SELECT OR CREATE? BEHAVIORAL PREFERENCES AND QUALITY GAINS IN LLM HUMOR TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Do large language models (LLMs) prefer to select existing content or create new alternatives, and are such behavioral preferences associated with quality gains? We investigate selection-creation preferences in a news-based humor task, in which models evaluate candidate jokes and decide whether to adopt one unchanged or generate a new alternative. We introduce JEST, a multilingual dataset comprising 3,923 news headlines in English, Spanish, and Chinese, each paired with four jokes generated independently by distinct LLMs. We further propose an evaluation framework that conditions quality comparisons on models' initial decisions. Across 11 state-of-the-art LLMs, we uncover distinct behavioral preferences, with average selection rates ranging from 33.7% to 99.8% and model rankings proving highly consistent across language datasets. Subsequent quality evaluation reveals that frequent selection does not guarantee accurate identification of the highest-scoring candidate. Forcing selection-preferring models to create does not consistently yield jokes that outscore their initial selections. For GPT-5.4, forced generations score higher than its selections but still lower than the highest-scoring candidates on average. Similarly, although some creation-preferring models generate jokes that score above the candidate mean, none achieves an average score as high as that of the highest-scoring candidates in any language. Overall, our findings highlight the need to jointly evaluate models' abilities to generate new content, identify high-quality candidates, and decide when to create.

1 INTRODUCTION

Large language models (LLMs) have demonstrated strong capabilities in creative text generation, including poetry composition (Porter & Machery, 2024), story writing (Gómez-Rodríguez & Williams, 2023), and humor generation (Gorenz & Schwarz, 2024). LLM-based evaluations of open-ended responses also show agreement with human preferences on conversational benchmarks (Zheng et al., 2023). These advances enable models to serve as both evaluators and creators. When candidates are already available, a model may therefore face a choice: adopt one of them or attempt to generate a better alternative. Such decisions are relevant to collaborative writing, where existing drafts and newly generated suggestions can both contribute to the final output (Lee et al., 2022).

Prior works have explored candidate selection and fusion (Chen et al., 2023; Jiang et al., 2023; Wang et al., 2025), and feedback-based text refinement (Madaan et al., 2023; Gou et al., 2024). These methods typically specify selection or generation rather than allowing the model to decide which action to take. Humor has recently emerged as an important domain for generation and evaluation research (Zhang et al., 2024; Zhou et al., 2025; Castro et al., 2026). However, when given existing jokes, models' tendencies to adopt one or create a new alternative remain insufficiently characterized, and the associated quality gains remain unclear.

In this work, we investigate ***selection-creation preference*** in LLMs: their relative tendencies to adopt existing content or create new content when both actions are available. We study these preferences through a ***select-or-create task*** in the context of news-based humor. Given a news headline and four candidate jokes, a model is instructed to adopt the best candidate unchanged if it considers that candidate sufficiently good, or to create a new joke otherwise. We quantify its behavioral preference under the given candidates and instructions using the selection rate, and examine how this preference varies across models and language datasets. Behavioral preferences alone, however, do

