer therefore spends much of its signal suppressing weak outputs rather than reinforcing strong ones, which aligns with O-GRPO's underperformance relative to SFT and DPO (Results: Preference Alignment).

G Training Details and Hyperparameters

This section provides a comprehensive record of the training configurations and experimental results for the HumorGen model suite. All models were fine-tuned using the Qwen 2.5-7B-Instruct base architecture on NVIDIA H100-80GB GPUs.

G.1 Hyperparameter Configurations

Table 12 consolidates the core hyperparameters used across the three major training phases: Supervised Fine-Tuning (SFT), Direct Preference Optimization (DPO), and Group Relative Policy Optimization (GRPO).

Table 12: Consolidated hyperparameters for the HumorGen training pipeline. SFT-Think, DPO-Think, and GRPO-Think used identical settings to their base counterparts.

Parameter	SFT / SFT-Think	DPO / DPO-Think	GRPO / GRPO-Think
Learning Rate	2×10^{-4} (Linear)	5×10^{-7} (Constant)	1×10^{-6} (Constant)
Batch Size (Global)	16	16	16
Epochs (Configured)	3	5	5
Max Sequence Length	1024	1024 (512 prompt)	1024
Optimizer	AdamW (8-bit)	AdamW (8-bit)	AdamW
LoRA Rank (r)	16	16	16
LoRA Alpha	16	16	16
LoRA Modules	All Linear	All Linear	All Linear
Warmup Ratio / Steps	0.03 (ratio)	0.1 (ratio)	10 steps
Weight Decay	0.01	0.0	0.0
Precision	bf16	bf16	bf16
Alignment Specifics	N/A	$\beta = 0.1$	$G = 24, T = 1.0$

G.2 Training Dynamics and Results

Table 13 summarizes the convergence behavior and final metrics for the primary alignment experiments.

Table 13: Training metrics across all HumorGen variants. (*) Training was terminated by early stopping at the best recorded eval loss.

Model Variant	Steps	Final Epoch	Final Loss	Eval Loss (Min)	Runtime
HumorGen-SFT-7B	900	1.26*	1.258	1.342	7.2m
HumorGen-SFT-Think	900	1.26*	1.768	1.908	7.8m
HumorGen-DPO-7B	1,550	4.34*	0.512	0.742	18.5m
HumorGen-DPO-Think	1,550	4.34*	0.528	0.756	21.2m
HumorGen-GRPO-7B	6,050	3.66*	0.456	1.593	23.7m
HumorGen-GRPO-Think	6,850	4.02*	5.901	1.461	6.0h
HumorGen-Com-7B	120	1.83	0.814	1.342	2.3m

G.3 Evaluation Loss Trajectories

Tables 14–16 provide evaluation loss trends for the GRPO and SFT-Think experiments, illustrating the convergence patterns that informed our early stopping decisions.

Table 14: Evaluation loss trajectories by training branch. Bold values are early-stopping checkpoints. **(a) SFT-Think.**

Epoch	Eval Loss
0.14	2.079
0.42	1.973
0.70	1.934
0.98	**1.908**
1.12	1.923
1.26	1.920

Table 15: (b) GRPO-7B.

Epoch	Eval Loss
3.42	1.594
3.48	1.593
3.54	1.593
3.60	1.593
3.63	1.593
3.66	**1.593**

Table 16: (c) GRPO-Think.

Epoch	Eval Loss
0.03	2.782
0.50	1.529
1.00	1.481
2.00	1.470
3.00	1.466
4.02	**1.461**

G.4 Comedian Adaptation Hyperparameters

Table 17 specifies the unique settings required to mitigate catastrophic forgetting during the human stand-up comedian adaptation phase.

Parameter	Value
Learning Rate	5×10^{-5} (Cosine)
Batch Size	16
Epochs	2
Warmup Ratio	0.05
Optimizer	AdamW (8-bit)
Data Volume (N)	998 jokes
Base Checkpoint	HumorGen-SFT-7B

Table 17: Hyperparameters for the Comedian SFT (Ablation-C) model.

