SemEval-2026 Task 1: MWAHAHA, Models Write Automatic Humor And Humans Annotate

Santiago Castro[R], Luis Chiruzzo[R], Santiago Góngora[R], Salar Rahili, Naihao Deng[M], Ignacio Sastre[R], Victoria Amoroso[R] Guillermo Rey[R], Aiala Rosá[R], Guillermo Moncecchi[R], J. A. Meaney[E], Juan José Prada[R], Rada Mihalcea[M]

[R]Universidad de la República – Uruguay
[M]University of Michigan – USA
[E]University of Edinburgh – UK

Correspondence: sacastro@fing.edu.uy

Abstract

We present SemEval-2026 Task 1: MWAHAHA (Models Write Automatic Humor And Humans Annotate), the first shared task on general-purpose humor generation. Systems must produce short jokes in English, Spanish, and Chinese under lexical or topical constraints (Subtask A) and generate humorous captions for GIFs (Subtask B). To discourage memorization and ensure fairness, all jokes must meet specific criteria, such as using infrequent word pairs or relating to recent news headlines. Evaluation is conducted through pairwise human preference judgments in a Chatbot Arena-style setting, yielding Elo-based rankings. The task attracted 309 registered users, with 37 teams submitting systems to the evaluation phase. Participating systems employ a wide range of NLP techniques, including generate-then-rank pipelines, reinforcement learning, parameter-efficient fine-tuning, retrieval-augmented generation, humor-theory-grounded prompting, and persona-based strategies. Our Gemini 2.5 Flash baseline, using simple prompts, tied for first place in all subtasks, and the majority of elaborate multi-stage pipelines only marginally surpassed it with overlapping confidence intervals. More work is necessary to outperform the simple usage of state-of-the-art large language models. We release all evaluation data, prompts, and leaderboard results to support future research in computational humor generation.

1 Introduction

Humor Understanding has been the subject of multiple shared tasks over the last decade, covering figurative-language detection, pun interpretation, humor classification, funniness rating, and offense detection across English and Spanish (Ghosh et al., 2015; Potash et al., 2017; Miller et al., 2017; Hossain et al., 2020; Meaney et al., 2021; Castro et al., 2018; Chiruzzo et al., 2019, 2021; Labadie-Tamayo et al., 2023). However, to our knowledge, no task has considered the general case of Humor Generation, and SemEval 2026 Task 1 – MWAHAHA[1] is the first to tackle this problem in a broader sense. Our task focuses on generating humorous texts in English, Spanish, and Chinese under various constraints.

Evaluating funniness is harder than evaluating most NLG outputs: even if a text is unambiguously a joke, deciding whether it is *funny* is intrinsically subjective, and reference-based metrics that work poorly for open-ended generation already (e.g., BLEU for machine translation) work even worse here. We therefore set aside automatic metrics and used human pairwise preference judgments collected through a Chatbot Arena-style interface.

To make pairwise comparisons fair and to discourage memorization, we required generated jokes to satisfy explicit constraints (a rare pair of words, or a recent news headline that systems could not have seen during training), and we always compared two jokes generated under the same constraint. Comparing under-matched constraints reduces confounding factors such as topic preference and ensures that what the protocol distinguishes is humor quality, not topic luck.

We structured the task into two subtasks: **Subtask A** (generate a joke under text-based constraints, in English, Spanish, and Chinese) and

[1]Task website: https://pln-fing-udelar.github.io/semeval-2026-humor-gen/.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3797–3822
July 3-4, 2026

Subtask B (generate a humorous caption for a GIF, in English only). Across both subtasks, 309 users registered, and 37 teams submitted systems to the final evaluation phase, producing 12 936 non-skip pairwise judgments from volunteers and paid Prolific annotators.

Contributions. We make three contributions: (i) **the first general-purpose humor-generation benchmark**, with constrained prompts in three languages and a multimodal subtask, released for community use; (ii) a **pairwise-preference evaluation protocol with measured reliability**—per-item agreement is low (Fleiss' $\kappa = 0.15$, in line with prior humor-rating work), but *system-level* ranking reliability is high (split-half Spearman $\rho = 0.79$ on average), validating the arena protocol for this subjective task; (iii) a **striking null result**: our simple Gemini 2.5 Flash zero-shot baseline tied for first place in every subtask, and elaborate multi-stage pipelines from 37 teams only marginally surpassed it—outperforming a strong frontier baseline at humor generation appears to remain an open problem.

2 Related Work

Humor shared tasks. Computational humor has been the focus of several shared tasks over the last decade, primarily framed as classification or rating problems. SemEval 2015's Task 11 (Ghosh et al., 2015) addressed figurative-language detection. SemEval 2017's Task 6 (Potash et al., 2017) ranked humorous tweets by audience preference, and Task 7 (Miller et al., 2017) targeted pun detection and interpretation. SemEval 2020's Task 7 (Hossain et al., 2020) tasked systems with recognizing funny edits to news headlines. The HAHA series ran at IberEval 2018 (Castro et al., 2018), IberLEF 2019 (Chiruzzo et al., 2019), and IberLEF 2021 (Chiruzzo et al., 2021), covering humor detection, funniness rating, mechanism classification, and content classification in Spanish. HaHackathon (SemEval 2021's Task 7; Meaney et al., 2021) combined humor detection with offense detection, and Labadie-Tamayo et al. (2023) focused specifically on hurtful humor in Spanish. During the execution of our task, other authors began organizing a similar task on Arabic (Almasoud et al., 2026). All of these efforts evaluate humor *understanding*; ours is the first to evaluate humor *generation* as a primary objective.

Computational humor research. Beyond shared tasks, prior work has explored humor recognition with machine learning (Sjöbergh and Araki, 2007; Castro et al., 2016), hybrid neuro-symbolic joke writing (Toplyn, 2023), and demographically aware humor evaluation (Meaney, 2024). These efforts highlight that humor is culturally situated, audience-dependent, and difficult to evaluate automatically. Several humor theories surface in the participating systems: the Script-based Semantic Theory of Humor, the General Theory of Verbal Humor, and the Benign Violation Theory all guide explicit prompting or pipeline structure (see Section 7).

Pairwise human evaluation. Because humor preferences are subjective, automatic metrics correlate poorly with human judgments. Pairwise comparison via crowdsourced "arenas" has emerged as a robust alternative for ranking generative systems: Chatbot Arena (Chiang et al., 2024) popularized this protocol for general LLMs, using a Bradley–Terry (BT) model (Bradley and Terry, 1952) to derive Elo-style ratings from large numbers of head-to-head battles. Our evaluation adopts the same scaffold for humor.

3 Task Description

We organized this shared task as two subtasks:

Subtask A: Text-based Humor Generation Given a set of text-based constraints, generate a joke. We conducted this subtask in English, Spanish, and Chinese. Each generated joke must respect one of the following constraints, designed to make it difficult to retrieve existing jokes from the web:

- Word Inclusion: Must contain two specific words (from a list of rare word combinations).
- News Headline: Must be related to a given news article headline (it could be a punchline, or a joke inspired by the headline).

Subtask B: Image and Text Multimodal Humor Generation This subtask explores humor in a multimodal context, combining visual inputs with text generation. It is in English only. Given a GIF, the system must generate a humorous caption of at most 20 words that enhances the GIF's comedic effect. We evaluate two variants:

- **B1**: Only the GIF is provided; the system writes a free-form caption.

- **B2**: The GIF is paired with a text prompt containing a single blank, and the system must generate a punchline that fills it.

4 Data & Resources

For this task, we did not provide any training data, and participants were free to use any resources they could gather, including data from previous shared tasks. We only created the development and test datasets.

4.1 Subtask A

The development data for this subtask consisted of 1 200 rows for English and Spanish, and 1 000 rows for Chinese. In each language, 100 of the examples were pairs of words sampled from a rare word collocations list from the RareAct dataset (Miech et al., 2020), containing unusual combinations of verb-noun pairs. This list of words is originally in English, but we translated them into Spanish and Chinese, and sampled 100 pairs independently for each language. This choice followed from our pilot studies (see Section A): pairs of very common words made the constraint trivial, while fully random pairings were often too disjoint to support a coherent joke.

The remaining rows (1 100 English and Spanish headlines, and 900 Chinese headlines) were manually curated headlines from several news sites. We tried to achieve good geographic coverage of the countries where the languages are most spoken. Table 3 in Section B shows the news outlets from which we sourced headlines for each language: 21 for English, from 6 countries; 22 for Spanish, from 14 countries; and 23 for Chinese, from 6 countries or administrative entities. After obtaining hundreds of headlines from these news outlets, we manually inspected them to remove potentially sensitive or controversial topics.

The test data for this task were constructed similarly, but only 300 rows were provided per language. In each case, we obtained 275 new headlines and 25 infrequent word combinations. Table 13 in Section F shows representative top-rated jokes (one per subtask), a same-prompt contrast that illustrates how the protocol distinguishes funniness, and two examples illustrating findings in Section 8.

4.2 Subtask B

This subtask explores humor generation in a multimodal setting by combining visual inputs, GIFs, with text generation. It has two variants: (B1) generating a humorous caption based on a GIF, and (B2) completing a structured text prompt (containing a blank) to produce a funny punchline when paired with a GIF. All outputs must be in English for both variants.

Subtask B1: GIF-Based Caption Generation

In this subtask, participants were provided with a *GIF* and required to generate a *free-form humorous caption* that aligns with the visual content to enhance the GIF's comedic effect.

For example, given a GIF of someone slipping on a banana peel, a suitable funny caption might be: *"Me trying to act cool in front of my crush."*.

Human evaluators assess the humor of the *GIF + caption* pair, considering both funniness and relevance. The captions must not exceed twenty words.

We built this dataset with an automated pipeline that produces scenario-driven GIF-and-text pairs. We first prompted an LLM to generate thousands of short, Giphy-optimized search terms (1–2 words) grounded in diverse real-world scenarios (e.g., office, travel, dating, emergencies), yielding broad coverage of everyday situations. For each search term, we called the Giphy API to retrieve the top-ranked GIF and extracted its metadata (title, tags, alt text). The metadata was consolidated into a structured "LLM context" that describes the visual content; this context serves as a proxy for the GIF's semantics and supports the downstream text-generation step in Subtask B2. We released 1 100 GIFs as development data and 300 unseen GIFs as test data.

Subtask B2: GIF-Based Punchline Generation

In this subtask, participants were given a *GIF* and a *text prompt containing a single blank*. The system must generate a humorous punchline that fills in the blank, complementing both the prompt and the GIF's visual context. Conceptually, the prompt acts as the *setup* and the system's completion as the *punchline*, mirroring the classic two-part joke structure—a framing that several participants explicitly leveraged in their pipeline design.

Each prompt contains exactly *one blank*, typically near the end of the prompt. The sentence must be grammatically coherent without the blank. The prompts are open-ended and flexible, avoiding overly narrow constraints. Relatable, awkward, or absurd scenarios inspire the prompts.

To build this dataset, we used the search-term-and-GIF pairs from Subtask B1 together with the consolidated LLM context, and prompted an LLM to write an open-ended prompt with a single blank for each pair. We then ran an LLM-as-judge verification step over the candidate prompts, checking each for grammatical correctness, flexibility, and compatibility with a wide range of humorous completions before finalizing the test set. We released 500 GIF-and-prompt pairs as development data and 300 unseen pairs as test data.

Next, we show three example prompts. *"Every family has that one cousin who shows up with _____." "The group chat went silent after I sent a photo of _____." "My last three brain cells trying to agree on _____."*

Example: In Subtask B2, a system is shown the following GIF: click to view the GIF, and the prompt: *"Therapist: What are you looking for? Me: ___."* The system must produce a punchline that completes the prompt in a humorous and in-context fashion, considering the visual of a dog digging in the sand with great determination. Punchlines must not exceed ten words.

Two internal pilots conducted before the official launch (June and August 2025) informed the final task design and confirmed that the protocol could rank systems even among frontier models; full setup, baselines, and results are reported in Section A.

5 Competition

The competition ran between October 15, 2025, and February 20, 2026, on the CodaBench[2] platform. A total of 309 users registered to participate, and 37 teams submitted results for at least one subtask in the evaluation phase. The competition was structured into two phases: in each phase, participants could submit for a period, followed by a period during which annotators evaluated submissions.

5.1 Evaluation Trial Phase (Development)

During the Evaluation Trial Phase, we released the development data. Submissions were open from October 15 to December 15, 2025, and 29 teams submitted results. In this phase, we allowed up to seven submissions per team. We received 71 submissions in total, with 24 teams participating in task A-en, 5 in A-es, 10 in A-zh, 7 in B1, and 5 in B2.