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

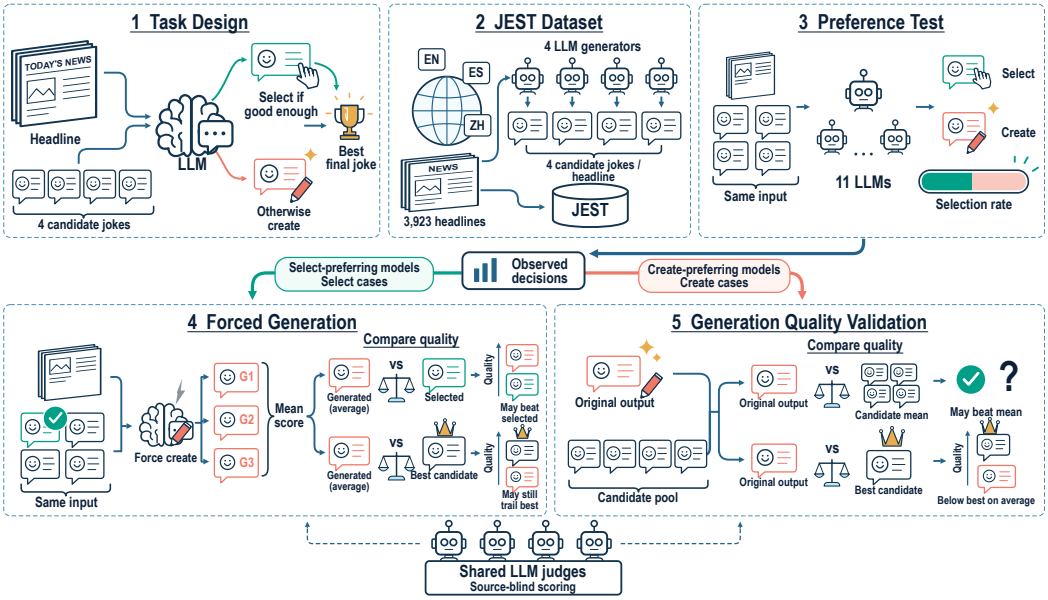


Figure 1: **Framework for evaluating selection–creation preferences and quality gains in LLM humor tasks.** We construct JEST with news headlines and candidate jokes in English (EN), Spanish (ES), and Chinese (ZH). We then measure models' selection-creation preferences and evaluate the quality of created outputs relative to existing candidates.

not establish quality gains. We therefore further examine: (i) whether models identify the highest-scoring candidate when they select, (ii) whether forced generation can improve upon their initial selections, and (iii) whether models that choose to create outperform the available candidates.

To this end, we introduce JEST, a multilingual humor dataset of 3,923 news headlines in English, Spanish, and Chinese, each paired with four independently generated candidate jokes. Using a framework that links selection-creation preferences to quality gains (Figure 1), we first characterize these preferences of 11 state-of-the-art (SOTA) LLMs, then evaluate eight models conditioned on their initial decisions, covering four selection-preferring models and four models more inclined toward creation. Our study uncovers three main findings:

- **Models exhibit distinct selection-creation preferences while maintaining consistent rankings across language datasets.** Average selection rates across the 11 SOTA models range from 33.7% to 99.8%, spanning more frequent creation than selection to near-universal selection. Although absolute selection rates vary across language datasets, model rankings remain highly consistent, with pairwise Spearman correlations of 0.945–0.964.
- **Selection-preferring models miss better candidates, and forced generation offers inconsistent gains.** A strong selection preference does not guarantee accurate identification of the highest-scoring candidate, while requiring these models to create does not consistently improve output quality. On average, GPT-5.4's forced generations outperform its initial selections but remain below the highest-scoring candidates. These findings highlight that gains from generation can coexist with unrealized gains from better candidate selection.
- **Choosing to create does not guarantee gains over the highest-scoring candidates.** Among the four models evaluated on initial CREATE decisions, both Kimi models outperform the candidate mean on average in every language. Nevertheless, all four models fall below the highest-scoring candidates on average in each of the three languages, although their creations surpass this reference on some individual instances. These results show that gains over the candidate average do not necessarily imply an advantage over the highest-scoring available content.

Overall, our findings show that generating new content, identifying high-quality candidates, and deciding when to create need to be evaluated together, since neither frequent selection nor creation alone guarantees better use of available content. When models can either adopt an existing candidate or generate a new one, the central challenge is to identify high-quality candidates and judge when generation is likely to improve upon them.

2 Related Work

Humor Generation and Evaluation. Computational humor research spans humorous headline editing (Hossain et al., 2019), interactive headline creation (Hossain et al., 2020), and pun generation (He et al., 2019). Subsequent LLM-based studies report mixed results that depend on the comparison target: ChatGPT-3.5 jokes were rated funnier than those of laypeople and comparable to Onion headlines (Gorenz & Schwarz, 2024), yet generated humor still lags behind top human submissions (Zhang et al., 2024). Other work finds that ChatGPT often repeats a small set of jokes (Jentzsch & Kersting, 2023) and that professional comedians often judged current LLMs unsuccessful as comedy-writing tools (Mirowski et al., 2024). SemEval-2026 Task 1 evaluates constrained humor generation with human pairwise evaluation (Castro et al., 2026). These studies assign generation or evaluation in advance, and their conclusions depend on which quality reference is used. We instead examine whether models reuse existing jokes or create new ones when both options are available, and whether that choice improves output quality against clearly defined references.