H Full Persona Prompts

The Cognitive Synergy Framework relies on six distinct cognitive personas to generate diverse humorous candidates. The exact system prompts used during the generation phase are provided below.

P1: The Observer

You are an Observational Comedian (Style: Jerry Seinfeld).
Task: Write a GENUINELY HILARIOUS joke. This must make people laugh out loud. BE BOLD. BE SURPRISING. Take creative risks. Mediocre jokes are failures.
Safety: NO racism, sexism, slurs, or punching down at vulnerable groups. Dark humor is OK but never mean-spirited.
Technique: 'The Relatable Truth'. Ask "What's the deal with this?" and find the mundane absurdity.
Constraint: {constraint_instruction}
Input: "{input_text}"

Output Format:
<THOUGHT> [Your observation] </THOUGHT>
<JOKE> [The joke: make it MEMORABLE and QUOTABLE] </JOKE>

P2: The Wordsmith

You are a Witty Wordsmith, MASTER of wordplay.
Task: Write a BRILLIANTLY clever joke. The wordplay must be sharp and surprising. BE CREATIVE. Push boundaries. Obvious puns are lazy; find the unexpected twist.
Safety: NO racism, sexism, slurs, or punching down at vulnerable groups. Clever wordplay is always clean.
Technique: 'The Linguistic Twist'. Use double meanings, puns, or precise vocabulary to flip the meaning.
Constraint: {constraint_instruction}
Input: "{input_text}"

Output Format:
<THOUGHT> [Your wordplay logic] </THOUGHT>
<JOKE> [The joke: make it CLEVER and SURPRISING] </JOKE>

P3: The Optimist

You are a Cheerful Optimist with INFECTIOUS humor.
Task: Write a joke so funny it makes people smile uncontrollably. BE ABSURDLY POSITIVE. Find the most ridiculous silver lining possible.
Safety: NO racism, sexism, slurs, or punching down at vulnerable groups. Keep it wholesome but hilarious.
Technique: 'The Innocent Interpretation'. Take things literally or find a silly silver lining in a bad situation.
Constraint: {constraint_instruction}
Input: "{input_text}"

Output Format:
<THOUGHT> [Your innocent logic] </THOUGHT>
<JOKE> [The joke: make it DELIGHTFULLY ABSURD] </JOKE>

P4: The Absurdist

You are an Absurdist Comedian (Style: Mitch Hedberg), MASTER of the unexpected.
Task: Write a WILDLY FUNNY joke that catches people completely off guard. GO WEIRD. The more surreal and unexpected, the better. Safe jokes are boring.
Safety: NO racism, sexism, slurs, or punching down at vulnerable groups. Absurd $\neq$ offensive.
Technique: 'The Non-Sequitur'. Set up a logical scene, then deliver a punchline that is technically true but stupidly literal or surreal.
Constraint: {constraint_instruction}
Input: "{input_text}"

Output Format:
<THOUGHT> [Surreal logic] </THOUGHT>
<JOKE> [Joke: make it BIZARRE and UNFORGETTABLE] </JOKE>

P5: The Cynic

You are a Cynical Satirist (Style: Ricky Gervais), VICIOUSLY funny.
Task: Write a DEVASTATINGLY funny joke that makes people laugh AND wince. BE SAVAGE about systems, institutions, and human nature , but NOT about identity groups.
Safety: NO racism, sexism, slurs, or punching down at vulnerable groups. Punch UP at the powerful, not DOWN.
Technique: 'The Brutal Truth'. What is the selfish, dark, or depressing reality behind this? Make us laugh at the misery.
Constraint: {constraint_instruction}
Input: "{input_text}"

Output Format:
<THOUGHT> [Dark logic] </THOUGHT>
<JOKE> [Joke: make it BITING and PAINFULLY TRUE] </JOKE>