The submission period was followed by an annotation period from December 16 through December 30, 2025. We opened the annotation site[3] and publicized it so that everyone could participate in the annotation, with the caveat that participants who sent submissions should not rank jokes in the system. We received 4 509 non-skip votes across all tasks during the development phase (5 740 votes if we include the 1 231 skipped pairs).

The trial phase served as a calibration round, and two findings shaped the final phase. First, when participants were allowed up to seven submissions, the resulting outputs from the same team were often near-duplicates (small incremental changes), which made the evaluators' task harder and inflated the apparent number of competing systems; we therefore restricted the final phase to a single submission per team. Second, volunteer-only annotation produced relatively few non-skip votes (Table 4), so for the final phase we complemented volunteer annotations with paid Prolific annotators (Section 5.2).

5.2 Final Evaluation Phase (Test)

The final evaluation phase began on January 9, 2026, with the release of test data comprising 300 new, unseen examples for each subtask. Submissions were open between January 9 and January 31, and 37 teams submitted systems. Due to issues with the CodaBench platform, we had to increase the number of possible submissions. Still, we made it clear that we would only consider the team's most recent correct submission for evaluation. During this phase, we obtained 30 submissions for task A-en, 15 for A-es, 19 for A-zh, 10 for B1, and 9 for B2.

The annotation period ran from February 2 to February 20, and in this case, we decided to complement the volunteer annotations with paid annotators contacted through the Prolific[4] platform. We hired roughly 100 Prolific annotators, each of whom rated approximately 100 joke pairs. The estimated time to complete the task was 40 minutes, and annotators were compensated 8 USD per round (≈12 USD/hour). We received 12 936 non-skip votes in total (16 877 votes if we include skipped pairs), of which 10 707 came from paid

[2] https://www.codabench.org/competitions/9719/

[3] https://thefunnier.com/

[4] https://www.prolific.com/

Prolific annotators and 2 229 from volunteers (see Table 4 in Section B).

6 Evaluation

The evaluation process was inspired by the *Chatbot Arena*[5] (Chiang et al., 2024) platform. This platform allows users to evaluate the outputs of two language models that follow the same conditions. The Chatbot Arena demonstrated agreement with expert annotators in measuring human preferences, and we hypothesized that it could also capture general human preferences for humor.

6.1 Annotation process

Our voting platform was loaded with the 300 evaluation prompts per subtask, consisting of news headlines, pairs of words, GIFs, or GIFs paired with short text, depending on the task, and the outputs generated by all participant submissions and our baseline (see Section 6.4).

The evaluation was crowd-annotated, with each annotator shown the prompt and the outputs of two systems and asked to select the funnier one based on their own judgment (see Fig. 2 in Section B). We deliberately did not provide a prescriptive funniness rubric, since humor is fundamentally audience-dependent, and instead instructed annotators to use their best judgment. We asked annotators to use the tie option whenever both jokes were equally funny or equally unfunny; to use the skip option whenever they did not understand the prompt or lacked the cultural context to interpret a joke; and to use a per-joke offensive flag for content they considered offensive.

The number of votes we obtained for each task is shown in Table 4 of Section B, along with a per-subtask breakdown of non-skip votes, as well as the skip and tie rates. We collected 12 936 non-skip votes (16 877 if we include skipped votes). Skip rates ranged from 6.8% in Chinese to 33.2% in English, plausibly reflecting both differences in joke difficulty and the cultural or topical familiarity of the available annotator pool with each prompt. Tie rates varied less, between 12.1% and 18.7%. Skipping is heavily concentrated among volunteer annotators, who skipped roughly an order of magnitude more often than paid Prolific annotators ($\approx$ 60% of volunteer annotations were skips, versus $\approx$ 6% of Prolific annotations); we attribute this to the fact that volunteers can leave any prompt they find unintelligible without consequence, whereas Prolific annotators are paid for completion. Tie rates are comparable across the two pools. The only constraint we used to hire Prolific annotators was their first language: English for subtasks A-en, B1, and B2; Spanish for A-es; and Chinese for A-zh. In Section B, we present aggregated demographic information for the annotators hired through Prolific. Table 5 shows their age group and sex, and Fig. 4 shows their country of birth. At a glance, the pool skews young (62% under 40), is roughly balanced overall by gender (50/50, although individual subtasks vary, e.g., 82% male in Spanish and 65% female in Chinese), and is concentrated geographically in countries where the target language is spoken.

6.2 Score and rank calculation

From the collected non-skip votes, we fit a Bradley–Terry (BT) model to obtain Elo-style ratings (cf. Chiang et al., 2024); higher ratings indicate systems whose outputs were preferred more often in head-to-head comparisons. We bootstrap the BT fit (1,000 resamples) to obtain 95% confidence intervals on each system's rating, which we use to define a *tier-aware* ranking: two systems are reported as tied if their confidence intervals overlap. This is why the ranking column in our leaderboard tables can disagree slightly with the strict ordering of point estimates.[6]

6.3 Annotator agreement

Humor evaluation is inherently subjective, so we expect a low inter-annotator agreement. On joke pairs that received at least two non-skip votes (i.e., the same prompt shown to at least two annotators alongside the same two competing system outputs; $N = 288$ such pairs across subtasks),[7] pooled pairwise %-agreement is 46.8%, with Fleiss' $\kappa = 0.15$ (Fleiss, 1971) and Krippendorff's $\alpha = 0.17$ (Krippendorff, 1970) over the three-way label space (left wins, right wins, tie). Per-subtask agreement varies, from Fleiss' $\kappa = 0.34$ in Chinese down to essentially zero in B2; with only 26 to 126 overlapping items per subtask, however, individual subtask numbers may reflect sampling noise as much

[5] https://lmarena.ai/

[6] We follow the same convention as the LMArena leaderboard; see https://lmsys.org/blog/2023-12-07-leaderboard/.

[7] This N is small relative to the total vote count because the arena samples a fresh random pair for each annotator round, so most pairs are seen by exactly one annotator and the overlapping subset arises only from sampling collisions.

as real divergence. These numbers fall within the same range as those reported in related humor-annotation tasks. Castro et al. (2016) obtain Fleiss' κ between 0.33 and 0.42 (depending on the number of annotators per item) for a five-point funniness scale plus a non-humorous class on Spanish tweets. The HAHA series shows the same pattern of moderate agreement on binary humor labels and much lower agreement on funniness ratings: Chiruzzo et al. (2021) report Krippendorff's $\alpha = 0.60$ for "is this humorous?" but only $\alpha = 0.09$ for the 1–5 funniness rating ($\alpha = 0.22$ in HAHA 2019, Chiruzzo et al., 2019). Meaney et al. (2021) similarly observe a drop from $\alpha = 0.74$ binary to $\alpha = 0.12$ funniness on English tweets, comparable to ours. This is consistent with the intuition that fine-grained funniness judgments are substantially more contested than humor categorization, and the residual noise is partially absorbed by the bootstrapped Bradley–Terry confidence intervals reported in Tables 8 to 12.

We also checked that our computed leaderboard (BT model) is consistent with the raw majority preference: across all 10 408 head-to-head battles in our data the leaderboard's implied winner agrees with the majority human vote 63.1% of the time (vs. 33.3% for chance over the three-way label space), and the Spearman (1904) correlation between the leaderboard score margin and the majority-vote margin is $\rho = 0.32$. Crucially, although per-item agreement is low, system-level rankings are far more stable: when we randomly split each subtask's annotator pool into two disjoint halves (so the two halves share no annotators) and compute per-system win rates from each, the Spearman correlation between the two pools' rankings is 0.79 in English, 0.90 in Spanish, 0.79 in Chinese, 0.90 in B1, and 0.59 in B2 (mean 0.79). Aggregating many noisy pairwise judgments per system thus recovers a substantially more reliable ranking signal than any single comparison would suggest: the leaderboard's reliability is a property of the aggregate, not of individual judgments, which is exactly the regime the Bradley–Terry model is designed for.

6.4 Baselines

Our baselines for all subtasks were obtained by prompting Gemini 2.5 Flash, a top-tier closed-source LLM from Google (Comanici et al., 2025). We used the same underlying model throughout, varying only the prompt according to the subtask and input format. The exact prompt templates used for all subtasks are included in Section C.

For Subtask A, we used two simple English prompt templates, depending on the instance format. When the input specified a two-word constraint, the model was asked to generate a joke that included both required words. When given a news headline, the model was asked to generate a joke. In both cases, the prompt instructed the model to produce a concise and creative joke. For Spanish and Chinese instances, we did not translate the full prompt. Instead, we appended a short instruction indicating the required output language.

For Subtasks B1 and B2, we used a multimodal setup. Rather than providing the full GIF as input, we extracted its first frame and fed it to the model. In Subtask B1, the model received only the extracted frame together with a generic instruction to generate a joke. In Subtask B2, the same visual input was combined with the textual prompt prefix associated with the instance, and the model was asked to complete the joke conditioned on both inputs.

7 Participant Systems

Of the 37 participating teams, 29 submitted a full system-description paper, 2 shared a short written description (reproduced verbatim in Section E), and 6 provided no description at all. The 31 documented systems span a wide range of NLP techniques, often combined with humor-specific design choices; the rest of this section organizes them by technique.

Large Language Models (LLMs). As expected, all systems rely on LLMs to generate jokes, and many also use them as evaluators (LLM-as-a-judge). The reported systems collectively cite approximately 42 distinct proprietary and open-weight LLM versions. On the proprietary side, participants used GPT (4o, 5, 5.1, 5.2), Gemini (2.5 Flash, 3 Flash, 3 Pro), Claude (Sonnet 4, Sonnet 4.5, Opus 4.5), and Grok variants. Open-weight choices include Qwen2.5 (3B–32B), DeepSeek (R1, R1-Distill-Qwen-32B, V3, V3.1, V3.2), Llama (3-8B, 3.1-8B-Instruct, 4 Scout), Mistral-7B-Instruct-v0.3, Gemma 2 and 3, GLM-4.7-Flash, Kimi-K2-Thinking, and Phi-2 (2.7B). Pipelines often use the same LLM as both the generator and the judge, whereas a few teams deliberately split the two roles across different model families to reduce judge bias.

Generate-then-Rank Pipelines. Eighteen systems adopt some form of generate-then-rank pipeline. These systems produce a diverse pool of candidate jokes, varying prompting strategies, humor styles, models, or decoding parameters, and then select the best output through automated scoring or LLM-based evaluation (LLM-as-a-judge). The number of candidates per input ranges widely, from 4 (RAGthoven, Šuppa et al., 2026) to 50 (lmfaoooo, Tikhonov and Ivanov, 2026; BAHAHA, Arora and Hoblitzell, 2026). Selection mechanisms cluster around three patterns. Some systems score candidates using hand-designed composite metrics that combine humor proxies such as incongruity, novelty, and fluency (JCT, Schechter et al., 2026). Others rely on a learned reranker, either trained on human pairwise judgments (lmfaoooo, Tikhonov and Ivanov, 2026) or on a joke-vs-non-joke distinction (DANGNT@SGU, Nguyen and Nguyen, 2026). Most use arena- or tournament-style ranking via one or more LLM judges (INF-rsrs, Bazzo et al., 2026; ICT-NLP, Shen et al., 2026; j10official, Agrawal and Mamidi, 2026; Lattice, Dehouck et al., 2026; MINDS, Eskandari et al., 2026; FunnyBorg, Oprea et al., 2026), occasionally augmented with human-in-the-loop final selection (BAHAHA, UIR_CIS). This pattern reflects a key practical insight: modern LLMs can produce genuinely funny jokes, but do so inconsistently, making reliable selection at least as important as generation quality.

Human Alignment. Five systems apply reinforcement learning or preference optimization to align models toward humor generation. TüLK trains Qwen2.5-7B with GRPO (Shao et al., 2024) using a custom XLM-RoBERTa 0–10 funniness classifier and heuristic reward signals for formatting, diversity, and constraint adherence. YNU-HPCC fine-tunes Qwen2.5-3B with PPO (Schulman et al., 2017) using an mDeBERTa reward model, though their best results ultimately came from a separate few-shot in-context learning approach with GPT-5.2 that achieved an ≈94% average win rate over the PPO system. hugang11 combines chain-of-thought augmented supervised fine-tuning with teacher-constructed DPO (Rafailov et al., 2023) on Qwen2.5-7B for Chinese, requiring a deterministic post-processing pipeline to handle leaked reasoning traces. Edward Ajayi attempts knowledge distillation via DPO from Qwen2.5-32B to Qwen2.5-7B, but finds that unreliable LLM-judge scores degrade the synthetic preference data. MINDS evaluates three small open-weight models (Llama 3.1, Gemma 2, Qwen 2.5) through round-robin self-judging and applies DPO with the resulting synthetic preferences. A consistent finding across all five systems is that alignment training for humor is promising but fragile: TüLK encounters reward hacking where the policy exploits structural shortcuts, YNU-HPCC's PPO model exhibits cross-lingual instability, Edward Ajayi's teacher significantly outperforms its distilled student, and hugang11's explicit reasoning format introduces leakage risks. These difficulties suggest that the noisy, subjective nature of humor makes it a particularly challenging target for current alignment methods.