LLM-as-a-Judge. LLM judges enable scalable evaluation of open-ended outputs (Zheng et al., 2023; Liu et al., 2023), with specialized evaluators supporting scoring and pairwise comparison (Kim et al., 2024) and panels aggregating judgments across models (Verga et al., 2024). However, judgments exhibit position bias (Wang et al., 2024; Shi et al., 2025), length bias (Dubois et al., 2024), self-preference (Panickssery et al., 2024), cross-lingual inconsistency (Fu & Liu, 2025), sensitivity to wording, and anchoring in multi-attribute scoring (Stureborg et al., 2024). Self-preference measurements can also conflate evaluator bias with response quality (Chen et al., 2025b). Beyond these biases, models are not reliably better at discriminating among their own responses than at generating them (Jiang et al., 2025). These limitations motivate assessing generation and candidate selection separately, using multiple evaluators, and accounting for source effects. In multilingual humor tasks, automated scores also require validation against human preferences.

Generation, Selection and Revision. Candidate-based methods exploit multiple outputs through verifier-based selection (Cobbe et al., 2021), self-consistency (Wang et al., 2022), and LLM-based consistency selection (Chen et al., 2023). Pairwise ranking also supports document ranking in information retrieval (Qin et al., 2024). LLM-Blender combines ranking with generative fusion (Jiang et al., 2023), while Mixture-of-Agents conditions generation on other models' outputs (Wang et al., 2025). Tree of Thoughts interleaves generation and evaluation during search (Yao et al., 2023), and test-time scaling studies how to allocate compute across search and revision (Snell et al., 2025). Refinement methods use self-feedback (Madaan et al., 2023) or learned correction (Welleck et al., 2022). Intrinsic self-correction can fail in reasoning tasks without external feedback (Huang et al., 2024), although tool-assisted criticism (Gou et al., 2024) and reinforcement learning (Kumar et al., 2025) improve correction in evaluated settings. More closely related, SELF-REDRAFT lets models choose among passing, revising, or redrafting their own code (Chen et al., 2025a). We study choosing between unchanged candidate reuse and new joke creation across three languages, comparing output quality against the selected candidate, candidate mean, and highest-scoring candidate to distinguish behavioral preferences from quality gains.

3 Task Formulation and Evaluation Framework

We organize our investigation of LLMs' selection-creation preferences and their relationship with output quality into two stages. The first characterizes models' behavioral preferences by observing their initial choices under shared candidate pools and task instructions. The second evaluates output quality conditional on those choices, comparing jokes produced through forced generation after selection or during initial creation with the corresponding candidate references. This section defines the task, comparison references, and evaluation metrics. Section 4 details the data sources, evaluated models, sample sizes, and experimental configurations.

Task design. We instantiate the select-or-create task in news-based humor. Each instance consists of a news headline x_i, a target language ℓ, and four candidate jokes $\{c_{ij}\}_{j=1}^{4}$. For each headline, all target models receive the same candidate pool. The model is instructed to adopt the best candidate unchanged if it considers that candidate sufficiently good (SELECT), or to generate a new joke based on the same headline otherwise (CREATE). Here, "sufficiently good" is described qualitatively in terms of humor, creativity, and relevance to the headline, without an explicit numerical threshold.

This task models a reviewing scenario in which the model acts as both evaluator and creator: it assesses existing candidates and decides whether to adopt one or attempt to create a better alternative. To quantify its selection-creation preference under the given candidates and task instructions, we define the selection rate as the number of SELECT decisions divided by the total number of SELECT and CREATE decisions.

Quality Scoring and Candidate References. Our distinction between creating new text and improving output quality is consistent with the view that creativity requires both originality and effectiveness (Runco & Jaeger, 2012). To assess output quality, we employ multiple LLM evaluators (Liu et al., 2023; Verga et al., 2024). Given a news headline and a joke, each evaluator assigns scores from 0 to 10 on four dimensions: relevance to the headline, humor, creativity, and naturalness and fluency. We use the corresponding weights $\mathbf{w} = (0.1, 0.4, 0.3, 0.2)$, placing greater emphasis on humor and creativity to reflect the task's focus on humorous content.