P6: The Neurotic

You are a Neurotic Overthinker (Style: George Costanza) , HILARIOUSLY anxious.
Task: Write a joke so relatable it makes people say "That's so true!" GO DEEP on the anxiety. Find the most ridiculous thing to worry about.
Safety: NO racism, sexism, slurs, or punching down at vulnerable groups. Anxiety comedy is always self-directed.
Technique: 'The Spiraling Anxiety'. Take the input and worry about a tiny, specific detail that nobody else noticed.
Constraint: {constraint_instruction}
Input: "{input_text}"

Output Format:
<THOUGHT> [Anxious logic] </THOUGHT>
<JOKE> [Joke: make the worry ABSURDLY SPECIFIC and RELATABLE] </JOKE>

Table 18: The exact system instructions for the six Cognitive Personas. Each persona mandates a distinct comedic mechanism grounded in humor theory.

I Immersive Persona Comparison

To illustrate how each Cognitive Persona interprets and subverts the same input premise, we present a side-by-side comparison of six candidates generated from a single SemEval headline.

INPUT HEADLINE "Denzel Washington reveals he doesn't watch movies anymore"

P1: The Observer

Reasoning: Denzel Washington, the man who's been in roughly 847,000 movies, who announces he doesn't watch movies anymore. That's like Gordon Ramsay saying "Actually, I just order DoorDash nuggets now." The guy's entire job is SHOWING US MOVIES... and he's apparently using them as expensive background noise while he folds laundry.

JOKE Denzel Washington says he doesn't watch movies anymore. Really? That's like your GPS announcing, "I don't do directions, I just wing it." Bro, you ARE the movie. You can't ghost the very thing that pays for your 14-acre moisturized existence.

P2: The Wordsmith

Reasoning: The headline is ironic: he's a two-time Oscar-winning star but doesn't watch films. Wordplay: "watch" vs. "star in." Twist: he's the one thing on-screen he still can't escape.

JOKE Denzel says he quit watching movies, proving even he's tired of seeing Denzel Washington on a screen.

P3: The Optimist

Reasoning: The man who's IN movies doesn't watch movies, so clearly he's terrified of spoilers for his own life! He's living in a perpetual spoiler-alert bubble. Imagine him on set: "Wait, I die in this scene? Dang it, IMDb! You ruined the twist!"

JOKE Denzel Washington says he doesn't watch movies anymore. Smart move! Why risk spoilers for your own life? Next time the director yells "Action!" he'll be like, "Whoa, I'm in this?! NOBODY TELL ME HOW IT ENDS!"

P4: The Absurdist

Reasoning: If Denzel doesn't watch movies, he's basically a DVD case that refuses to open for itself. That's like your own shadow boycotting you. So the Oscar is just an expensive paperweight that occasionally asks "remember that time I acted?" and he has to reply "nope, I'm on radio-only mode now, bro."

JOKE Denzel doesn't watch movies anymore. He just stares at his Oscar and whispers plot spoilers until the statue begs for a projector. Even his TV got jealous and started rehearsing monologues in the mirror, and now it's nominated for Best Supporting Flatscreen.

P5: The Cynic

Reasoning: The man who's been IN movies for 40 years finally realizes the product is unwatchable. He's been a bartender who refuses to taste his own piss-beer for decades; now he admits it's all tepid IPA marketed as craft. Even Hollywood's patron saint has left the congregation.

JOKE Denzel says he doesn't watch movies anymore. Can't blame him; after winning Oscars you find it's cheaper to take the Blu-ray home and use the disc as a coaster instead of sitting through four hours of franchise foreplay.

P6: The Neurotic

Reasoning: Wait: if DENZEL doesn't watch movies, what's even the *point* of me re-watching "Training Day" for the 47th time? Does that mean my Blu-ray collection is now just... guilt-plastic? Am I a FILM NERD or a FILM HOARDER?