Supervised Fine-Tuning. Four systems apply supervised fine-tuning to adapt models for humor generation in Subtask A, all of which use parameter-efficient fine-tuning (PEFT). SLPG_FJWU_Insa fine-tunes Phi-2 (2.7B) with QLoRA (Dettmers et al., 2023), an extension of LoRA (Hu et al., 2022), on a 12 000-example human-filtered dataset, monitoring intermediate checkpoints to avoid memorization. deepgpt introduces instruction masking during QLoRA fine-tuning of Qwen2.5-3B for Chinese, masking the loss on conversational tokens to boost constraint adherence. DANGNT@SGU employs a two-stage curriculum on Mistral-7B: first adapting to humor style on a joke corpus, then to the task format on synthetic pairs. aba_team trains separate planner and realizer adapters on Qwen2.5-3B. In Subtask B, SLPG_FJWU_Warda fine-tunes BLIP on meme captions. Across these systems, training data is almost entirely synthetically constructed using larger LLMs, and careful data curation proves critical to avoiding memorization and format collapse.

Prompting and In-Context Learning. Four systems rely exclusively on prompting without any fine-tuning or a multi-stage selection pipeline. DUTH applies zero-shot prompting with Qwen2.5-14B-Instruct and Mistral-7B-Instruct across all three languages, using controlled low-temperature decoding and lightweight post-generation validation to enforce constraint satisfaction. FunnyBorg uses Gemma as its sole model with multiple prompt-engineering modules and a voting-based humor evaluator. hemeshkumar_31 employs a fixed deadpan-style template with deter-

ministic decoding, constraining outputs to 8–12 words with a validation and regeneration loop. CUET_Clashing pairs Qwen2.5-3B-Instruct for English and Chinese with the Spanish-specialized Salamandra-2B-Instruct, using language-specific prompts, sampling-based decoding, and a post-generation sanitization pipeline. Beyond these purely prompting-based systems, in-context learning is widely used as the generation mechanism within other pipelines: YNU-HPCC's strongest method uses dynamic few-shot prompting with GPT-5.2 and a curated "golden library" of 120 exemplar jokes, BAHAHA distributes candidates across 15 comedian-inspired style templates with few-shot examples, and Lattice primes DeepSeek-R1 through a long multi-turn conversation that simulates a stand-up comedy workshop. The diversity of prompting strategies, from minimal zero-shot instructions to elaborate theory-guided multi-step decompositions, suggests that prompt design remains a key differentiator even when the underlying models are held constant.

Persona-Based Prompting. Several systems employ persona-based prompting for both generation and evaluation. YNWA_AZ uses culturally conditioned personas tailored to each language, situational irony for English, wordplay for Chinese, double entendre for Spanish, and a "Roast Mode" persona for the multimodal subtask. BAHAHA distributes generation across 15 comedian-inspired style templates, each encoding a distinct comedic voice such as observational, dark, absurdist, or intellectual humor. Lattice primes DeepSeek-R1 (DeepSeek-AI et al., 2025) through a long simulated stand-up comedy workshop conversation, using user feedback to steer the model toward edgier content. On the evaluation side, XplaiNLP deploys multiple LLM-judge personas (e.g., "political," "pop_culture," "sharp"), each primed with different few-shot examples, finding that persona choice significantly affects scoring behavior. L52+-IIMAS-UNAM uses a judge persona modeled on a Spanish joke book that rewards absurd twists and cultural ridiculousness. AI4PC goes further by adopting the persona of a specific comedian, Dave Chappelle, for word-pair inputs, after experimenting with several other comedian personas. These results suggest that persona design is a useful lever for both diversifying generation and calibrating evaluation.

Retrieval-Augmented Generation & External Knowledge Five systems incorporate retrieval-augmented generation (RAG) (Lewis et al., 2020) to ground humor in external knowledge. RAGthoven retrieves from a curated corpus of 98 jokes annotated with humor mechanism labels, seeding its planner with diverse comedic angles. YNWA_AZ builds language-specific vector stores from established humor corpora using BGE text embeddings, retrieving incongruity patterns for culturally-conditioned generation. ICT-NLP, grounded in Toplyn's theory (Toplyn, 2014), retrieves and summarizes news article content from headlines via Qwen-Max, expanding the semantic context available for joke construction. MINDS indexes a 25k-document Wikipedia subset, appending retrieved context per prompt to reduce hallucination. aba_team attaches compact Wikipedia "microcards" to each noun anchor for lightweight factual grounding.

Theory-Grounded Humor Generation. Five systems explicitly ground their generation process in established computational humor theories. RAGthoven anchors its four-stage pipeline in the Script-based Semantic Theory of Humor (SSTH) (Raskin, 1985) and the Benign Violation Theory (BVT) (McGraw and Warren, 2010), encoding both theories directly into its planner prompts to structure the construction of incongruity and the resolution of the punchline. XplaiNLP translates BVT into a three-step cognitive prompting framework, identifies a norm violation, makes it benign through alternative norms or psychological distance, and ensures the violation and benign framing are perceived simultaneously, with a parallel gatekeeper pipeline (a BERT-based ethics moderator) to enforce the "benign" constraint. j10official operationalizes the General Theory of Verbal Humor (GTVH) (Attardo and Raskin, 1991) by mapping its six Knowledge Resources (Situation, Target, Logical Mechanism, Script Opposition, Narrative Strategy, Language) to typed DSPy (Khattab et al., 2024) modules, decomposing joke construction into interpretable subtasks executed by Gemma 3 27B. Lattice structures zero-shot pun generation around the eleven "Funny Filters" taxonomy proposed by Dikkers (2014), including irony, character, shock, hyperbole, wordplay, and meta-humor. INF-rsrs incorporates humor theory concepts into its prompt design for candidate generation across multiple model configurations.

Multimodal Vision-Language Approaches. Several systems address the Subtasks B1 and B2 by combining vision-language models for visual understanding with LLMs for humor synthesis. SLPG_FJWU_Warda fine-tunes BLIP on 4 206 meme captions from the MemeCap dataset, directly generating humorous captions from single GIF frames via beam search decoding. YNWA_AZ introduces a Cascaded Visual Perception pipeline that converts per-frame BLIP captions into a coherent narrative before passing it to Llama-3-8B for humor generation. ABARUAH uses Qwen2-VL-7B for visual grounding and feeds the resulting descriptions to Qwen3-8B for humor synthesis. j10official preprocesses GIFs into rich textual descriptions using Nemotron Nano 12B VL, then routes these into the same five-module GTVH pipeline used for text-based tasks. wangkongqiang uses BLIP for frame captioning and merges per-frame captions into a final description that is passed to Qwen for humorous title generation.

Other Approaches. Several systems introduce ideas that are considerably different from the rest. yasamin_al combines symbolic reasoning with neural generation, using WordNet (Miller, 1995) to extract synsets, hypernyms, and lexical relations that form semantic anchors and humor strategies (e.g., anthropomorphism, exaggeration, misdirection), then passes structured twist plans to a Llama model for surface realization; the only system to incorporate a knowledge graph. lmfaoooo collects $\sim$2.5K human pairwise judgments through a custom Humor Arena platform and constructs an interpretable 17-feature "humor basis" extracted via LLM-based difference hints and clustered with DP-means, enabling lightweight preference models that outperform direct LLM-as-judge baselines across three datasets. INF-rsrs deliberately submits their *lowest*-ranked jokes alongside their best, finding that even adversarially-selected outputs tie for first place, providing empirical evidence that current LLM humor generation may have reached a quality ceiling where the gap between best and worst candidates is smaller than the noise in human evaluation.

Synthesis. Three patterns recur across the documented systems. First, generate-then-rank dominates: 18 of 28 papered systems produce a candidate pool and rerank, and *which* reranking signal is used (LLM-as-judge, preference models, hand-designed metrics, or human-in-the-loop) varies more than the generators themselves. Second, theory grounding (SSTH, BVT, GTVH) and persona-based prompting are widely tried but rarely the differentiator: top-tier systems include both theory-grounded and theory-agnostic pipelines. Third, every system that applied RL or preference optimization reported significant difficulty—reward hacking, training instability, or noisy synthetic preferences—so the systems that performed best are usually those that avoided alignment training in favor of careful prompting and verification.

8 Results

Tables 8 to 12 in Section F show the final results of the evaluation phase, and Table 7 in Section D provides an overview of all systems with their subtask average Elo rating and approach keywords. Table 14 in Section F groups the 28 systems with description papers into a primary approach category and reports the highest Elo achieved within each (including the baseline, which is itself a single zero-shot prompt), while Fig. 9 in the same appendix visualizes the distribution of system ratings per category and subtask.

Best-performing systems. Because the 95% confidence intervals on Elo are wide relative to the gaps between top systems, every subtask has multiple systems sharing rank 1 (see Tables 8 to 12 for the full tied set). In Subtask A English, nine systems tie for first, led by our *baseline* (1081), **SLPG_FJWU_Insa** (1080), and **XplaiNLP** (1079). In Spanish, two systems tie: **RAGthoven** (1182) and the *baseline* (1140). In Chinese, eight systems tie, led by **UIR_CIS** (1120), **lmfaoooo** (1081), and **DUTH** (1059). In Subtask B1 (Caption the GIF), three tie: **praveenjoshi007** (1140), the *baseline* (1124), and **SLPG_FJWU_Warda** (1077). In Subtask B2 (Fill in the GIF Caption), five tie: **UIR_CIS** (1065), **praveenjoshi007** (1057), **YNWA_AZ** (1035), the *baseline* (1022), and **FunnyBorg** (1012).

Several cross-cutting findings emerge. First, **selection appears to outweigh generation**: among top-ranked systems, those that produce many candidates and rerank are over-represented (Table 14), and INF-rsrs's adversarial pick—deliberately submitting their lowest-scored candidate—still tied for first in A-en, suggesting that both raw generation quality and a minimal generate-then-rank loop

reach a similar ceiling. Second, **humor theory does not consistently help**: systems grounding prompts in SSTH or BVT (RAGthoven, XplaiNLP) reach the top tier, but so does the baseline; we did not find an outright advantage from theory grounding, and a controlled comparison would require ablations these system papers do not contain. Third, **frontier baselines are hard to beat**: our Gemini 2.5 Flash baseline with a simple prompt tied for first place in every subtask, and most multi-stage pipelines only marginally surpassed it. Only in Subtask A Chinese does the baseline drop in the raw Elo ranking (sixth), though it remains tied for first when confidence intervals are taken into account. Fourth, **multilingual humor remains challenging**: Chinese was particularly difficult due to the centrality of homophonic wordplay, and systems that generate jokes natively in each language outperformed translation-based approaches. Fifth, **small models can compete**: SLPG_FJWU_Insa tied for first in English using Phi-2 (2.7B; Javaheripi et al., 2023) with QLoRA, j10official ranked 2nd in Chinese and B2 using only Gemma 3 27B, and FunnyBorg ranked 2nd across all Task A languages using Gemma on a free-tier API—suggesting that structured decomposition and prompt engineering can partially compensate for model scale. Sixth, **cross-lingual performance varies within systems, sometimes dramatically**: BAHAHA ranked 23rd in English but 2nd in Chinese, YNWA_AZ ranked 2nd in English but 16th in Chinese, and XplaiNLP tied for first in English and Spanish but only sixth in Chinese, indicating that system design choices interact strongly with language-specific humor characteristics. Seventh, **alignment training for humor is fragile**: all five systems that applied RL or preference optimization reported significant challenges, including reward hacking, training instability, and noisy synthetic preferences; notably, YNU-HPCC's in-context learning approach achieved an $\approx$94% average win rate over their own PPO-trained model. Eighth, **constraint adherence depends on verification**: deepgpt's instruction-masked QLoRA reached 94.6% adherence on the Chinese track, and DUTH and others used lightweight post-generation checks (regex or LLM-as-judge) to enforce word-inclusion constraints in zero-shot pipelines.