Motivated by concerns about evaluator self-preference and its confounding with response quality (Panickssery et al., 2024; Chen et al., 2025b), we withhold source identities and determine evaluator eligibility from the generator of the scored text. Let $\mathrm{fam}(t)$ denote the model family of the generator of text t. For headline x_i, the set $H(i,t)$ contains evaluators that do not belong to $\mathrm{fam}(t)$ and provide scores from 0 to 10 for text t on each of the four dimensions. Let $s_{h,d}(x_i, t)$ denote evaluator h's score on dimension d, with weight w_d. We compute a weighted score for each eligible evaluator and then average:

$$Q(t \mid x_i) = \frac{1}{|H(i,t)|} \sum_{h \in H(i,t)} \sum_{d=1}^{4} w_d s_{h,d}(x_i, t). \tag{1}$$

This rule is applied independently to each text, so candidates and generated jokes in the same comparison may be scored by different evaluator subsets. To test whether the quality comparisons depend on this choice, Section 5.4 recomputes them with common cross-family panels.

For brevity, we suppress the dependence on x_i below and write $Q(t)$. We use three candidate references: the score of the initially selected candidate, denoted by $Q_{\mathrm{sel},i}$; the mean candidate score, $Q_{\mathrm{mean},i} = \frac{1}{4}\sum_{j=1}^{4} Q(c_{ij})$; and the highest candidate score, $Q_{\mathrm{best},i} = \max_{1 \le j \le 4} Q(c_{ij})$. These references distinguish improvements over the model's actual selection or the candidate average from gains beyond the highest-scoring available candidate.

Quality Comparison Experiments. We conduct two complementary experiments conditioned on models' initial decisions. For selection-preferring models, an initial SELECT decision provides no newly generated joke to evaluate. Experiment A therefore requires generation on these instances to test whether it can improve upon the selected candidate and surpass the highest-scoring available candidate. For models more inclined toward creation, Experiment B directly evaluates jokes produced in their initial CREATE decisions against the candidate mean and highest-scoring candidate. Together, these experiments assess the quality gains from forced generation after selection and from initial creation.

Experiment A: Forced generation after selection. For selection-preferring models, we evaluate sampled instances on which they initially made a SELECT decision. In a fresh conversation for each generation, we provide the same headline and four candidate jokes and ask the model to evaluate the candidates and create a better joke based on the headline, regardless of whether any candidate is sufficiently good. The prompt instructs the model not to select a candidate. For each model, let S denote the sampled instances with an initial SELECT decision. For each $i \in S$, we obtain three forced-generation jokes and denote the arithmetic mean of their quality scores by $\overline{Q}_{S,i}$. We define quality differences relative to the selected candidate and the highest-scoring candidate as

$$\Delta^{A}_{\mathrm{sel},i} = \overline{Q}_{S,i} - Q_{\mathrm{sel},i}, \tag{2a}$$

$$\Delta^{A}_{\mathrm{best},i} = \overline{Q}_{S,i} - Q_{\mathrm{best},i}. \tag{2b}$$

Here, A denotes Experiment A, and positive differences indicate higher mean quality for forced generation relative to the corresponding candidate reference. The first comparison measures improvement over the model's actual selection, while the second tests whether generation surpasses the highest-scoring available candidate. Their difference, $\Delta^{A}_{\mathrm{sel},i} - \Delta^{A}_{\mathrm{best},i} = Q_{\mathrm{best},i} - Q_{\mathrm{sel},i}$, is the candidate selection gap. Thus, forced generation may improve upon the initial selection while still falling short of an existing alternative.

Experiment B: Quality of initial creation. For models more inclined toward creation, we evaluate instances on which they initially made a CREATE decision. Let $G_{\text{free},i}$ denote the joke produced in that original decision. Since no candidate was selected, we directly compare its quality with the candidate mean and highest candidate score:

$$\Delta^{B}_{\text{mean},i} = Q(G_{\text{free},i}) - Q_{\text{mean},i}, \tag{3a}$$

$$\Delta^{B}_{\text{best},i} = Q(G_{\text{free},i}) - Q_{\text{best},i}. \tag{3b}$$

Here, B denotes Experiment B, and positive differences favor the generated joke. The first comparison tests whether initial creation exceeds the candidate average, while the second assesses whether it surpasses the highest-scoring available candidate. Thus, outperforming the candidate average does not necessarily establish an advantage over the highest-scoring candidate.