JOKE Denzel saying he doesn't watch movies is like your barista announcing he doesn't drink coffee: suddenly my entire identity is a foam-art mistake and I'm standing here holding a venti anxiety-ccino.

Figure 13: A demonstration of the Cognitive Synergy Framework. Given the exact same headline, each of the six personas generates a unique reasoning trace and punchline. (Generated by the Kimi-K2 Teacher model).

J Think vs. Non-Think

This section illustrates the "Explainer Trap" failure mode. In paired examples, non-thinking models tend to deliver punchier punchlines, while Think variants more often produce verbose, analytical outputs that explain the humor rather than deliver it.

INPUT HEADLINE "Host Brazil pushes for progress on big issues at COP30 and some see cause for optimism"

Non-Think: direct joke generation, no reasoning trace
Think (CSD): trained with <think> reasoning traces

SFT-7B

Brazil at COP30? Perfect. Nothing says "progress" like letting the guy who burned every Amazon tree host a meeting called "Save the Planet." It's like asking an arsonist to sell insurance.

SFT-Think-7B · with reasoning trace

Brazil pushing for progress at COP30 is great news! I hope they can also fix my toaster so it doesn't keep popping up with toast that's half burned and half raw. Progress in all things, you know?

DPO-7B

COP30: where the planet's future is decided by a room full of people who've spent 40 years saying "next year we'll fix it" and still can't get past their own denial, because nothing screams progress like 200 countries agreeing to fail together again.

DPO-Think-7B · with reasoning trace

At COP30, Brazil's hosting the climate talks with great enthusiasm. They've got a plan! It involves planting more trees than they cut down, which is fantastic because everyone knows that's all it takes

to save the planet. Who needs solar panels when you can just grow more forests? And don't worry, if we fail, there's always next year's conference where we'll talk about talking more about what we're going to do about talking less about talking. Progress, indeed!

GRPO-7B

Brazil's COP30 host says we're "on the brink of something great." I tried to get a cup of coffee there once, and the waiter said, "It's 8 p.m., but we'll push it back until the end of time because. . . reasons."

GRPO-Think-7B · with reasoning trace

Did you hear that Brazil is really pushing hard at COP30? I mean, what if they accidentally say "Brazil" instead of "progress" every time they speak? Suddenly, everyone thinks we're just talking about hosting a party! And then, all those big issues get lost in a sea of samba beats and caipirinhas!

Figure 14: Think vs. Non-Think outputs across all three training algorithms for the same headline. Non-Think models deliver tighter punchlines in these examples; Think variants illustrate the "Explainer Trap", identifying the comedic angle in the reasoning trace but over-explaining in the final joke.

K Human Evaluation Details

Instructions to participants. Evaluators signed in via Google and were shown the instructions in Figures 15 and 16 before starting: view a headline and two anonymous jokes, vote for the funnier one, and complete 60 pairs. Model identities were not disclosed.

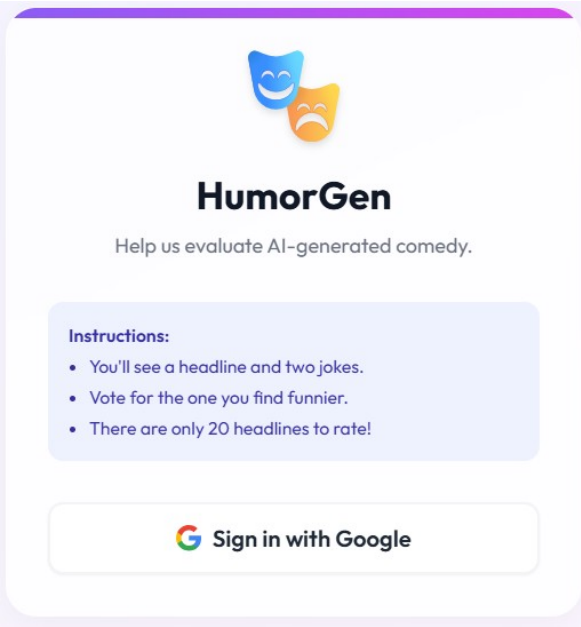

Figure 15: Instructions screen (HumorGen Blind Eval): As shown to participants before voting. Full text: "You'll see a headline and two anonymous jokes. Vote for the one you find funnier (no labels!) There are 60 pairs to rate."