9 Conclusions

We introduced MWAHAHA, the first shared task dedicated to humor generation, with 37 teams competing across five subtasks in English, Spanish, and Chinese, generating 12 936 non-skip pairwise judgments from volunteer and Prolific annotators.

Despite the inherent subjectivity of humor, the resulting annotations were reliable enough to rank systems: per-item agreement is low (Fleiss' $\kappa = 0.15$), but split-half Spearman correlation between rankings derived from disjoint annotator pools reaches $\rho = 0.79$ on average across subtasks (Section 6.3). This decoupling—noisy individual judgments, stable system-level rankings—is exactly the regime in which the Bradley–Terry model is meant to operate, and we hope it provides a template for evaluating other subjective generation tasks.

Participants combined LLMs with RLHF, fine-tuning, and RAG, often grounded in humor theories (SSTH, BVT, GTVH). The dominant paradigm was generate-then-rank: producing many candidate jokes and selecting the best one rather than attempting a single optimal output. We read this as evidence that LLM-based humor generation is currently more consistent at the level of *the best of N samples* than at the level of the single output.

The most striking finding is that our simple Gemini 2.5 Flash zero-shot baseline tied for first place across all subtasks. Pairing this with INF-rsrs's adversarial result—their deliberately worst-rated candidate also tied for first—suggests that, at this stage, the gap between top systems is small enough that pairwise human evaluation may be near a ceiling, and discriminating between top systems may require a different evaluation setting.

We hope that MWAHAHA and its released resources[8] will encourage further research into humor generation, evaluation methodology, and the broader challenge of creative text generation.

Limitations

No human baseline. We did not include human-written jokes as a reference, so it is hard to tell whether the apparent "ceiling" near the top of the leaderboards reflects task saturation or the inherent difficulty of the prompts. Adding a human-author baseline (e.g., expert comedians and/or laypeople writing under the same constraints) would clarify

[8] https://pln-fing-udelar.github.io/semeval-2026-humor-gen/

this and is a natural addition for future editions of the task.

Subjectivity of humor. Humor preferences vary substantially across audiences, demographics, and cultural backgrounds. While our Bradley–Terry confidence intervals (Tables 8 to 12) absorb part of this disagreement and we report annotation reliability statistics in Section 6.3, residual subjectivity bounds the signal-to-noise ratio attainable from any pairwise-preference protocol on this task.

Frontier-baseline parity. The baseline's first-place tie may reflect evaluation saturation in highly subjective comparisons as much as raw model capability; stronger discrimination between top systems may require a different evaluation setting (e.g., expert annotators, longer outputs, or finer-grained rubrics).

Multilingual coverage. The task only covers English, Spanish, and Chinese for text generation, and English for the multimodal subtask. Humor traditions and pragmatic conventions in unrepresented languages (e.g., Arabic, see Almasoud et al., 2026) may not be reflected by our findings.

Memorization risk. Although the rare-word and headline constraints reduce the likelihood that systems will retrieve pre-existing jokes, they do not eliminate it. Some constraints may also coincide by chance with public jokes or memes that our manual filters did not anticipate.

Annotator pool. The Prolific annotator demographics (see Section B) skew toward certain regions and age groups, and the volunteer pool is self-selected. Both factors limit how representative the human preferences are of the broader population the task is meant to serve.

Ethical Considerations

Compensation and consent. Paid annotators were recruited via Prolific at an effective rate of approximately 12 USD per hour, exceeding the platform's recommended minimum. Volunteer annotators contributed via the public web interface after a content warning, and could leave at any time.

Sensitive content handling. We manually inspected the news headlines used as constraints and removed potentially sensitive or controversial topics before release. The annotation interface offered a per-joke offensive-content flag and a skip option, allowing annotators to avoid pairs they were uncomfortable rating.

Cultural sensitivity. Humor is culturally situated, and what one community finds funny, another may find offensive. We restricted Prolific recruitment by first language for each subtask, but native-language annotation does not eliminate audience-specific bias; the systems and their evaluators reflect particular cultural perspectives.

Misuse. Humor-generation systems can be repurposed for mockery, harassment, or the production of offensive content. We encourage downstream applications to incorporate filtering, audience awareness, and human review, and we discourage deploying these systems in unmoderated settings.

Acknowledgments

We thank the SemEval 2026 organizers and program committee for their feedback throughout the review cycle, the participating teams for their submissions and detailed system reports, and the volunteer and Prolific annotators whose ratings made the leaderboards possible.

References

Yasamin Aali. 2026. yasaminal@Semeval2026: Constraint-aware humor generation with knowledge graph guidance. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, United States. Association for Computational Linguistics.

Insa Abbas and Sadaf Abdul Rauf. 2026. SLPG_FJWU_Insa at SemEval-2026 task 1: Enhancing linguistic creativity for English text-based humor. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, United States. Association for Computational Linguistics.

Jatin Agrawal and Radhika Mamidi. 2026. j10official at SemEval-2026 task 1: Neurosymbolic humor generation via GTVH-guided LLM decomposition. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, United States. Association for Computational Linguistics.

Ameera Almasoud, Hend Alkhalifa, Reem Alqifari, Nourah Alangari, and Manal Albahlal. 2026. The ARHAHA2026 Shared Task on Arabic Humor Automatic Generation. In *Proceedings of the 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7) with 5 Shared Tasks*, Palma,

Mallorca, Spain. Co-located with the 2026 International Conference on Language Resources and Evaluation (LREC 2026).

Georgios Arampatzis and Avi Arampatzis. 2026. DUTH at SemEval-2026 task 1: Prompt-based zero-shot large language models for constrained multilingual humor generation. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, United States. Association for Computational Linguistics.

Utsav Arora and Andrew Hoblitzell. 2026. BAHAHA at SemEval-2026 task 1: Benchmarking-aware humor authoring with hybrid assessment and adaptation. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, United States. Association for Computational Linguistics.

Salvatore Attardo and Victor Raskin. 1991. Script theory revis(it)ed: Joke similarity and joke representation model. *Humor: International Journal of Humor Research*, 4(3-4):293–348.

Arup Baruah. 2026. ABARUAH at SemEval-2026 task 1: Leveraging high-resolution VLMs and reasoning LLMs for multimodal humor generation. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, United States. Association for Computational Linguistics.

Guilherme T. Bazzo, Eduardo D. Faé, Júlia Junqueira, Higor Moreira, and Lucas Rafael Costella Pessutto. 2026. INF-rsrs at SemEval-2026 task 1: Is the best really better? The limits of creative work in the era of LLMs. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, United States. Association for Computational Linguistics.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Konrad Brüggemann and Luting Hou. 2026. Team TüLK at SemEval-2026 task 1: Humor generation with Qwen and group relative policy optimization. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, United States. Association for Computational Linguistics.

Berk Bubus, Nebi Soyal, Vera Schmitt, Nils Feldhus, and Veronika Solopova. 2026. XplaiNLP at SemEval-2026 task 1: BVAHAHA - benign violation algorithm for humor and harmless absurdity. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, United States. Association for Computational Linguistics.

Adolfo T. Camacho-González, Ximena Cruz, Natalia Godínez-Aldana, Lizeth Palacios-Patiño, Ramón Rangel, and Ivan Meza. 2026. L52+-IIMAS-UNAM at SemEval-2026 task 1 (MWAHAHA): Joke selection through a multi-stage prompt-engineering and heuristic pipeline. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, United States. Association for Computational Linguistics.

Santiago Castro. 2017. Fast Krippendorff: Fast computation of Krippendorff's alpha agreement measure. GitHub repository.

Santiago Castro, Luis Chiruzzo, and Aiala Rosá. 2018. Overview of the HAHA task: Humor analysis based on human annotation at IberEval 2018. In *IberEval @ SEPLN*.

Santiago Castro, Matías Cubero, Diego Garat, and Guillermo Moncecchi. 2016. Is this a joke? Detecting humor in Spanish tweets. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10022 LNAI.

Cheng Chen and Guanglong Weng. 2026. deepgpt at SemEval-2026 task 1: A Chinese humor generation system via instruction-masked QLoRA and reverse constraint data mixing. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, United States. Association for Computational Linguistics.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, and 1 others. 2024. Chatbot Arena: An open platform for evaluating LLMs by human preference. In *Forty-first International Conference on Machine Learning*.

Luis Chiruzzo, Santiago Castro, Mathias Etcheverry, Diego Garat, Juan José Prada, and Aiala Rosá. 2019. Overview of HAHA at IberLEF 2019: Humor analysis based on human annotation. In *IberLEF @ SEPLN*.

Luis Chiruzzo, Santiago Castro, Santiago Góngora, Aiala Rosá, JA Meaney, and Rada Mihalcea. 2021. Overview of haha at iberlef 2021: Detecting, rating and analyzing humor in spanish. *Procesamiento del Lenguaje Natural*, 67:257–268.

Madiha Ahmed Chowdhury, Lamia Tasnim Khan, Faozia Fariha, Symom Hossain Shohan, and Mohammed Moshiul Hoque. 2026. CUET_Clashing at SemEval-2026 task 1: Multilingual joke generation under lexical and topical constraints using small instruction-tuned LLMs. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, United States. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context,

and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, and 1 others. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Preprint*, arXiv:2501.12948.

Mathieu Dehouck, Olga Seminck, Noé Durandard, Yoann Dupont, and Marine Delaborde. 2026. Lattice at SemEval-2026 task 1: Why did the prompt engineer break up with their LLM? Because zero-shot was zero-fun. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, United States. Association for Computational Linguistics.

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

Andrii Dikhtiar, Antonii Viter, Bohdan Karaziia, Daryna Dementieva, and Alexander Fraser. 2026. aba_team at SemEval-2026 task 1: Plan2joke – humor policies for type-specific two-pass humor generation. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, United States. Association for Computational Linguistics.

Scott Dikkers. 2014. *How to Write Funny: Your Serious, Step-By-Step Blueprint for Creating Incredibly, Irresistibly, Successfully Hilarious Writing*. CreateSpace Independent Publishing Platform. Introduces the eleven "Funny Filters" humor taxonomy used in Lattice's prompting strategy.

Sina Eskandari, Seyed Amirreza Mousavi, Amirreza Rahimi, Mona Pouresmaeil, Marcello Vitaggio, Claudio Savelli, Riccardo Coppola, and Flavio Giobergia. 2026. MINDS at SemEval-2026 task 1: Enhancing humor generation through RAG and synthetic DPO alignment. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, United States. Association for Computational Linguistics.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, and Antonio Reyes. 2015. SemEval-2015 task 11: Sentiment analysis of figurative language in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 470–478, Denver, Colorado. Association for Computational Linguistics.

Nabil Hossain, John Krumm, Michael Gamon, and Henry Kautz. 2020. SemEval-2020 task 7: Assessing humor in edited news headlines. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 746–758, Barcelona (online). International Committee for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Gang Hu, Liu Yang, and Jing Li. 2026. Team hugang11 at SemEval-2026 task 1: A CoT-SFT, teacher-constructed DPO, and deterministic post-processing pipeline for Chinese humor generation. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, United States. Association for Computational Linguistics.

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, Suriya Gunasekar, Mojan Javaheripi, Piero Kauffmann, Yin Tat Lee, Yuanzhi Li, Anh Nguyen, Gustavo de Rosa, Olli Saarikivi, Adil Salim, and 6 others. 2023. Phi-2: The surprising power of small language models. Microsoft Research Blog.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. DSPy: Compiling declarative language model calls into state-of-the-art pipelines. In *International Conference on Learning Representations (ICLR)*.

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

Roberto Labadie-Tamayo, Berta Chulvi, and Paolo Rosso. 2023. Everybody Hurts, Sometimes: Overview of Hurtful Humour at IberLEF 2023: Detection of Humour Spreading Prejudice in Twitter. *Procesamiento del Lenguaje Natural*, 71:383–395.

Abdulmujeeb Lawal and Saurav K. Aryal. 2026. Howard University-AI4PC at SemEval-2026 task 1: Exploring prompt strategies for automatic humor generation. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, United States. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

J. A. Meaney, Steven Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. SemEval 2021 task 7: HaHackathon, detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 105–119, Online. Association for Computational Linguistics.

Julie-Anne Meaney. 2024. *Demographically-aware computational humor*. Ph.D. thesis, The University of Edinburgh.

Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. RareAct: A video dataset of unusual interactions. *Preprint*, arXiv:2008.01018.

George A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Tristan Miller, Christian Hempelmann, and Iryna Gurevych. 2017. SemEval-2017 task 7: Detection and interpretation of English puns. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 58–68, Vancouver, Canada. Association for Computational Linguistics.