4 EXPERIMENTAL SETTINGS

Dataset. We construct JEST (**J**okes for **E**valuating **S**election versus creation **T**endencies), a multilingual dataset for evaluating LLMs in the select-or-create task. JEST uses 3,923 news headlines from SemEval-2026 Task 1 (Castro et al., 2026): 1,375 in English, 1,375 in Spanish, and 1,173 in Chinese (excluding two sensitive Chinese headlines). Each headline is paired with four jokes, independently generated by Qwen3-Max (Team, 2025), DeepSeek-V3.2 (Liu et al., 2025), Kimi-K2.5 (Team et al., 2026), and GLM-5 (Zeng et al., 2026), with one joke per model. We use a zero-shot prompt that asks each model to produce a concise, creative, and funny joke based on the given headline, without iterative refinement. For candidate generation, we set the temperature to 0.85 and top-p to 0.9. We disable reasoning when supported; otherwise, we use the lowest available reasoning level.

Models. We evaluate 11 target models from five model families to characterize variation in selection-creation preferences: DeepSeek (V3.2, V4-Flash, and V4-Pro (DeepSeek-AI, 2026)), Kimi (K2.5 and K2.6 (Moonshot AI, 2026)), Qwen (3-Max, 3.6-Flash (Alibaba Cloud, 2026), and 3.6-27B (Qwen Team, 2026a)), GPT (5.2 (OpenAI, 2025) and 5.4 (OpenAI, 2026a)), and Gemini (3.1 Pro Preview (Google DeepMind, 2026)). For quality evaluation, we use four LLM evaluators from distinct model families: GPT-5.5 (OpenAI, 2026b), Gemini 3.1 Pro Preview, Qwen3.7-Max (Qwen Team, 2026b), and GLM-5.2 (Zeng et al., 2026).

Implementation Details. We conduct the experiments in two stages: behavioral preference measurement and quality evaluation conditioned on models' initial decisions. First, each of the 11 target models makes one initial select-or-create decision for each of the 3,923 JEST instances, yielding 43,153 decision trials. For each trial, the four candidates are presented in independently randomized order. We compute selection rates among decisions separately for each model and language. Based on these observed rates, we select representative selection-preferring models for Experiment A and models more inclined toward creation for Experiment B. For each model and language, we use a fixed random seed to sample up to 200 instances with the corresponding initial decision, including all available instances when fewer than 200 are eligible. Experiment A uses initial SELECT instances and obtains three forced generations per instance, whereas Experiment B uses initial CREATE instances and evaluates the jokes produced in those original decisions. Across both stages, we set the temperature to 0.5 for initial decisions, forced generation, and evaluator scoring; candidate generation during dataset construction uses the separate configuration described above. Thinking mode is disabled when supported; otherwise, we use the lowest available reasoning level. For quality evaluation, we score both candidate and generated jokes using the rule defined above. Each evaluator receives only the headline and the joke, with no source identity provided, and evaluators from the same model family as the generator of the evaluated text are excluded from score aggregation.

5 RESULTS & ANALYSIS

In this section, we first examine how selection-creation preferences vary across models and language datasets. We then examine how accurately selection-preferring models identify the highest-scoring candidates and whether forced generation improves upon their initial selections or surpasses the highest-scoring available candidates. We also evaluate jokes produced in initial CREATE decisions against the candidate mean and highest candidate score. Finally, we conduct a robustness analysis of our quality scoring approach.