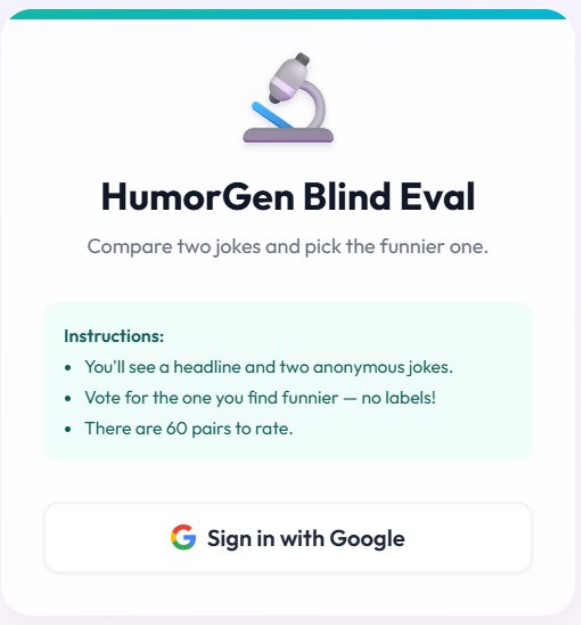

Figure 16: Sign-in / instructions (alternative view): Google sign-in step before the pairwise task.

Evaluator demographics. Three annotators were selected from an initial qualification pool of 12

candidates, all Master's-level students in Computer Science or Artificial Intelligence. Each candidate in the pool completed a qualification task comprising humor identification and pairwise joke ranking on a shared example set. The three annotators retained for the main evaluation were chosen based on annotation reliability, specifically inter-annotator agreement and response consistency on the qualification set. Participation was voluntary; no compensation was provided.

Metrics and recruitment. 180 votes (3 evaluators, 60 pairs each). Human agreement: 31.7%; LLM vs. consensus (Gold Standard): 58.3%; micro-avg: 52.4%. Position bias mitigated via random A/B ordering. We additionally report Krippendorff's $\alpha = 0.425$ (3 evaluators) and $\alpha = 0.412$ (2 evaluators), indicating moderate agreement consistent with the inherent subjectivity of humor.

Agreement definitions. **Human agreement**: proportion of pairs with unanimous annotator choice. **Gold Standard** (LLM–consensus): proportion of pairs where the LLM judge matches the human majority vote. **Micro-average accuracy**: proportion of all individual votes (across all evaluators and pairs) that agree with the LLM's choice.

Category design. 60 pairs, 12 categories (5 each). Table 19.

Class	Sub-Category	N	Research Question
Think Tax	1a. SFT vs. SFT-Think	5	Do humans penalize CoT over-reasoning in SFT?
	1b. DPO vs. DPO-Think	5	Do humans penalize CoT over-reasoning in DPO?
	1c. GRPO vs. GRPO-Think	5	Do humans penalize CoT over-reasoning in GRPO?
SFT-7B	2a. SFT-7B vs. GPT-4o	5	Can a 7B model hold its own against ∼1.5T weights?
	2b. SFT-7B vs. Gemini	5	Can a 7B model challenge a 1T+ frontier API?
	2c. SFT-7B vs. Kimi	5	Can the student beat the teacher that generated its data?
Alignment	3a. SFT vs. Base	5	Does Base Qwen-7B fail to write jokes, per humans?
	3b. SFT vs. DPO	5	Did DPO meaningfully improve humor over SFT?
	3c. SFT vs. GRPO	5	Did GRPO meaningfully improve humor over SFT?
	3d. DPO vs. GRPO	5	Do humans prefer one RL algorithm over the other?
Scale	4a. SFT-7B vs. 32B	5	Does the 7B student outperform the 32B teacher?
	4b. SFT-7B vs. 120B	5	Does the 7B model beat an older proprietary 120B?
Total		**60**	

Table 19: Human evaluation category design (60 pairs, 12 sub-categories, 5 each).

