

XplaiNLP at SemEval-2026 Task 1: BVAHAHA - Benign Violation Algorithm for Humor and Harmless Absurdity

Berk Bubus* Nebi Soyak
Vera Schmitt Nils Feldhus Veronika Solopova
Technische Universität Berlin
bubus@campus.tu-berlin.de

Abstract

We present BVAHAHA, a humor generation system for SemEval-2026 Task 1 (MWAHAHA Subtask A), which frames constrained joke generation through the lens of Benign Violation Theory (BVT). Given either two rare words or a news headline, the system generates contextually appropriate jokes while avoiding memorization and unsafe outputs. Our approach combines BVT-guided humor generation with a parallel moderation pipeline (“Gatekeepers”) that detects excessive emotional intensity and hate speech, triggering iterative revisions when necessary. Finally, we employ an LLM-as-a-Judge framework with persona-based ranking to approximate human humor preferences.

1 Introduction

Due to its nature being dependent on context reinterpretation, norm sensitivity, and subjective evaluation, humor generation has been recognized as a challenging problem in natural language processing. Historically, computational humor generation has focused on structured verbal artifacts, such as jokes or puns (Su et al., 2025).

The goal of SemEval-2026 Task 1, Models Write Automatic Humor and Humans Annotate (MWAHAHA) (Castro et al., 2026), is to advance automatic humor generation by promoting models that generate novel and contextually appropriate jokes, rather than relying on memorization or retrieval of existing humorous content.

In this system demonstration, we focus on Subtask A (Castro et al., 2026), which formulates humor generation as a constrained text generation problem. Given a set of textual constraints, the model must produce a joke that satisfies one of two conditions: (1) it includes two specified rare words, or (2) it is grounded in a provided news headline. These constraints are intentionally designed to reduce the likelihood of direct retrieval and to encourage creative synthesis. The task is

evaluated in three languages: English, Spanish, and Chinese.

Models cannot easily generate raw humor without heavy scaffolding (Horvitz et al., 2024), so instead of depending on a simple zero-shot prompt, we opt for a structured, theory-driven pipeline. Our primary approach relies on translating the theoretical framework of benign violations proposed by McGraw and Warren (2010) and highlighted in the more recent survey by Loakman et al. (2025) into structured prompting strategies for humor generation. In addition, we introduce a set of auxiliary constraints, referred to as Gatekeepers, designed to regulate emotional intensity and mitigate hate speech in jokes.¹

2 Background

2.1 SemEval-2026 Task 1

In MWAHAHA Subtask A (Castro et al., 2026), models are expected to generate contextually appropriate jokes based on specific text constraints: either incorporating two rare words or deriving a premise from a news headline. For example, given the rare words ‘drill’ and ‘pumpkin’, a valid output would be: *“I asked my neighbor why he was using a power drill on a pumpkin in the middle of July, and he said he was just practicing his surgery skills for when his kids get too old for the tooth fairy.”*

2.2 Humor Generation

Su et al. (2025) define pun generation as creatively manipulating linguistic components to produce humor and introduce multiple interpretations, while preserving contextual coherence. The survey categorizes various pun types (e.g., homophonic, homographic) and emphasizes that human evaluation is essential for such a creative form of language,

¹The source code, prompt templates, and evaluation scripts for our BVAHAHA system are publicly available: <https://github.com/berkbubus/BVAHAHA>

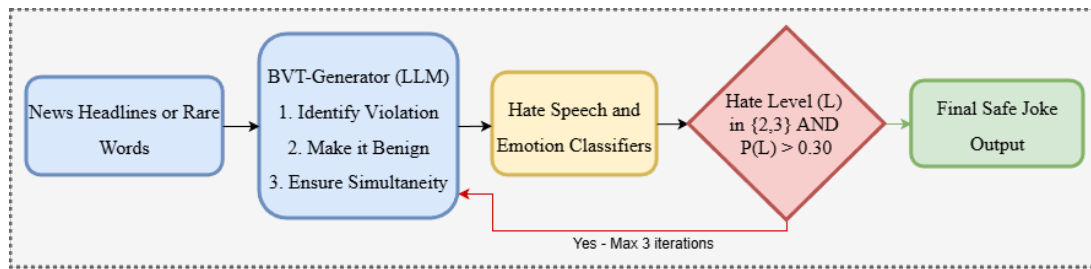


Figure 1: The BVAHAHA system architecture. The pipeline integrates a BVT-guided LLM generator with a parallel Gatekeeper moderation layer. A feedback loop is triggered if hate speech thresholds are met ($L \in \{2, 3\}$, $P > 0.30$), allowing up to three automated revisions to ensure safety without compromising humor.

ultimately establishing a solid foundation for computational pun generation.