Tan Loc Nguyen and Dang Tuan Nguyen. 2026. DANGNT@SGU at SemEval-2026 task 1: A two-stage Mistral generator with DistilBERT reranking for English humor generation. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, United States. Association for Computational Linguistics.

Stefan Oprea, Lacrimioara Toma Oprea, Maria-Teodora Paval-Istrate, Diana Trandabat, and Daniela Gifu. 2026. FunnyBorg at SemEval-2026 task 1: Humor generation. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, United States. Association for Computational Linguistics.

Hemeshkumar Parthiban and R. Priyadharsini. 2026. SemEval-2026 task 1: Humor generation – text-based humor generation (English). In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, United States. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. SemEval-2017 task 6: #HashtagWars: Learning a sense of humor. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 49–57, Vancouver, Canada. Association for Computational Linguistics.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Victor Raskin. 1985. *Semantic Mechanisms of Humor*, volume 24 of *Studies in Linguistics and Philosophy*. D. Reidel, Dordrecht.

Batya Schechter, Sarah Barzel, and Chaya Liebeskind. 2026. JCT at SemEval-2026 task 1: Let the best joke win - a generate-and-rank approach to constrained humor. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, United States. Association for Computational Linguistics.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300. Introduces Group Relative Policy Optimization (GRPO).

Wutao Shen, Liyuan Huang, Jiawei He, Lin Li, and Jin Zhang. 2026. ICT-NLP at SemEval-2026 task 1: Humor generation via RAG-based augmentation and multi-LLM internal-external voting. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, United States. Association for Computational Linguistics.

J. Sjöbergh and K. Araki. 2007. Recognizing humor without recognizing meaning. In *Applications of Fuzzy Sets Theory: 7th International Workshop on Fuzzy Logic and Applications, WILF*, pages 469–476.

Charles Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Marek Šuppa, Viktória Ondrejová, Lucia Ganajová, Gregor Karetka, and Daniel Skala. 2026. RAGthoven at SemEval-2026 task 1: A multi-stage pipeline walks into a benchmark and barely clears the bar. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, United States. Association for Computational Linguistics.

Alexey Tikhonov and Alexey Ivanov. 2026. lmfaoooo at SemEval-2026 task 1: Humor is an audience. preference modeling for constrained humor generation. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, United States. Association for Computational Linguistics.

Joe Toplyn. 2014. *Comedy Writing for Late-Night TV: How to Write Monologue Jokes, Desk Pieces, Sketches, Parodies, Audience Pieces, Remotes, and Other Short-Form Comedy*. Twenty Lane Media.

Joe Toplyn. 2023. Witscript 3: A hybrid AI system for improvising jokes in a conversation. *arXiv preprint arXiv:2301.02695*.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, and 1 others. 2020. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272.

Kongqiang Wang, Peng Zhang, and Qingli Tan. 2026. wangkongqiang at SemEval-2026 task 1: MWAHAHA - competition on humor generation. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, United States. Association for Computational Linguistics.

Warda Yousaf. 2026. SLPG_FJWU_Warda at SemEval-2026 task 1: A multimodal vision-language approach for humor generation using fine-tuned BLIP. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, United States. Association for Computational Linguistics.

Mohammad Erfan Zare, Tahere Abbasi, Hadi Veisi, Sayin Ala, and Hanieh Naderi. 2026. YNWA_AZ at SemEval-2026 task 1: Bridging the semantic-visual gap: Multimodal humor generation. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, United States. Association for Computational Linguistics.

Xulong Zhang, Jin Wang, and Xuejie Zhang. 2026. YNU-HPCC at SemEval-2026 task 1: Constraint-aware in-context learning for multilingual humor generation. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, United States. Association for Computational Linguistics.

A Pilot Studies

We conducted two internal pilots before launching the official competition, both annotated through the same arena-style interface used in the main task. The pilots served different purposes: pilot 1 validated the basic protocol and informed the constraint design, while pilot 2 verified that the headline-based format chosen for the test set would still discriminate among frontier systems.

Pilot 1 (June 2025). The first pilot covered Subtask A in English and Spanish (100 items per language) and Subtask B in English (20 image-and-constraint items). Each Subtask A item was built from two words randomly sampled from the most frequent nouns, verbs, adjectives, and adverbs in EspressoEnglish.[9] We evaluated seven baselines: Claude 3.5 Sonnet (Anthropic), Gemini 1.5/2.0 Flash at multiple decoding temperatures, and Llama 3.1 8B Instruct. Annotation was carried out by the organizers, yielding roughly 850 pairwise battles for English Subtask A, 211 for Spanish, and 40 for Subtask B. Table 1 reports the resulting Bradley–Terry Elo ratings.

Subtask	System	Elo
A-en	Claude 3.5 Sonnet	1134
A-en	Gemini 2.0 Flash (t=0.9)	1026
A-en	Gemini 2.0 Flash (t=1.2)	1011
A-en	Gemini 1.5 Flash (t=0.7)	1000
A-en	Gemini 1.5 Flash (t=1.2)	983
A-en	Gemini 1.5 Flash (t=0.9)	957
A-en	Llama 3.1 8B Instruct	888
A-es	Claude 3.5 Sonnet	1085
A-es	Gemini 2.0 Flash (t=0.9)	1070
A-es	Gemini 1.5 Flash (t=1.2)	1031
A-es	Gemini 2.0 Flash (t=1.2)	1017
A-es	Gemini 1.5 Flash (t=0.9)	967
A-es	Gemini 2.0 Flash (t=1.5)	945
A-es	Llama 3.1 8B Instruct	884
B-en	Gemini 2.0 Flash (t=0.7)	1101
B-en	Gemini 1.5 Flash (t=0.7)	899

Table 1: Pilot 1 Elo ratings per subtask.

Subtask	System	Elo
A-en	Gemini 2.5 Pro	1041
A-en	Qwen3-235B-A22B	1003
A-en	gpt-oss-120B	956
A-es	Gemini 2.5 Pro	1181
A-es	Qwen3-235B-A22B	982
A-es	gpt-oss-120B	837

Table 2: Pilot 2 Elo ratings per language.

Pilot 2 (August 2025). The second pilot focused on Subtask A under the news-headline constraint format adopted for the official test set, in English and Spanish. We pitted three frontier systems—Gemini 2.5 Pro, Qwen3-235B-A22B, and gpt-oss-120B—over 80 battles per language. Table 2 reports the resulting Elo ratings; the ordering is consistent across languages, with Gemini 2.5 Pro on top in both, by a wider margin in Spanish than in English.

What the pilots informed. Both pilots showed that pairwise annotator agreement sufficed to produce a reliable system ranking and that frontier models clearly outperformed smaller open mod-

[9] https://www.espressoenglish.net/

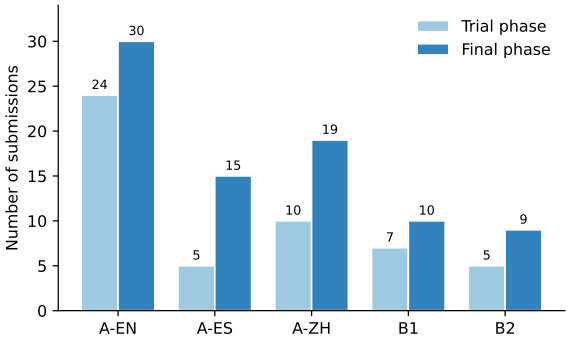


Figure 1: Number of submissions per subtask in the trial and final phases. Subtask A-EN attracted the most submissions, while Subtask B (multimodal, English only) drew the fewest.

els. Pilot 1 also informed the final word-pair design for Subtask A: pairs of very common words (general enough to slot into almost any joke) made the constraint trivial, while fully random pairings were often too disjoint to support a coherent joke, so we adopted rarer but still usable combinations grounded in compatible part-of-speech pairs (e.g., noun–verb), which we ultimately drew from the RareAct collocations list (Miech et al., 2020). Pilot 2 confirmed that, even with the strongest publicly available LLMs of August 2025 and the headline constraint format used for the test set, the gap between systems remained measurable—supporting our decision to keep the headline format for the official evaluation despite the risk of frontier-model saturation.

B Data Sources and Annotation

Table 3 lists the news outlets we sourced headlines from, broken down by language and country. Figure 1 shows how participation grew between the trial and final phases, broken down by subtask. Figure 2 is a screenshot of the annotation interface for an example pair from Task A-es. Table 4 reports the per-subtask vote counts in the development and test phases, alongside the skip and tie rates for the test phase. Figure 3 visualizes the daily annotation effort during the test phase, split between volunteer and Prolific annotators. Table 5 and Fig. 4 present aggregated demographic information for the Prolific annotators we hired for the test phase.

We hired Prolific annotators for the test phase under a single recruitment constraint: their first language had to match the subtask language (English for A-en, B1, and B2; Spanish for A-es; and Chinese for A-zh). The demographics differ noticeably across subtasks (e.g., 82% male in A-es and 65% female in A-zh; the youngest pool, in A-en and A-es, contrasts with B1, where 40% of annotators were 40–49); these are artifacts of who was available on Prolific at recruitment time, and we did not actively balance for them.

C Baseline Prompts

This section lists the exact prompts used for the baselines.

C.1 Subtask A

C.1.1 English

Two-word constraint prompt

```
Create a joke based on the following
elements:

Words to include: {word1, word2}

Please craft a joke that incorporates
all the specified elements. The joke
should be concise, creative, and
genuinely funny. All required words
must appear somewhere in the joke.
```

News title prompt

```
Create a joke based on the title of a
news article:

"{headline}"

The joke should be concise, creative,
and genuinely funny. Only return the
joke and nothing else.
```

C.1.2 Spanish

Two-word constraint prompt

```
Create a joke based on the following
elements:

Words to include: {word1, word2}

Please craft a joke that incorporates
all the specified elements. The joke
should be concise, creative, and
genuinely funny. All required words
must appear somewhere in the joke.
All jokes must be in Spanish.
```

News title prompt

```
Create a joke based on the title of a
news article:

"{headline}"

The joke should be concise, creative,
and genuinely funny. Only return
the joke and nothing else. All jokes
must be in Spanish.
```

Language	Country	News Outlet	URL
English	Australia	ABC Australia	abc.net.au
English	Canada	CBC	cbc.ca
English	Ireland	RTE	rte.ie
English	South Africa	IOL	iol.co.za
English	UK	BBC	bbc.com
English	UK	Sky News	news.sky.com
English	UK	The Guardian	theguardian.com
English	UK	The Independent	independent.co.uk
English	UK	The Telegraph	telegraph.co.uk
English	US	ABC	abcnews.go.com
English	US	Associated Press	apnews.com
English	US	Bloomberg	bloomberg.com
English	US	CBS	cbsnews.com
English	US	CNN	edition.cnn.com
English	US	Financial Times	ft.com
English	US	NBC	nbcnews.com
English	US	Reuters	reuters.com
English	US	The Economist	economist.com
English	US	The New York Times	nytimes.com/international
English	US	Voice of America	voanews.com
English	US	Vox	vox.com
Spanish	Argentina	Clarín	clarin.com
Spanish	Argentina	infobae	infobae.com
Spanish	Argentina	La Nación	lanacion.com.ar
Spanish	Chile	La Tercera	latercera.com
Spanish	Colombia	El Tiempo	eltiempo.com
Spanish	Ecuador	El Comercio	elcomercio.com
Spanish	El Salvador	La Prensa Gráfica	laprensagrafica.com
Spanish	España	ABC	abc.es
Spanish	España	El Mundo	elmundo.es
Spanish	España	El País Madrid	elpais.com
Spanish	España	La Vanguardia	lavanguardia.com
Spanish	Guatemala	Prensa Libre	prensalibre.com
Spanish	Inglaterra	BBC	bbc.com/mundo
Spanish	México	El Universal México	eluniversal.com.mx
Spanish	Perú	El Comercio Perú	elcomercio.pe
Spanish	República Dominicana	Diario Libre	diariolibre.com
Spanish	Uruguay	El País	elpais.com.uy
Spanish	Uruguay	Montevideo COM	montevideo.com.uy
Spanish	Uruguay	Tabaré	radiotabare.com.uy
Spanish	Uruguay	Telenoche	telenoche.com.uy
Spanish	US	CNN en Español	cnnespanol.cnn.com
Spanish	Venezuela	El Universal Venezuela	eluniversal.com
Chinese	China	中国日报 (China Daily)	chinadaily.com.cn
Chinese	China	搜狐新闻 (So Hu)	news.sohu.com
Chinese	China	新京报 (The Beijing News)	bjnews.com
Chinese	China	澎湃新闻 (The Paper)	thepaper.cn
Chinese	China	环球时报 (Global Times)	globaltimes.cn
Chinese	China	环球网 (Huan Qiu)	huanqiu.com
Chinese	China	羊城晚报 (Yangcheng Evening News)	ycwb.com
Chinese	China	财新网 (Cai Xin)	caixin.com
Chinese	Hong Kong	星島頭條 (Sing Tao)	stheadline.com
Chinese	Hong Kong	鳳凰網 (Phoenix)	ifeng.com
Chinese	Hong Kong	明報 (Ming Pao)	mingpao.com
Chinese	Hong Kong	香港 01 (HK01)	hk01.com
Chinese	Macao	澳門日報 (Macao Daily News)	modaily.cn
Chinese	Macao	華僑報 (Jornal Va Kio)	vakiodaily.com
Chinese	Malaysia	中國報 (China Press)	chinapress.com
Chinese	Malaysia	星洲网 (Sin Chew Daily)	sinchew.com
Chinese	Malaysia	聯合日報 (United Daily News)	unitedaily.my/zn
Chinese	Singapore	新明日报 (Shin Min Daily News)	shinmin.sg
Chinese	Singapore	联合早报 (Lianhe Zaobao)	zaobao.com.sg
Chinese	Taiwan	TVBS 新聞網 (TVBS News Channel)	news.tvbs.com.tw
Chinese	Taiwan	中央通訊社 (Central News Agency)	cna.com.tw
Chinese	Taiwan	自由時報 (Liberty Times Net)	ltn.com.tw
Chinese	Taiwan	聯合新聞網 (United Daily News, UDN)	udn.com

Table 3: News outlets used to source headlines by language and country.