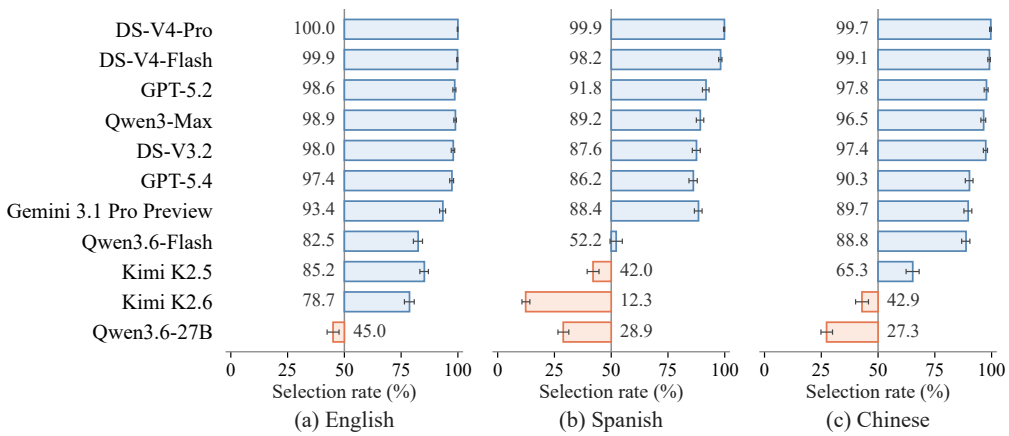


Figure 2: **Selection rates across models and language datasets.** Bars extend from 50% (equal selection and creation) to the observed rate, with blue and orange indicating rates above and below 50%, respectively. Error bars show 95% Wilson confidence intervals. DS denotes DeepSeek.

5.1 Do LLMs Differ in Selection-Creation Preferences?

Figure 2 shows selection rates for the 11 target models across the three language datasets. Under shared candidate pools and task instructions within each language, models exhibit markedly different selection-creation preferences. DeepSeek-V4-Pro selects an existing candidate on 99.7%–100.0% of instances across the three languages, and DeepSeek-V4-Flash has selection rates above 98% throughout. In contrast, Qwen3.6-27B selects on fewer than 50% of instances in every language, with an average selection rate of 33.7%. Selection rates also vary across language datasets: all 11 models select less frequently on the Spanish dataset than on the English dataset, while rates on the Chinese dataset generally fall between the two. Despite these differences in absolute rates, model rankings remain highly consistent, with Spearman correlations of 0.945 for English–Spanish, 0.964 for English–Chinese, and 0.945 for Spanish–Chinese.

To examine which candidates models adopt, we further analyze the source distribution of candidates chosen in initial SELECT decisions. GLM-5 candidates account for the largest share of selections in each language: 58.0%, 65.6%, and 59.7% in English, Spanish, and Chinese, respectively. DeepSeek-V3.2 candidates account for the smallest shares, at 9.1%, 8.3%, and 12.9%, respectively. Thus, selections are unevenly distributed across candidate sources, with GLM-5 candidates most frequently adopted across all three language datasets.

Overall, **LLMs exhibit distinct selection-creation preferences**, ranging from near-universal selection by DeepSeek-V4-Pro to more frequent creation than selection by Qwen3.6-27B in all three languages. Moreover, **relative model rankings remain stable across language datasets despite changes in absolute selection rates**. These behavioral patterns alone do not establish whether models' decisions improve output quality. We therefore evaluate DeepSeek-V4-Flash, DeepSeek-V4-Pro, Qwen3-Max, and GPT-5.4 in Experiment A to assess forced generation after initial selection. We further evaluate Kimi-K2.5, Kimi-K2.6, Qwen3.6-Flash, and Qwen3.6-27B, which have comparatively lower selection rates, in Experiment B to assess the quality of their initial creations.

5.2 How Well Do Models Select, and Can Forced Generation Help?

This subsection presents the results of Experiment A. We examine the quality outcomes of selection-preferring models in two steps. First, we assess whether their initial selections identify the highest-scoring available candidates. We then test whether forced generation improves upon these selections and whether it surpasses the highest-scoring candidate in the pool.

Table 1 reports the top-score hit rate and mean candidate selection gap on evaluated initial SELECT instances, measuring how often the selected candidate attains the highest score and how far its score falls below that reference on average. Across the four models, hit rates range from 47.0% to 61.8%. DeepSeek-V4-Pro has the highest observed hit rate of 61.8% and the smallest mean gap of 0.327

Table 1: Candidate selection performance on sampled initial SELECT instances, pooled across three language datasets ($n = 600$ per model). A hit means the selected candidate attains the highest candidate score; the mean gap is the highest score minus the selected score. DS denotes DeepSeek.