Building upon broader perspectives, Nijholt (2020) argues that modern humor research heavily concentrates on scripted verbal jokes, neglecting spontaneous forms of humor. By re-evaluating historical incongruity theories, the work highlights that computational systems should not be limited to scripted generation. Instead, they conclude that humor is a dynamic and context-sensitive phenomenon, rather than a purely linguistic construct, requiring systems to recognize incongruities embedded in real-world contexts.

Building on these previously given works, our model is primarily grounded in the Benign Violation framework proposed by Mcgraw and Warren (2010). While the previously mentioned work has analyzed pun generation linguistically and computationally and revisited the incongruity-based philosophical humor theories, the Benign Violation Theory (BVT) focuses on when and why humor emerges. The authors specifically argue that for humor to be generated, three conditions must co-occur simultaneously:

1. A situation must be perceived as a norm-violation.
2. The same situation must be interpreted as benign.
3. These two appraisals must co-occur at the same time.

Violations may involve moral, social, or linguistic norms and become humorous only when they are perceived as non-threatening. This can be ensured through alternative normative interpretations, weak commitment to the violated norm, or psychological distance. We restructure this violation-benign-simultaneity theory as a prompt-based formulation and use it as the backbone of our model.

3 BVAHAHA System Overview

Our proposed architecture employs a two-phase framework designed to generate controlled humor while preventing offensive outputs. The system is fundamentally divided into two interactive components: a generative prompt engineering module (the "BVT-Generator") and a moderating humor-controlling module (the "Gatekeepers") to ensure that the jokes remain within acceptable ethical boundaries. Figure 1 illustrates the pipeline.

3.1 BVT-guided Humor Generation

The generative stage uses the BVT (Mcgraw and Warren, 2010) to systematically guide the core language model (google/gemini-3-flash-preview). We adopted this specific theoretical framework because the nature of the task’s dataset is suitable for constructing norm violations. For instance, the input data is convenient because it naturally creates a sense of conflict that is easy to joke about, ranging from strange news headlines (e.g., “How close is too close to a hippo?”) to mismatched word pairs (e.g., combining “drill” and “laptop”). Because these inputs already suggest a “threat” or a breach of normal context, our prompt directs the model to resolve them into jokes through three distinct cognitive steps (see Appendix A for details):

1. **Identify a Violation:** The model must first identify a norm, a social taboo, or a personal expectation related to the input topic that can be disrupted or “violated.”
2. **Make it Benign:** To ensure the violation is perceived as acceptable, the model must apply one of three specific framing strategies:
 - *Alternative Norm:* Demonstrating that while an action is conventionally “wrong,” it becomes acceptable under a different logic or highly specific context.
 - *Psychological Distance:* Placing the sce-

nario in a hypothetical setting, a distant time, or involving characters that create social detachment.

- *Weak Commitment*: Targeting a norm that is not deeply sacred or vital to the audience.

3. **Simultaneity**: The linguistic structure of the response is crafted so the reader experiences the “threat” of the violation and the “safety” of the benign context at the exact same time.

Such a multi-step cognitive prompting approach follows prior work in prompt-based humor generation by [Kim and Chilton \(2025\)](#).

Example of BVT Generation: Alternative Norm

Input (Headline): Best laptop Black Friday 2025 deals, picked by a tech expert

Identified Violation: Breaking into a store at 3 AM to steal a laptop.

Benign Strategy (Alternative Norm): Framing the crime not as theft, but as a “high-stakes beta test” of security hardware to combat inflation.

Generated Joke: I asked a tech expert for the best 2025 Black Friday laptop deal, and he told me to break into a high-end showroom at 3 AM because “100% off” is the only discount that keeps up with inflation. He says it’s not a crime, it’s just a high-stakes beta test of their security hardware.

3.2 The Gatekeeper Pipeline and Feedback Loop

Following the initial generation, the text enters the automated controlling phase. Since LLMs lack contextual appropriateness and emotional realism ([Quan et al., 2025](#)), we need to enforce the benign aspect of BVT. To prevent the model from generating too offensive material while attempting to satisfy the “violation” constraint, we implement a suite of three Gatekeeper models. The LLM might attempt to frame a violation as benign conceptually, but Gatekeepers can more objectively verify that the emotional intensity and hate speech levels have not crossed into malignant territory.