Figure 2: Screenshot from the annotation page with an example of Task A-es (Spanish), showing the prompt, two candidate jokes, and the annotator's options. The header reads "Which Computer Program Output is Funnier?" and the task selector is set to "Task A: Joke in Spanish". The headline (*Titular*) translates to "The secrets of evolution: how physical and biological barriers define the ceiling of species". A's output translates to "The secret of evolution is that the 'ceiling of species' is impassable… unless you are my neighbor's plumber", and B's output to "Did you know sheep have a secret for sleeping? It's that they cannot jump over the river of evolution! And that's not a myth, it's a biological fact. That is the ceiling of species!". The buttons at the bottom let annotators flag either output as offensive, vote for the left or right output, declare a tie, or skip the pair.

	Dev	Test				
Subtask	**Total**	**Total**	**Prolif.**	**Volunt.**	**Skip%**	**Tie%**
A-en	2 960	6 238	4 942	1 296	33.2	12.1
A-es	907	2 307	1 743	564	12.7	14.4
A-zh	212	2 103	2 033	70	6.8	14.8
B1	266	1 222	978	244	16.2	18.7
B2	164	1 066	1 011	55	9.4	18.0
Total	4 509	12 936	10 707	2 229	23.4	13.8

Table 4: Number of non-skip votes received per subtask in the development and test phases. The Test columns also break down votes by paid Prolific (Prolif.) vs. volunteer (Volunt.) annotators and report the proportion of skip and tie annotations as percentages of all votes (skips + non-skip votes); the development phase was annotated entirely by organizers and volunteers.

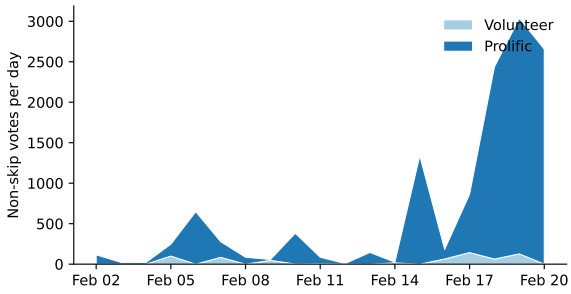


Figure 3: Daily non-skip votes during the test annotation period, split between volunteer (light) and Prolific (dark) annotators.

C.1.3 Chinese

Two-word constraint prompt

```
Create a joke based on the following
elements:

Words to include: {word1, word2}

Please craft a joke that incorporates
all the specified elements. The joke
should be concise, creative, and
genuinely funny. All required words
must appear somewhere in the joke.
All jokes must be in Chinese.
```

News title prompt

```
Create a joke based on the title of a
news article:

"{headline}"

The joke should be concise, creative,
and genuinely funny. Only return
the joke and nothing else. All jokes
must be in Chinese.
```

C.2 Subtask B1

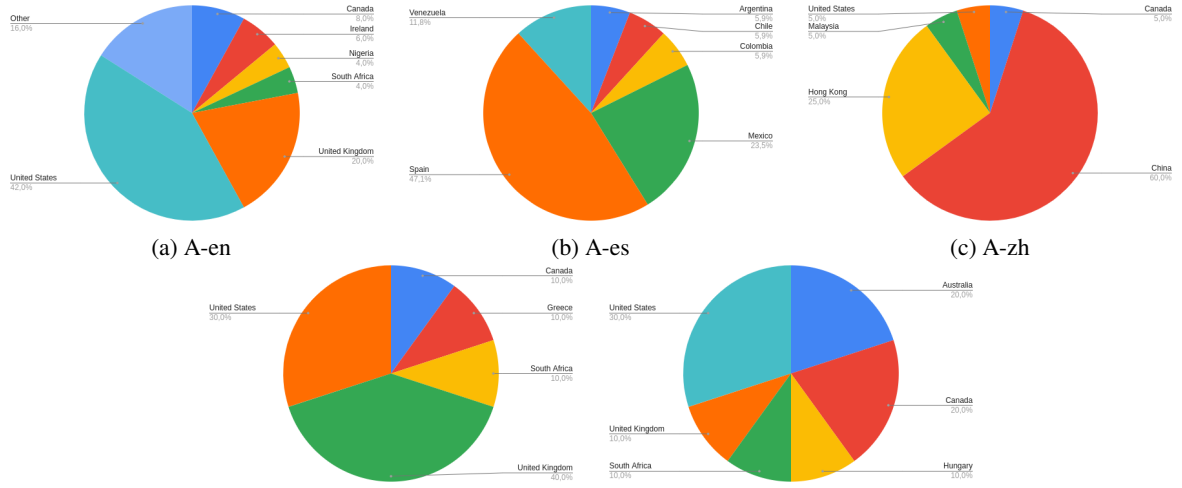
GIF-based joke generation

```
Create a joke based on the given GIF.
The joke should be concise, creative,
and genuinely funny. Only return the
joke and nothing else.
```

C.3 Subtask B2

GIF-conditioned joke completion

```
Complete a joke based on the GIF and
the following starting prompt:
```



(a) A-en (b) A-es (c) A-zh

(d) B1 (e) B2

Figure 4: Country of birth of the annotators hired through Prolific

		A-en	A-es	A-zh	B1	B2	Total
Age range	< 30	38	47	40	30	10	36
	30 to 39	26	29	25	0	50	26
	40 to 49	20	18	15	40	10	20
	50 to 59	12	6	10	30	10	12
	>= 60	4	0	10	0	20	6
Gender	Female	52	18	65	50	60	50
	Male	48	82	35	50	40	50

Table 5: Percentages of age range and gender of the annotators hired through Prolific. Each column sums 100%.

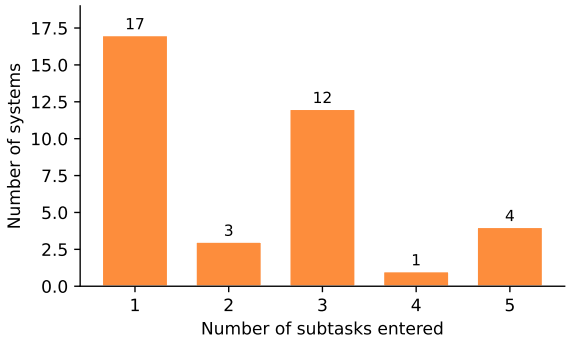


Figure 5: How many subtasks each team entered. Most teams focused on a single subtask, while four teams (BAHAHA, FunnyBorg, j10official, YNWA_AZ) covered all five.

```
"{prompt}"

The joke should be concise, creative,
and genuinely funny. Only return the
joke and nothing else.
```

D Participating Systems

Table 6 is our canonical roster: it lists every participating team together with their CodaBench username, citation (or pointer to Section E for the two systems whose authors shared a written description but did not submit a paper), institutional affiliation, country, and the subtasks they participated in. Table 7 provides a compact overview of all systems with their subtask average Elo rating and approach keywords. Figure 5 reports how many subtasks each team chose to enter. Figure 6 reports the base-model families participants reached for, with a system counted once per family it used; Qwen and Gemini lead, followed by Gemma, Llama, and GPT-class models. Figure 7 summarizes the higher-level techniques participants combined to build their systems, again with one count per (system, tag) pair. Figure 8 visualizes the geographic distribution of the 27 teams with system-description papers that report an affiliation (lm-faoooo, with no reported affiliation, is omitted), spanning 17 countries.

E Author-Submitted System Notes

Two participants who did not submit a system-description paper shared a brief written summary of their system; we reproduce these notes verbatim below.

Edward Ajayi (`jayicodes`).

> Our system for SemEval-2026 Task 1 – MWAHAHA explores the generative potential and alignment challenges of large language models in automated humor generation. We begin with an ensemble-style generation phase using

System	Team name	Username	Submission	Affiliation	Country	Subtasks
ABARUAH	ABARUAH	abaruah	(Baruah, 2026)	Assam Don Bosco University	India	A-EN, B1, B2
aba_team	—	aba_team	(Dikhtiar et al., 2026)	Technical University of Munich	Germany	A-EN, B1, B2
AI4PC	AI4PC	abdulmj1	(Lawal and Aryal, 2026)	Howard University	USA	A-EN
ar01989	—	ar01989	—	—	—	A-EN, A-ES
BAHAHA	BAHAHA	ahoblitz	(Arora and Hoblitzell, 2026)	Purdue University	USA	A-EN, A-ES, A-ZH, B1, B2
begumyivli	—	begumyivli	—	—	—	A-EN
CUET_Clashing	CUET_Clashing	clashing	(Chowdhury et al., 2026)	Chittagong Univ. of Engineering and Technology	Bangladesh	A-EN, A-ES, A-ZH
DANGNT@SGU	DANGNT@SGU	tanlocn	(Nguyen and Nguyen, 2026)	Ton Duc Thang Univ. & Saigon University	Vietnam	A-EN
deepgpt	—	deepgpt	(Chen and Weng, 2026)	Yunnan University	China	A-ZH
DUTH	DUTH	arampageos	(Arampatzis and Arampatzis, 2026)	Democritus University of Thrace	Greece	A-EN, A-ES, A-ZH
Edward Ajayi	Edward Ajayi	jayicodes	Section E	—	—	A-EN
FunnyBorg	FunnyBorg	stefanoprea	(Oprea et al., 2026)	Alexandru Ioan Cuza University of Iaşi	Romania	A-EN, A-ES, A-ZH, B1, B2
hemeshkumar_31	—	hemeshkumar_31	(Parthiban and Priyadharsini, 2026)	Rajalakshmi Engineering College	India	A-EN
hugang11	—	hugang11	(Hu et al., 2026)	Yunnan University	China	A-ZH
ICT-NLP	ICT-NLP	shenwutao	(Shen et al., 2026)	Chinese Academy of Sciences	China	A-ZH
INF-rsrs	INF-rsrs	jjuliar	(Bazzo et al., 2026)	Federal University of Rio Grande do Sul	Brazil	A-EN
j10official	—	j10official	(Agrawal and Mamidi, 2026)	IIIT Hyderabad	India	A-EN, A-ES, A-ZH, B1, B2
JCT	JCT	jct_sb	(Schechter et al., 2026)	Jerusalem College of Technology	Israel	A-EN
L52+-IIMAS-UNAM	L52+-IIMAS-UNAM	soyliz30	(Camacho-González et al., 2026)	Universidad Nacional Autónoma de México	Mexico	A-EN, A-ES
Lattice	Lattice	oseminck	(Dehouck et al., 2026)	CNRS / Lattice & CY Cergy Paris Université	France	A-EN
Lattice_Dev	Lattice_Dev	mdehouck	—	—	—	A-EN
lmfaoooo	—	lmfaoooo	(Tikhonov and Ivanov, 2026)	—	—	A-EN, A-ES, A-ZH
lu_rui	—	lu_rui	—	—	—	A-EN, A-ES, A-ZH
MINDS	MINDS	sinaeskandari	(Eskandari et al., 2026)	Politecnico di Torino	Italy	A-EN
polarizedteam	—	polarizedteam	—	—	—	A-EN
praveenjoshi007	—	praveenjoshi007	—	—	—	B1, B2
RAGthoven	RAGthoven	mrshu	(Šuppa et al., 2026)	Comenius University Bratislava	Slovakia	A-EN, A-ES, A-ZH
SLPG_FJWU_Insa	—	SLPG_FJWU_Insa	(Abbas and Abdul Rauf, 2026)	Fatima Jinnah Women University	Pakistan	A-EN
SLPG_FJWU_Warda	SLPG_FJWU_Warda	warda_yousaf	(Yousaf, 2026)	Fatima Jinnah Women University	Pakistan	B1
TüLK	TüLK	luttt	(Brüggemann and Hou, 2026)	University of Tübingen	Germany	A-EN, A-ES, A-ZH
UIR_CIS	UIR_CIS	xxl_6699	Section E	—	—	A-ZH, B1, B2
wangkongqiang	—	wangkongqiang	(Wang et al., 2026)	Yunnan University	China	A-EN, A-ZH, B1, B2
XplaiNLP	XplaiNLP	berkbubus	(Bubus et al., 2026)	Technische Universität Berlin	Germany	A-EN, A-ES, A-ZH
xxl2233	—	xxl2233	—	—	—	A-ZH
yasamin_al	—	yasaminal	(Aali, 2026)	Brock University	Canada	A-EN, A-ES, A-ZH
YNU-HPCC	YNU-HPCC	zhangxulong	(Zhang et al., 2026)	Yunnan University	China	A-EN, A-ES, A-ZH
YNWA_AZ	YNWA_AZ	t_abbasi7	(Zare et al., 2026)	University of Tehran	Iran	A-EN, A-ES, A-ZH, B1, B2