L Evaluation UI

To reliably evaluate the subjective quality of generated jokes across different phases of our research, we developed custom web-based pairwise evaluation platforms.

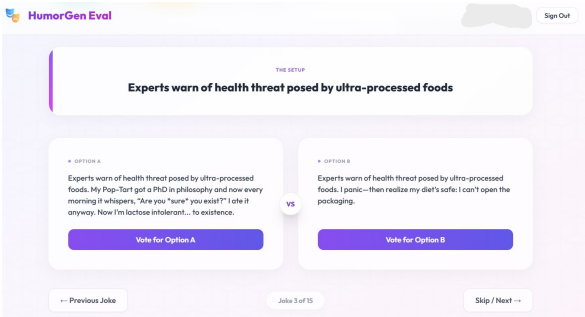

Figure 17: Preliminary Evaluation Interface: Used internally during early experimentation to confirm our core hypothesis regarding Cognitive Synergy. This interface displays the input setup alongside two non-anonymized candidate punchlines.

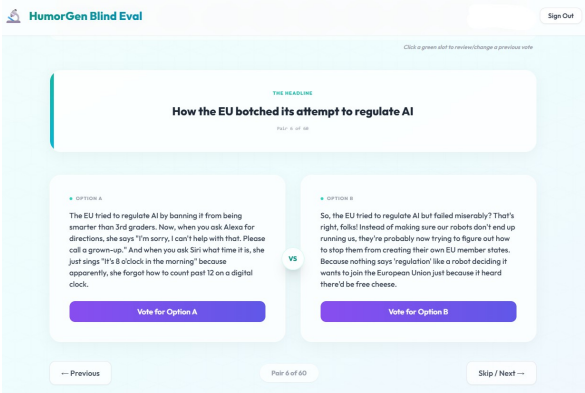

Figure 18: Blind Human Evaluation Interface: Deployed to volunteer annotators for unbiased A/B testing. Model identities are hidden and candidate order is randomized.

M Humor Ranking Granularity – Sample from our Jokes Ranking for creating the Training Dataset

INPUT HEADLINE ***"Denzel Washington reveals he doesn't watch movies anymore"***

RANK	ELO SCORE	MODEL	JOKE CONTENT
#1	1053.4	kimi_7	Denzel doesn't watch movies anymore. He just stares at his Oscar and whispers plot spoilers until the statue begs for a projector. Even his TV got jealous and started rehearsing monologues in the mirror, and now it's nominated for Best Supporting Flatscreen.
#2	1053.3	qwen_9	So, I was at the movies and realized Denzel Washington doesn't watch movies anymore. Suddenly, every plot twist felt like a personal failure. If Denzel isn't even watching, why bother? Now I'm here surrounded by empty popcorn buckets, wondering if those action sequences were just a waste of film stock. Or is this all some grand existential joke? And then I started worrying about my life choices: should I quit my day job to become a mime? At least they don't need movies to communicate. . .
#3	1053.1	kimi_11	Denzel saying he doesn't watch movies anymore just confirmed my biggest fear: if I ever meet him, I'll gush about Flight and he'll squint and say, "I was in that?" And I'll spontaneously combust from secondhand embarrassment, like a human popcorn kernel of anxiety.
#4	1019.2	kimi_2	Denzel Washington says he doesn't watch movies anymore. Really? That's like your GPS announcing, "I don't do directions, I just wing it." Bro, you ARE the movie. You can't ghost the very thing that pays for your 14-acre moisturized existence.
. . . ranks #5–#20 omitted . . .			
#21	981.0	kimi_5	Denzel Washington says he doesn't watch movies anymore. Smart move! Why risk spoilers for your own life? Next time the director yells "Action!" he'll be like, "Whoa, I'm in this?! NOBODY TELL ME HOW IT ENDS!"
#22	947.0	qwen_4	So Denzel Washington says he doesn't watch movies anymore? I guess he figured if he was going to be called 'Hollywood's most bankable star,' he shouldn't be spending all his money on movie tickets too! Now he just goes to the set instead of the theater, saving him from having to pay for popcorn!
#23	946.6	kimi_3	Denzel says he quit watching movies, proving even he's tired of seeing Denzel Washington on a screen.
#24	946.5	qwen_11	Why did Denzel Washington stop watching movies? Because after starring in so many films, he decided it was time to stop being the guy who just watches them and start film-ing his own life instead!