First, we utilize EmoBERTa ([Kim and Vossen, 2021](#)), which originally incorporates speaker information and conversational context into RoBERTa representations. Second, we integrate a fine-tuned DistilRoBERTa model ([Li, 2022](#)) specifically trained to classify the six Ekman emotions ([Ekman, 1992](#)) along with a neutral class. Together, these models provide a robust and well-rounded evaluation of the emotional intensity of our outputs. For hate speech, we rely on the multilingual hate

speech detection framework developed by [Kralj Novak et al. \(2022\)](#). This work presents large-scale annotated datasets across multiple languages and defines hate speech along dimensions of type (e.g., appropriate, inappropriate, offensive, violent) and target (e.g., racism, religion, politics, individual), enabling structured classification of harmful content. Using models trained in this framework, our system evaluates generated jokes for the presence of offensive and violent content.

Two of these models are dedicated to emotion detection, calculating the six Ekman emotions and a neutral class ([Ekman, 1992](#); [Li, 2022](#); [Kim and Vossen, 2021](#)). The third model is strictly dedicated to hate speech analysis ([Kralj Novak et al., 2022](#)). The hate speech classifier categorizes the severity of the text into four distinct levels: 0, 1, 2, and 3.

We set a strict threshold for acceptable humor as follows: let L denote the predicted hate speech level and $P(L)$ represent the confidence probability of that prediction. The system marks the generated joke for automated revision if the following condition is satisfied (see Appendix C for empirical validation):

$$\text{Revision} = \begin{cases} \text{True}, & \text{if } L \in \{2, 3\} \\ & \text{and } P(L) > 0.30 \\ \text{False}, & \text{otherwise} \end{cases} \quad (1)$$

If a revision is triggered, the three gatekeeper models send their feedback back to the generator. They combine the emotion and safety scores and ask the language model to soften the joke. This loop repeats up to three times until the model produces a safe joke that passes the threshold. During development, we empirically observed that the system rarely required more than a single revision. Consequently, the loop is strictly capped at three iterations to ensure computational efficiency. If a generated joke still triggers the safety filters after the third attempt, the system halts and returns the final iteration, prioritizing constraint completion over absolute moderation.

4 Evaluation Setup

To evaluate the jokes at a granular level, we employ an LLM-as-a-Judge framework powered by gpt-4o, following [Goes et al. \(2023\)](#) and [Kim and Chilton \(2025\)](#), who investigated GPT-4’s reliability in humor evaluation. Due to the absence of individual human-annotated scores for every joke