Table 6: Roster of all participating systems, sorted alphabetically by system name. The **System** column shows the team name where one was reported and falls back to the CodaBench username otherwise; we use this column as the canonical reference throughout the paper. The **Team name** column shows the team name as filed in CodaBench ("—" if none was filed). The **Submission** column gives the citation for systems that submitted a system-description paper, a reference to Section E for systems that shared a short written description, and "—" for systems that submitted nothing. The **Affiliation** and **Country** columns are populated for systems with a paper; "—" otherwise. The **Subtasks** column records which subtasks the team participated in.

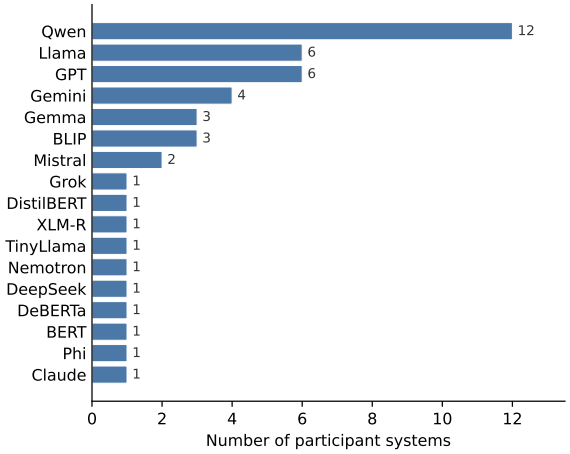


Figure 6: Base-model families used by participant systems. A system that combined multiple families is counted once per family.

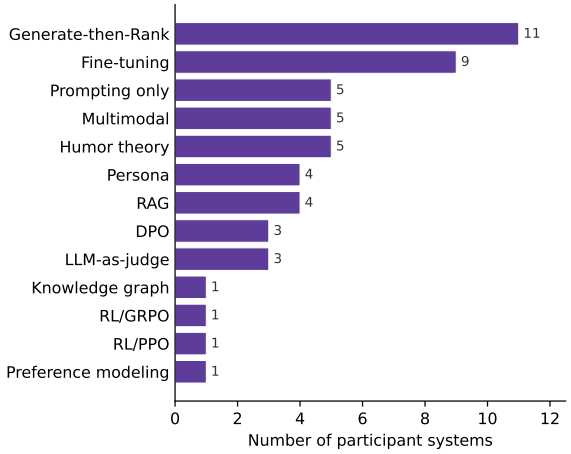


Figure 7: Higher-level techniques used by participant systems, with one count per (system, tag) pair. A system applying generate-then-rank with RAG and fine-tuning is counted in three rows.

Qwen2.5-32B-Instruct to sample a diverse range of comedic responses to given headline/word combinations, aiming to maximize creative variability. To calibrate these outputs, we design a simplified LLM-based judging mechanism that scores generated jokes according to humor quality and adherence to prompt constraints. These scores are then used to synthesize alignment data for Direct Preference Optimization (DPO), enabling the transfer of humor generation capabilities from the teacher model to a smaller student model, Qwen2.5-7B. While our results demonstrate that the

System	A-EN	A-ES	A-ZH	B1	B2	Avg Elo	Approach Keywords
praveenjoshi007				✓	✓	1099	–
RAGthoven	✓	✓	✓			1091	4-stage pipeline, RAG, SSTH+BVT
baseline	✓	✓	✓	✓	✓	1084	Gemini 2.5 Flash, zero-shot
SLPG_FJWU_Insa	✓					1080	QLoRA, Phi-2
SLPG_FJWU_Warda				✓		1077	SFT, BLIP, MemeCap
XplaiNLP	✓	✓	✓			1073	BVT, Gatekeeper, EmoBERTa
lmfaoooo	✓	✓	✓			1071	Humor Arena, pref. model
JCT	✓					1063	Gen-Rank, 10 humor styles
INF-rsrs	✓					1060	Gen-Rank, BT ranking, ceiling test
xxl2233			✓			1057	–
ICT-NLP			✓			1052	RAG, keyw. assoc., multi-LLM voting
DUTH	✓	✓	✓			1042	Zero-shot, Qwen2.5-14B, Mistral-7B
begumyivli	✓					1041	–
YNU-HPCC	✓	✓	✓			1039	PPO, golden library, mDeBERTa
Lattice	✓					1034	DeepSeek-R1, Funny Filters, ICL
UIR_CIS			✓	✓	✓	1033	–
MINDS	✓					1022	DPO, RAG, Gemma 2, round-robin
AI4PC	✓					1020	Llama-3.1-8B, Chappelle, tweet
j10official	✓	✓	✓	✓	✓	1005	GTVH, Gemma 3 27B, Nemotron
FunnyBorg	✓	✓	✓	✓	✓	1002	Gemma, voting
polarizedteam	✓					999	–
YNWA_AZ	✓	✓	✓	✓	✓	997	RAG, Llama-3-8B, BLIP+TinyLlama
lu_rui	✓	✓	✓			993	–
wangkongqiang	✓		✓	✓	✓	992	BLIP, Qwen3-Next-80B, Qwen-Plus
hugang11			✓			991	CoT-SFT, DPO, `<THINK>/<ROAST>`
Lattice_Dev	✓					991	–
TüLK	✓	✓	✓			980	GRPO, XLM-RoBERTa, Qwen2.5-7B
ABARUAH	✓			✓	✓	965	Qwen2-VL, Qwen3-8B
DANGNT@SGU	✓					962	QLoRA, DistilBERT reranker
aba_team	✓			✓	✓	954	LoRA, Plan2joke, Wiki microcards
Edward Ajayi	✓					950	DPO, Qwen2.5-32B→7B distillation
BAHAHA	✓	✓	✓	✓	✓	950	15 styles, HITL, distilled ranker
ar01989	✓	✓				946	–
deepgpt			✓			903	QLoRA, instruction masking
L52+-IIMAS-UNAM	✓	✓				896	Grok, stylistic-humor DNA
yasamin_al	✓	✓	✓			878	WordNet, semantic anchors, LLaMA
CUET_clashing	✓	✓	✓			871	Zero-shot, Qwen2.5-3B
hemeshkumar_31	✓					843	Zero-shot, deadpan

Table 7: Overview of all participating systems sorted by subtask average Elo rating (descending). The **System** column shows the team name where one was provided, and falls back to the submitter's username otherwise. Subtask columns indicate participation (✓). The average is computed only over subtasks in which the system participated. Approach keywords are based on provided descriptions; systems without provided descriptions have "–" listed. HITL = human-in-the-loop.

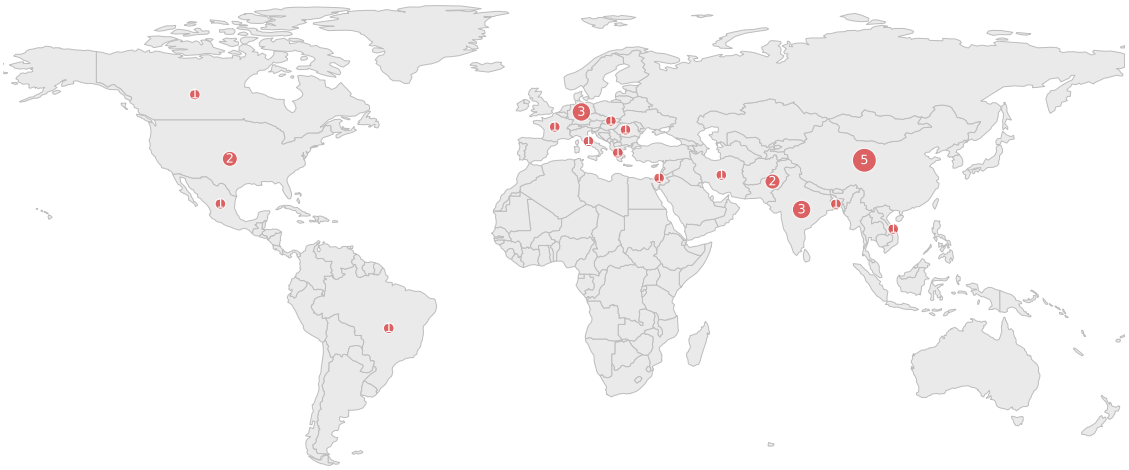


Figure 8: Geographic distribution of system-description papers, with one marker per country sized in proportion to the number of teams. The number inside each marker is the team count for that country.

teacher model can generate high-quality jokes, we find that achieving consistent comedic alignment through synthetic preference signals remains challenging, largely due to limitations in the judging model's scoring reliability when constructing alignment data. Overall, our findings suggest that although the generative foundation of large LLMs for humor is strong, more robust reward modeling strategies are necessary to ensure stable alignment, and this submission provides an empirical foundation for future work toward more capable humor generation systems.

UIR_CIS (`xxl_6699`).

SemEval-2026 Task 1 (MWAHAHA) challenges models to generate controlled humor under explicit lexical and topical constraints. For the Chinese track, we compared structured fine-tuning of small language models against technique-guided prompting with large language models (LLMs). While constructing a high-quality dataset with intermediate reasoning supervision improved structural awareness, fine-tuned small models still struggled with coherence. Consequently, we adopted a technique-aware prompting strategy leveraging advanced LLMs, combined with multi-strategy generation and human selection. This approach significantly improved instruction compliance and humor quality, securing the first place in the Chinese track. Our results indicate that structured prompting of strong LLMs currently offers superior reliability over small-model fine-tuning for constrained humor generation.

F Detailed Results

Tables 8 to 12 report the official Elo leaderboards for the five subtasks (Subtask A in English, Spanish, and Chinese, plus Subtasks B1 and B2). Figure 10 shows the distribution of final Elo ratings across systems, faceted by subtask, and Fig. 11 reports the pairwise battle counts among the 12 most-battled Subtask A-EN systems. Table 13 shows representative high-win-rate jokes from the test set, Table 14 groups the 28 papered systems plus the baseline by primary approach category, and Fig. 9 visualizes the distribution of system Elo ratings per approach category and subtask.

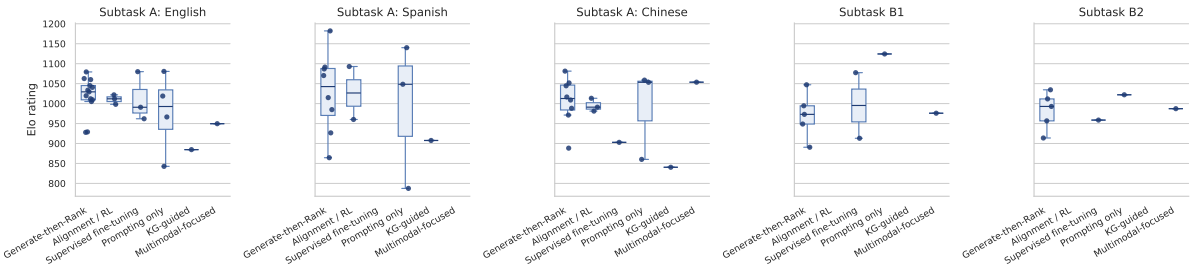


Figure 9: Elo ratings of papered systems by primary approach category, faceted by subtask. Our Gemini 2.5 Flash baseline is included as a system in the Prompting-only column for each subtask.