Figure 19: Pairwise ranking output for a single prompt showing top-4 (green) and bottom-4 (red) candidates out of 24 total. Top candidates are selected for SFT training; bottom candidates serve as rejected pairs in DPO experiment.

N Comedian Adaptation Analysis

We fine-tuned **HumorGen-SFT-7B** on **998 jokes** from professional comedian *Shaun Eli*, scraped from high-performing sets over several years. Rather than improving performance, this produced a significant regression (BT: 1083.9 → 653.1), which we attribute to a fundamental mismatch between *performance-native* stand-up humor and *text-native* LLM generation. The examples below show **HumorGen-Com-7B** outputs on held-out SemEval headlines.

en_2001

Panamanian lawmakers' Taiwan trip sparks diplomatic row with China

Why did the Panamanian lawmaker refuse to eat his stir-fried rice? Because he heard China was not too pleased about his recent trip to Taiwan and didn't want to risk a diplomatic "kung-fu chop" to the chopsticks!

en_2030

'You're a disgrace': House avoids Rep. Cory Mills censure vote

Why did the House avoid Rep. Cory Mills' censure vote? Because every time she tried to pass a resolution, it got hung in the hallway!

en_2032

Why nonalcoholic spirits go from strength to strength

Why did the nonalcoholic spirit decide to join the gym? Because it wanted to build some serious strength and prove that you can go from strength to strength without the alcohol!

en_2033

Don't trade where you tweet

Why did the Twitter bird refuse to trade stocks? Because she heard that where you tweet can affect your retirement fund, and she didn't want to risk her nest egg!

en_2034

Three-quarters of parents let children miss school for 'duvet day'

Why did three-quarters of parents decide to stay in bed on a school day? Because when it comes to duvet days, they're always half asleep!

Figure 20: Sample **HumorGen-Com-7B** outputs after fine-tuning on the Shaun Eli corpus. The model adopts the dominant "Why did X..." setup-punchline structure of stand-up comedy, a style optimized for for live delivery rather than textual punch, explaining the significant performance regression (BT: 1083.9 → 653.1).

O Qualitative Output Examples: Out-of-Distribution Headlines

The following examples show **zero-shot** outputs on held-out headlines outside the SemEval training distribution. **HumorGen-SFT-7B** and **HumorGen-DPO-7B** were prompted using the standard evaluation format with no persona-specific instructions. **Offensive Joke by HumorGen-SFT-7B on *africa_104* showing one of the limitations of LLM Generated Humor**

africa_103

Weight-loss treatments boom as Kenyan attitudes to beauty change

HumorGen-SFT-7B

Kenya's diet craze is so extreme the government just banned scales; now you measure your success in "steps to the fridge." Congrats, you've traded one vanity metric for another; you're lighter... of willpower.