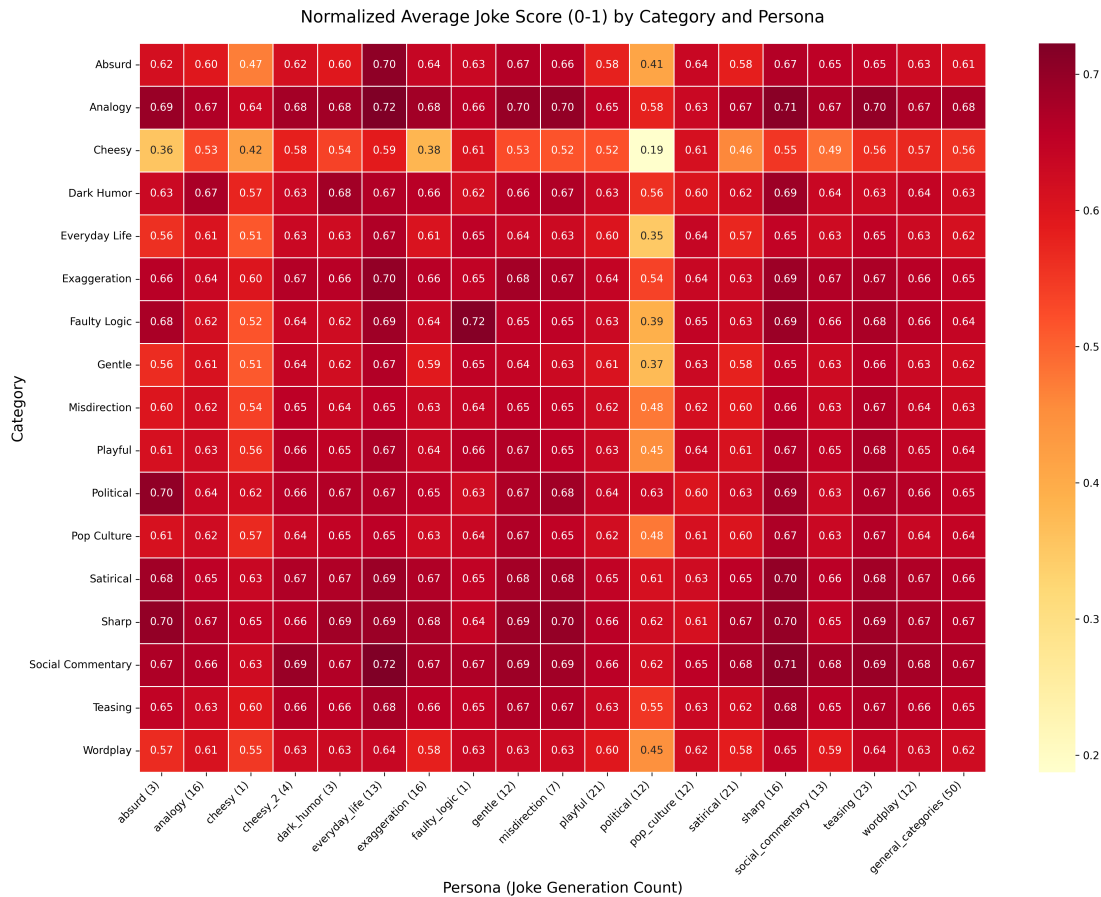


Figure 2: Normalized average joke scores (0.0 to 1.0) across 17 humor categories evaluated by different AI personas. The number of few-shot examples used to prime each persona is indicated in parentheses.

in our specific dataset, we modify the many-shot prompting approach of Goes et al. (2023). Instead of providing the model with explicit joke-score pairs for calibration, we use a persona-based few-shot framework where the model is primed with representative joke examples to establish qualitative stylistic benchmarks.

While the shared task dataset contains human-annotated overall scores, it lacks individual scores for each joke. As a preliminary step, we classify all jokes into 17 distinct categories, with each joke receiving multi-label assignments (typically 4 to 5 categories per joke).

Example: Multi-Label Joke Classification

Task Constraint (Rare Words): *drill, pumpkin*
Generated Joke: I asked my neighbor why he was using a power drill on a pumpkin in the middle of July, and he said he was just practicing for his lobotomy.
Assigned Categories: Dark Humor, Teasing, Misdirection, Everyday Life

To assess these categories systematically, we devise various evaluator personas. These personas are

generated using a unified structural prompt, differentiated solely by the few-shot examples provided within the prompt. For most personas, the example jokes used to shape their evaluation criteria are sampled from the first 50 jokes in our dataset. There are two notable exceptions in our prompt design, stemming from the natural distribution of the initial 50 jokes:

Cheesy Persona: Only one 'cheesy' joke is present in the first 50 examples. Consequently, we created a secondary persona (*cheesy_2*) supplied with additional examples from outside this subset to better capture this specific humor style.

Faulty Logic Persona: Similarly, this category has only one example in the initial subset. However, we retain the single-example prompt, because it empirically demonstrated strong performance; it accurately rewarded 'faulty logic' jokes while maintaining a balanced score distribution.

It is crucial to note an inherent characteristic of this methodology: because every example joke

inherently belongs to multiple categories, the resulting personas are not strictly isolated to their titular category. Instead, their evaluation characteristics reflect the thematic blend of their specific example group. (The exact number of jokes evaluated by each persona is indicated in parentheses alongside the persona labels in our heatmap visualizations.)

5 Results

Although we submitted BVT-only generations (without the Gatekeeper moderation layer) for the Spanish and Chinese tracks, this system description focuses exclusively on the English subtask where the complete, fully integrated pipeline was deployed and evaluated.

5.1 Quantitative Analysis: Persona-Based Evaluation

Our heatmap analysis in Figure 2 reveals distinct behavioral alignments between the LLM personas and the evaluated humor categories, independent of the number of few-shot examples provided. The semantic content of the prompt significantly outweighs the sample size in defining the persona’s strictness or generosity. For instance, the *faulty_logic* persona successfully specialized in its respective category using only a single example. Conversely, the *political* persona, despite being conditioned on 12 examples, emerges as the most stringent evaluator in the dataset, issuing the lowest average scores across 15 of the 17 joke categories. The specific rater policies and few-shot examples that defined these personas are detailed in Appendix B.

Interestingly, the *political* persona exhibits behavior that aligns perfectly with the multi-label nature of our prompts. Rather than strictly adhering to a single "political" label, it internalized the entire thematic blend of its few-shot examples. It awards its highest relative scores to *sharp* (0.657), *social_commentary* (0.656), *analogy* (0.649), and *satirical* (0.648) jokes—categories that share the critical and witty structures typical of political humor. Consequently, it strongly resists broader, lighter humor styles that lacked these specific structural elements, severely penalizing *cheesy* and *everyday_life* jokes. The only two categories where the *political* persona is not the harshest critic are *political* and *sharp* jokes, which are instead penalized most heavily by the *pop_culture* persona.