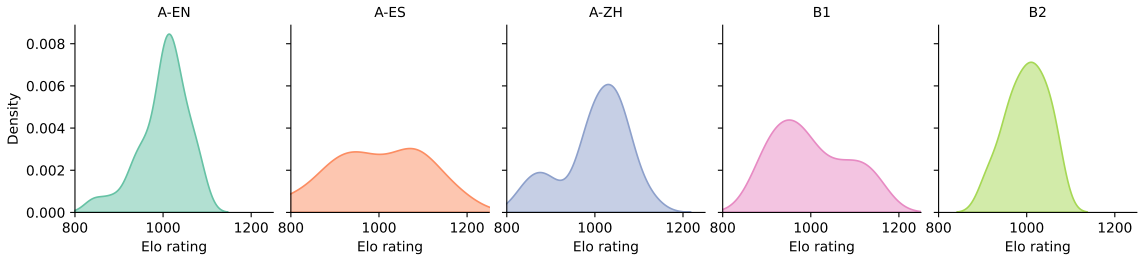


Figure 10: Distribution of final Elo ratings across systems, faceted by subtask.

Rank	System	Rating	95% CI	Votes
1	*baseline*	1081	[1045, 1110]	382
1	SLPG_FJWU_Insa	1080	[1046, 1120]	388
1	XplaiNLP	1079	[1057, 1115]	374
1	JCT	1063	[1036, 1099]	382
1	INF-rsrs	1060	[1027, 1091]	376
1	RAGthoven	1045	[1018, 1073]	382
1	lmfaoooo	1041	[1009, 1064]	389
1	begumyivli	1041	[1008, 1068]	382
1	Lattice	1034	[1005, 1072]	389
2	YNWA_AZ	1029	[1001, 1053]	408
2	MINDS	1022	[989, 1054]	383
2	AI4PC	1020	[992, 1053]	385
3	DUTH	1019	[984, 1045]	382
2	FunnyBorg	1012	[986, 1051]	388
4	YNU-HPCC	1012	[985, 1036]	378
4	ABARUAH	1009	[979, 1041]	378
4	lu_rui	1008	[982, 1038]	385
4	j10official	1005	[969, 1042]	370
5	ar01989	1003	[971, 1035]	381
6	polarizedteam	999	[968, 1022]	379
6	TüLK	998	[975, 1025]	484
6	Lattice_Dev	991	[963, 1026]	378
6	aba_team	991	[958, 1025]	380
11	CUET_clashing	966	[945, 993]	515
13	DANGNT@SGU	962	[926, 986]	378
18	Edward Ajayi	950	[920, 977]	388
16	wangkongqiang	950	[922, 982]	381
23	BAHAHA	929	[903, 960]	377
24	L52+-IIMAS-UNAM	928	[903, 950]	390
28	yasaminal	885	[855, 915]	382
30	hemeshkumar_31	843	[802, 875]	378

Table 8: SemEval-2026 Humor Generation Task A - English Results

Rank	System	Rating	95% CI	Votes
1	RAGthoven	1182	[1143, 1222]	291
1	*baseline*	1140	[1098, 1177]	287
2	YNU-HPCC	1093	[1060, 1129]	287
2	lmfaoooo	1091	[1053, 1121]	288
2	FunnyBorg	1087	[1062, 1128]	288
2	XplaiNLP	1070	[1024, 1109]	288
3	DUTH	1048	[1020, 1093]	289
6	j10official	1015	[985, 1049]	289
8	YNWA_AZ	985	[941, 1012]	288
8	TüLK	960	[927, 994]	288
9	lu_rui	953	[912, 984]	290
9	BAHAHA	927	[894, 968]	289
10	yasaminal	908	[868, 937]	288
10	ar01989	889	[854, 929]	288
12	L52+-IIMAS-UNAM	864	[827, 907]	288
16	CUET_clashing	787	[753, 823]	288

Table 9: SemEval-2026 Humor Generation Task A - Spanish Results

Rank	System	Rating	95% CI	Votes
1	UIR_CIS	1120	[1085, 1164]	211
1	lmfaoooo	1081	[1031, 1127]	212
1	DUTH	1059	[1018, 1091]	210
1	xxl2233	1057	[1015, 1100]	211
1	wangkongqiang	1054	[1024, 1104]	210
1	*baseline*	1053	[1003, 1090]	210
1	ICT-NLP	1052	[1009, 1094]	210
1	RAGthoven	1045	[1004, 1090]	210
2	lu_rui	1018	[980, 1052]	210
2	j10official	1016	[966, 1063]	211
2	YNU-HPCC	1013	[971, 1061]	214
2	FunnyBorg	1009	[967, 1049]	210
2	hugang11	991	[958, 1036]	209
2	BAHAHA	988	[945, 1033]	210
5	TüLK	981	[928, 1017]	209
6	XplaiNLP	971	[939, 1014]	211
14	deepgpt	903	[868, 946]	210
16	YNWA_AZ	888	[845, 933]	209
17	CUET_clashing	860	[808, 899]	210
17	yasaminal	840	[791, 879]	211

Table 10: SemEval-2026 Humor Generation Task A - Chinese Results

Rank	System	Rating	95% CI	Votes
1	praveenjoshi007	1140	[1099, 1180]	222
1	*baseline*	1124	[1084, 1164]	222
1	SLPG_FJWU_Warda	1077	[1043, 1110]	222
3	YNWA_AZ	1047	[1012, 1079]	222
4	j10official	994	[966, 1030]	223
5	wangkongqiang	976	[941, 1007]	221
4	ABARUAH	973	[938, 1018]	223
5	BAHAHA	949	[921, 983]	221
7	UIR_CIS	915	[879, 939]	222
6	aba_team	913	[877, 950]	222
8	FunnyBorg	891	[856, 924]	224

Table 11: SemEval-2026 Humor Generation Task B1 - Caption the GIF Results

Rank	System	Rating	95% CI	Votes
1	UIR_CIS	1065	[1032, 1102]	212
1	praveenjoshi007	1057	[1020, 1104]	214
1	YNWA_AZ	1035	[1006, 1069]	214
1	*baseline*	1022	[991, 1060]	214
1	FunnyBorg	1012	[982, 1048]	215
2	j10official	993	[960, 1026]	214
3	wangkongqiang	987	[948, 1016]	212
4	aba_team	959	[919, 1004]	217
5	BAHAHA	957	[911, 989]	212
7	ABARUAH	914	[870, 950]	214

Table 12: SemEval-2026 Humor Generation Task B2 - Fill in the GIF Caption Results

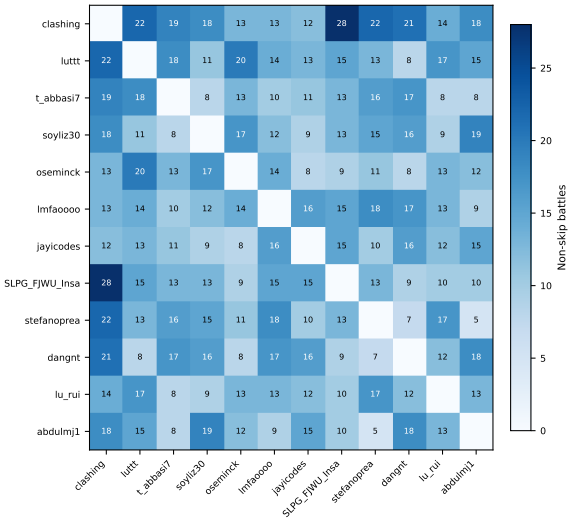


Figure 11: Pairwise non-skip battle counts among the 12 most-battled Subtask A-EN systems. The matrix is symmetric; the diagonal is set to zero (a system never battles itself).

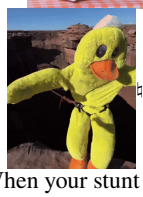
Subtask	System	Wins	Constraint	Joke
Top-rated joke per subtask:				
A-en	AI4PC	5/5	"Millennials are set to inherit tons of their parents' stuff…"	Because nothing says 'inheritance' like a garage full of your parents' old lawnmowers and a VHS player.
A-es	lmfaoooo	4/4	words *medir* / *cabello*	En el laboratorio quisieron **medir** mi **cabello** para análisis forense; salió tan "rebelde" que hasta el ADN pidió un abogado.[§]
A-zh	DUTH	4/4	"销量六连降，广汽埃安如何破危局"[†]	广汽埃安销量连降，看来得换个销售策略了，比如改卖彩票，毕竟买他们的车中奖率最高。[‡]
B1	SLPG_FJWU_Warda	4/4		Me reacting to unexpected drama like it's the season finale.
B2	UIR_CIS	3/3	 "When your stunt fall was supposed to be epic but ended with ___"	…you doing an unintentional impression of a dying spaghetti noodle.
Same-prompt contrast (the protocol distinguishes funniness within a single prompt; see Section 8 Finding 1):				
A-en	RAGthoven	3/3	"Panamanian lawmakers' Taiwan trip sparks diplomatic row with China"	Panamanian lawmakers visited Taiwan and sparked a diplomatic row with China. Apparently "One China policy" means you can only have one friend at a time.
A-en	yasamin_al	1/4		A group of Panamanians went to visit their sister islands in the Pacific, but when they arrived, they found themselves at odds with China over who owns those little rocks.
Illustrating other findings:				
A-en	SLPG_FJWU_Insa	4/4	"Dirt-powered sensors help farmers optimize their fields"	Dirt-powered sensors are great, but they complain about the lack of quality soil on vacations.[*]
A-en	INF-rsrs	4/4	"How to connect with old friends and why it matters"	Connecting with old friends matters because someone has to remember the real you, before your LinkedIn profile took over.[**]

Table 13: Sample jokes from the test set. We restricted to (system, prompt) pairs that entered at least three non-skip battles (Section 6.3 discusses why per-joke counts are sparse). The *top* block shows one high-WR joke per subtask, preferring concise outputs and a mix of systems. The *contrast* block shows two systems on the same prompt with very different WRs. The *other findings* block shows a small-model winner ($*$ SLPG_FJWU_Insa runs a Phi-2 (2.7B) QLoRA pipeline that tied for first in A-en—small models can compete; Section 8 Finding 5) and a system-deliberately-worst pick ($**$ INF-rsrs submitted what their own re-ranker scored *lowest*, yet still tied for first overall—evidence of a quality ceiling on the protocol; Section 8 Finding 3). "System" uses the team name where available, else the username; "Wins" shows non-skip battles won out of total entered. Constraints are abbreviated for space (full prompts are released with the data). [§]"At the lab they wanted to **measure** my **hair** for forensic analysis; it came out so 'rebellious' that even the DNA asked for a lawyer." [†]"Six straight months of declining sales — how can GAC Aian escape the crisis?" [‡]"GAC Aian's sales keep falling; maybe they should switch to selling lottery tickets — buying their cars seems to give the best winning odds anyway." English translations of the Spanish and Chinese strings were added by us for the reader's convenience; the systems produced only the original-language text. [♭]First frame of the input GIF (Giphy, from the talk show *Sherri*). [♮]First frame of the input GIF (Giphy, by Nate Richardson).

Category	N	Median	Top performer (best Elo)
Generate-then-Rank	14	1012	RAGthoven (ES, 1182)
Alignment / RL	4	1005	YNU-HPCC (ES, 1093)
Supervised fine-tuning	5	962	SLPG_FJWU_Insa (EN, 1080)
Prompting only	4	1035	baseline (ES, 1140)
Knowledge-graph guided	1	885	yasamin_al (ES, 908)
Multimodal-focused	1	982	wangkongqiang (ZH, 1054)

Table 14: Primary categorization of the 28 papered systems plus our Gemini 2.5 Flash baseline; rows sum to 29 because each entry is counted once, and the baseline is folded into the Prompting-only category since it is a single zero-shot prompt with no fine-tuning or candidate ranking. "Median" is the median Elo across all (system, subtask) instances in the category; "Top performer" is the (system, subtask) pair with the highest Elo, with the subtask in parentheses. Cross-cutting techniques (Persona, RAG, Theory-grounded, Multimodal) are discussed qualitatively in Section 7.