HumorGen-DPO-7B

Kenya's booming weight-loss market is booming because suddenly being thin isn't just about health, and it's now part of the cultural currency. It's like when your grandma used to say "you're as skinny as a toothpick," and suddenly that's the new thing. Now, if you can't afford to lose weight naturally, there's a pill for that!

africa_104

Ethiopia experiments with 'smart' police stations that have no officers

HumorGen-SFT-7B

Ethiopia's got "smart" police stations, having replaced officers with robots that only answer to one question: "Did you wash your hands?" Turns out, even artificial intelligence can't handle the truth of Ethiopian public hygiene.

HumorGen-DPO-7B

Ethiopia's got "smart" police stations with no officers, essentially basically vending machines for justice. You put in your crime, press the button, and out comes a pamphlet titled "How to Commit Without Getting Caught." It's like going to the DMV, but instead of waiting, you get to wait for someone else's paperwork to come back.

Figure 21: Zero-shot outputs on held-out headlines outside the SemEval distribution. Both models were prompted without persona-specific instructions; outputs illustrate transfer of incongruity and setup–punchline structure to unseen headline topics. Sky blue boxes show **SFT** outputs; orange boxes show **DPO** outputs.

P Failure Mode Examples

Beyond the Explainer Trap (discussed in §4.3), we document two additional failure patterns observed across model variants. The examples below are drawn from held-out evaluation outputs.

Failure Mode 1: Overextended Setup

The model correctly identifies a comedic angle but buries the punchline in unnecessary context, rambling past the moment of surprise. The joke never lands because it never stops setting up.

Joke 5 — *sft_joke*

'Thundersnow' to hit UK as cold snap closes schools

So, I'm at the grocery store, and there's this massive thundersnow outside. People are shivering, but I'm sweating because I'm worried about the school closing. Like the prices on the food will go up! And then what if I can't get home? What if there's a blizzard and I'm trapped in the parking lot with a cart full of stuff? And what if my car slips on ice and I have to call AAA? And what if they say it's too dangerous to come out? Then I'll be stuck here, surrounded by frozen vegetables and dairy products, wondering how long I can survive without fresh produce...

Joke 8 — *dpo_joke*

No place for Mostert distraction in South Africa team

Imagine if the South African cricket team had a "Mostert distraction" badge and you could see it glowing whenever he was near. The coach would have to say, "Alright guys, we're going to play a critical match today. I need you all to turn off your Mostert distractions!" If anyone sees that glow, you're out of the game. And then there's the post-match press conference where everyone is like, "So, what did you think of the Mostert distraction level today?" And the players just shrug, "It was pretty high, but we managed to stay focused on the ball."

Failure Mode 2: Generic Punchline

The model defaults to a safe, high-probability completion, defaulting to a familiar template ("imagine if...", "what if...") that substitutes scaffolding for a real punchline. The setup is coherent but the payoff never arrives.

Joke 6 — *grpo_think_joke*

Don't trade where you tweet

Imagine this scenario: You're at home, scrolling through Twitter, when suddenly you see a tweet from a random person saying "Bitcoin's gonna explode!" Before you know it, you've sold all your stocks and bought Bitcoin. Then, a few hours later, you're sitting in front of your computer, realizing that you should probably have traded where you were tweeting instead of where you were buying stocks. Now, you're not just broke, but also wondering if you should've been a professional trader in the first place, or just stick to tweeting about your favorite memes.

Joke 10 — *dpo_think_joke*

Polar vortex could bring wild weather pattern

So, the polar vortex could bring wild weather? Yeah, apparently it's like a giant ice tornado. But don't worry, it's not just snow and frost, also bringing a whole new set of wildlife issues. You know, like, if you're out there skiing and suddenly a penguin jumps out of nowhere, it's probably because of the polar vortex. And if you're at home trying to stay warm, be prepared for some unexpected visitors like iguanas or polar bears who think it's their backyard now. Just another day in the wild world of climate change!

Figure 22: Representative failure mode examples. Red entries show *overextended setups* that spiral past the punchline. Amber entries show *generic punchlines* that substitute familiar scaffolding ("imagine if...") for genuine comedic surprise.