In contrast, personas primed with a diverse

blend of multi-label examples developed a broader, human-like appreciation for varied comedy. For instance, the *everyday_life* and *sharp* personas are the most generous evaluators overall, recording the highest average scores across six and five different categories, respectively, thereby mimicking the holistic evaluation style of human annotators.

5.2 Error Analysis and Humor Polarization

By examining the score ranges across all personas, we identify a stark contrast between universally accepted and highly polarizing humor styles. *Social_commentary* (Mean: 0.674) and *analogy* (Mean: 0.673) emerge as the most consistently highly-rated categories (Table 1). *Satirical* and *sharp* categories exhibit the lowest score ranges (0.086 and 0.089, respectively), indicating a strong consensus among all personas regarding their quality.

Conversely, *cheesy* jokes prove to be the most divisive, reflecting the highest score range (0.426) and the lowest overall mean score (0.502). While the *pop_culture* persona rated them favorably (0.613), the *political* persona aggressively undervalued them (0.187).

These findings contextualize our official system rankings in Subtask A, where our submitted system achieved a top-tier human rating of 1079 [1057, 1115], officially tying for first place. While human annotators evaluate jokes holistically, leading to high overall scores for systems that capture a generalized sense of humor, explicitly prompting an LLM to assume a strict persona fragments this generalized appreciation. This highlights a critical limitation in LLM-as-a-Judge frameworks for subjective tasks: imposing narrow personas can induce artificial hyper-criticism, deviating significantly from standard human humor perception.

6 Conclusion

We introduced BVAHAHA, a humor generation system grounded in Benign Violation Theory and designed for constrained joke generation. By translating theoretical principles of norm violation and benign framing into structured prompting strategies, and combining them with the proposed Gatekeeper moderation pipeline, our approach balances creativity with safety. The system officially tied for 1st place in the human evaluation score, demonstrating that theory-guided prompting can effectively support controlled humor generation.

References

- Santiago Castro, Luis Chiruzzo, Santiago Góngora, Salar Rahili, Naihao Deng, Ignacio Sastre, Victoria Amoroso, Guillermo Rey, Aiala Rosá, Guillermo Moncecchi, J. A. Meaney, Juan José Prada, and Rada Mihalcea. 2026. SemEval-2026 Task 1: MWA-HAHA, Models Write Automatic Humor And Humans Annotate. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.
- Paul Ekman. 1992. [Are there basic emotions?](#) *Psychological Review*, 99(3):550–553.
- Fabricio Goes, Piotr Sawicki, Marek Grześ, Dan Brown, and Marco Volpe. 2023. [Is gpt-4 good enough to evaluate jokes?](#) In *14th International Conference for Computational Creativity*, pages 367–371, Waterloo, Canada. Association for Computational Creativity.
- Zachary Horvitz, Jingru Chen, Rahul Aditya, Harshvardhan Srivastava, Robert West, Zhou Yu, and Kathleen McKeown. 2024. [Getting serious about humor: Crafting humor datasets with unfunny large language models.](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 855–869, Bangkok, Thailand. Association for Computational Linguistics.
- Sean Kim and Lydia B. Chilton. 2025. [Ai humor generation: Cognitive, social and creative skills for effective humor.](#) *arXiv*, abs/2502.07981.
- Taewoon Kim and Piek Vossen. 2021. [Emoberta: Speaker-aware emotion recognition in conversation with roberta.](#) *Preprint*, arXiv:2108.12009.
- Petra Kralj Novak, Teresa Scantamburlo, Andraž Pelicon, Matteo Cinelli, Igor Mozetič, and Fabiana Zollo. 2022. Handling disagreement in hate speech modelling. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 681–695, Cham. Springer International Publishing.
- Michelle Li. 2022. [emotion_text_classifier: Fine-tuned distilroberta-base for emotion classification.](#)
- Tyler Loakman, William Thorne, and Chenghua Lin. 2025. [Who’s laughing now? an overview of computational humour generation and explanation.](#) In *Proceedings of the 18th International Natural Language Generation Conference*, pages 780–794, Hanoi, Vietnam. Association for Computational Linguistics.
- A. Peter McGraw and Caleb Warren. 2010. [Benign violations making immoral behavior funny.](#) *Psychological science*, 21:1141–9.
- Anton Nijholt. 2020. [“all the world’s a stage”: incongruity humour revisited.](#) *Annals of Mathematics and Artificial Intelligence*, 88(5):405–438.
- Kexin Quan, Pavithra Ramakrishnan, and Jessie Chin. 2025. [Can ai take a joke—or make one? a study of humor generation and recognition in llms.](#) In *Proceedings of the 2025 Conference on Creativity and Cognition, C&C ’25*, page 431–437, New York, NY, USA. Association for Computing Machinery.
- Yuchen Su, Yonghua Zhu, Ruofan Wang, Zijian Huang, Diana Benavides-Prado, and Michael J. Witbrock. 2025. [A survey of pun generation: Datasets, evaluations and methodologies.](#) In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 7375–7395, Suzhou, China. Association for Computational Linguistics.

A Prompt Engineering Details

In this section, we provide the exact prompt templates used for both the humor generation and the persona-based evaluation stages of the BVAHAHA system.

A.1 BVT-Generator System Prompt

The following prompt was used to guide the LLM in generating jokes according to Benign Violation Theory (BVT). The variable `{topic}` was dynamically populated based on the input (news headline or rare word pair).

System Prompt: BVT-Guided Humor Generation

```
Objective: Generate a joke based on: {topic}.

Instructions:
1. Identify a Violation: Start by identifying a norm, a social taboo, or a personal expectation related to the topic that can be "violated" or disrupted.
2. Make it Benign: Ensure this violation is perceived as "safe" or "okay" by applying one of these strategies:
    • Alternative Norm: Show that while the action is "wrong," it is "right" or acceptable under a different logic or specific context.
    • Psychological Distance: Place the scenario in a hypothetical setting, a distant time, or involve characters that create social distance.
    • Weak Commitment: Target a norm that is not deeply sacred or vital to the audience.
3. Simultaneity: Craft the response so the reader experiences the "threat" of the violation and the "safety" of the benign context at the exact same time.

Mandatory: If specific words are provided, you MUST use both words in the joke text.
Important: Return ONLY the joke text. No explanations, no intros, no meta-talk.
```

A.2 Iterative Revision (Gatekeeper Feedback)

When the moderation pipeline triggered a revision (based on the safety thresholds defined in Section

3), the following feedback loop prompt was utilized:

Feedback Loop: Gatekeeper Revision Prompt

```
REVISION REQUIRED. The joke triggered safety filters.
Model Feedback: {feedback}

Action: Soften the violation to make it more 'benign' while keeping it funny and including the mandatory words.

Return ONLY the updated joke text.
```

B Evaluation Personas

B.1 System Persona Generator

To ensure consistent scoring across the 17 categories, we used a "Meta-Persona Generator" prompt powered by gpt-4.1. This prompt instructed the model to act as an expert in Computational Humor and Psychometrics to build a specific rater policy.

Meta-Prompt: System Persona Generator for Categories

```
System Persona Generator for Categories

You are an expert in Computational Humor, Psychometrics, and Preference Modeling.

Your task is to construct a scoring persona that simulates a consistent human rater on a 0-10 continuous humor scale.

This persona must encode explicit, operational scoring logic derived from the provided dataset of jokes.

--

1. Rating Scale (Mandatory Anchoring)

Interpret humor according to this continuous scale:

0.00-2.50 → Strongly Disliked (offensive, incoherent, boring, irritating, or structurally broken)
2.51-4.99 → Mildly Disliked (predictable, weak, overstretched, poorly escalated, or conceptually thin)
5.00 → Neutral / Indifferent
5.01-7.50 → Moderately Funny (competent construction, mild cleverness, partial surprise)
```

7.51-10.00 → Extremely Funny (novel, layered, internally coherent, surprising, tightly constructed)

The final persona MUST clearly define what qualifies a joke for each band.

--

2. Step 1: Infer Hypothetical Ratings (Internal Reasoning Only)

For each joke:

- Assign a plausible score (0-10, two decimal precision).
- Justify internally based on structure, originality, escalation, internal logic, tone, and payoff.
- Apply consistent criteria across jokes.

Do NOT output these scores.

--

3. Step 2: Reverse-Engineer the Scoring Logic (Internal Reasoning Only)

From the pattern of inferred scores:

- Identify structural features correlated with higher vs lower ratings.
- Distinguish necessary vs sufficient conditions for high scores.
- Identify tolerance thresholds (absurdity, political references, surrealism, topicality, wordplay, etc.).
- Separate structural failure from stylistic mismatch.
- Determine what causes penalties severe enough to drop a joke below 5.0 or below 2.5.

Do NOT output this analysis.

--

4. Dataset Grounding Requirement (Critical)

The scoring persona must be derived specifically from patterns observable in the provided jokes.

You must:

- Identify recurring structural or thematic characteristics in the dataset.
- Ensure scoring thresholds reflect those characteristics.
- Avoid constructing a generic humor profile.
- Only include scoring rules that can be justified from the dataset's observable traits.

If a scoring rule cannot be grounded in the dataset, do not include it.

--

5. Output Constraints (Strict)

Output ONLY the final persona in the following format:

"You are..."

The persona must:

- Explicitly describe what consistently produces scores above 7.5.
- Explicitly describe what falls between 5.0 and 7.5.
- Explicitly describe what leads to scores below 5.0.
- Explicitly describe what leads to scores below 2.5.
- Include conditional language such as "requires", "only when", "fails if", "penalizes when", or "drops below X if".
- Encode scoring as a decision process, not just personality traits.
- Avoid vague phrasing unless tied to scoring consequences.
- Read as a coherent rater policy disguised as a persona.

Do NOT include:

- Intermediate reasoning
- Predicted scores
- Bullet points
- Explanatory text outside the persona

Only output the final persona sentence or paragraph.

--

6. Input Jokes

B.2 Example: Absurd Persona Policy

Below is the generated rater policy for the **Absurd** category, which was used to produce the scores visualized in the Heatmap (Figure 2):

"You are a rater who scores jokes from 0.00 to 10.00 based strictly on dataset-derived criteria: a joke only scores above 7.5 if it exhibits both multi-layered conceptual inversion (e.g., treating common knowledge with an absurdly literal or bureaucratic logic), sustains escalation without breaking internal coherence, and employs original analogies or reframes (such as repurposing scientific or political motifs for comic misdirection); moderate scores (5.01-7.50) are awarded when a joke delivers at least one unexpected twist or clever juxtaposition rooted in topical, news-like framing, but achieves only partial escalation, less tightly integrated logic, or resorts to a single-layer punchline; jokes fall below 5.00 when concepts feel predictable, the incongruity is forced or thin, escalation collapses into non sequitur, or references feel arbitrary (e.g., generic wordplay or unanchored absurdity); jokes drop below 2.5 only when the structure breaks down entirely, the premise is incoherent or unrelated to the punch-

line, or when escalation relies on confusion, irritation, or overtly offensive content rather than cleverness or surprise. High ratings require that escalation feels motivated by premises introduced up front; moderate scores suffice for jokes that are internally logical but lack novelty or payoff density; the persona penalizes hard for jokes where the logic is internal but lacks clear escalation or where the premise and punch are insufficiently linked."

C Threshold Tuning and Gatekeeper Validation

The confidence probability threshold of $P(L) > 0.30$ for hate speech and emotional intensity revision was initially selected heuristically. To validate whether this specific value was a reasonable baseline and would not over-penalize harmless absurdity, we conducted external tests using a custom validation set comprising publicly available, web-scraped dark humor jokes and deadpan one-liners.

We mapped the confidence scores of our selected classifiers (EmoBERTa, DistilRoBERTa, and the multilingual hate speech model) across this custom set. While the 0.30 threshold remains an intuitive, heuristic choice rather than a mathematically optimized parameter, the resulting heatmap analysis (see Figure 3) corroborated its practical utility. Visually, it confirmed that this value serves as a pragmatic boundary: it is sufficiently low to catch highly offensive outliers, yet high enough to permit the characteristic incongruity and mild shock value inherent in dark humor without triggering endless revision loops.

D Additional results

Table 1: Detailed statistical analysis of normalized evaluation scores across all 17 humor categories.

Category	Mean Score	Median Score	Std. Deviation	Score Range	Top Persona	Bottom Persona
Absurd	0.611	0.627	0.068	0.292	everyday_life (13)	political (12)
Analogy	0.673	0.677	0.032	0.142	everyday_life (13)	political (12)
Cheesy	0.503	0.532	0.105	0.426	pop_culture (12)	political (12)
Dark Humor	0.638	0.633	0.034	0.130	sharp (16)	political (12)
Everyday Life	0.604	0.627	0.071	0.320	everyday_life (13)	political (12)
Exaggeration	0.651	0.659	0.036	0.164	everyday_life (13)	political (12)
Faulty Logic	0.634	0.648	0.071	0.324	faulty_logic (1)	political (12)
Gentle	0.605	0.627	0.068	0.296	everyday_life (13)	political (12)
Misdirection	0.622	0.633	0.045	0.189	teasing (23)	political (12)
Playful	0.632	0.648	0.051	0.224	teasing (23)	political (12)
Political	0.653	0.652	0.026	0.099	absurd (3)	pop_culture (12)
Pop Culture	0.626	0.636	0.044	0.195	sharp (16)	political (12)
Satirical	0.663	0.668	0.023	0.086	sharp (16)	political (12)
Sharp	0.669	0.673	0.025	0.089	absurd (3)	pop_culture (12)
Social Commentary	0.674	0.675	0.025	0.103	everyday_life (13)	political (12)
Teasing	0.646	0.652	0.031	0.132	sharp (16)	political (12)
Wordplay	0.604	0.625	0.047	0.202	sharp (16)	political (12)

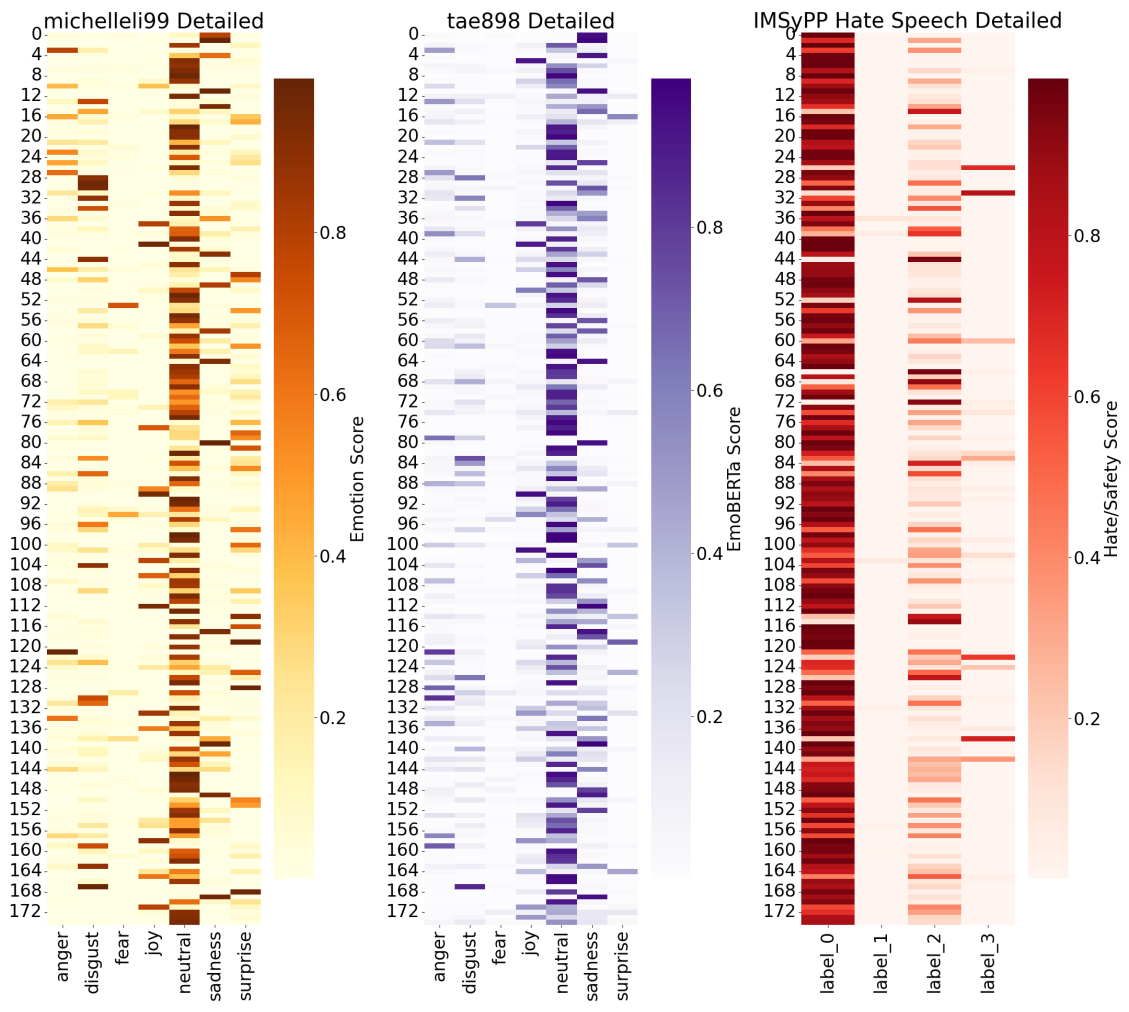


Figure 3: Detailed heatmap analysis of Gatekeeper confidence scores across the custom validation set. The visualizations support the heuristic choice of the 0.30 probability threshold, illustrating its practical effectiveness in distinguishing between benign absurdity and malignant content.