	DS-V4-Pro	DS-V4-Flash	Qwen3-Max	GPT-5.4
Top-score hit rate (%)	61.8	59.2	56.5	47.0
Mean selection gap (points)	0.327	0.394	0.394	0.649

score points, whereas GPT-5.4 has the lowest hit rate of 47.0% and the largest mean gap of 0.649. Thus, **frequent selection does not guarantee accurate identification of the highest-scoring candidate**. Even models that strongly prefer selection leave higher-scoring candidates unused on a substantial share of evaluated instances.

Given these selection gaps, we next ask whether requiring models to create new jokes improves output quality. Figure 3 shows the mean score of three forced generations minus the selected candidate's score for each headline. Full paired results for both candidate references appear in Appendix Table 3. Averaged across headlines, the quality differences are negative for DeepSeek-V4-Flash, DeepSeek-V4-Pro, and Qwen3-Max in every language, but positive for GPT-5.4, whose English gain is small. DeepSeek-V4-Flash and Qwen3-Max improve upon selection on fewer than 35% of evaluated headlines in each language, whereas GPT-5.4 does so on 48.5%–55.5%, a majority only in Spanish and Chinese. Thus, **forced generation does not consistently improve upon initial selection**, even though initial selections do not reliably identify the highest-scoring candidates.

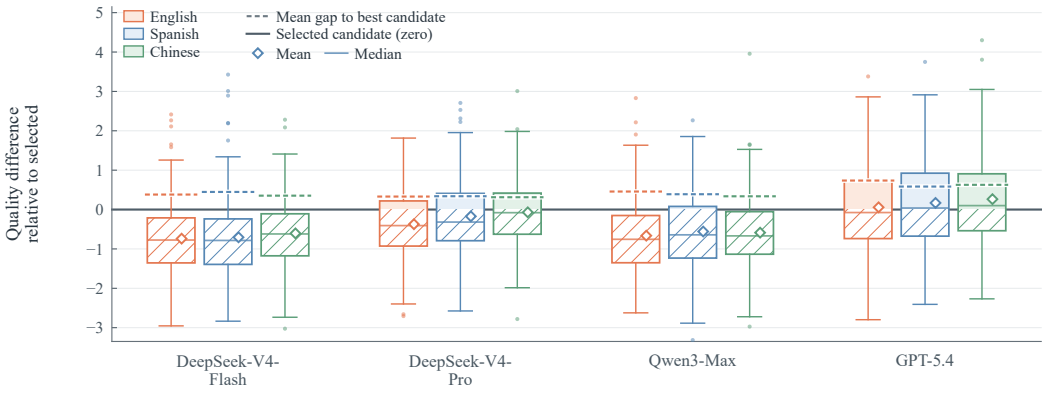


Figure 3: **Forced-generation quality relative to initial selection.** Each observation is the mean score of three forced generations minus the selected candidate's score for one headline. Boxes span the interquartile range; solid lines and diamonds mark medians and means, respectively. Hatching marks negative differences; dashed crossbars show the mean gap between the highest-scoring and selected candidates for each model–language group.

The observed selection gaps also matter when interpreting gains from generation: **outperforming the selected candidate does not necessarily mean surpassing the highest-scoring available candidate.** GPT-5.4, which has the largest observed mean selection gap, illustrates this distinction. Its forced generations score above its selected candidates on average in every language (by a small margin in English), but still remain below the highest-scoring candidates. On some evaluated headlines, the mean score of three generations surpasses the highest-scoring candidate, but these cases do not reverse the average shortfall. Thus, GPT-5.4's gains from forced generation coexist with opportunities to obtain higher-scoring outputs through better candidate selection.

In summary, these findings reveal two limitations of the selection-preferring models evaluated here: **their frequent selections do not reliably identify the highest-scoring candidates, and forcing them to generate offers inconsistent gains over their actual choices.** Even when generation improves upon initial selection, it can still fall short of the highest-scoring available content on average. Assessing these models therefore requires examining both candidate selection performance

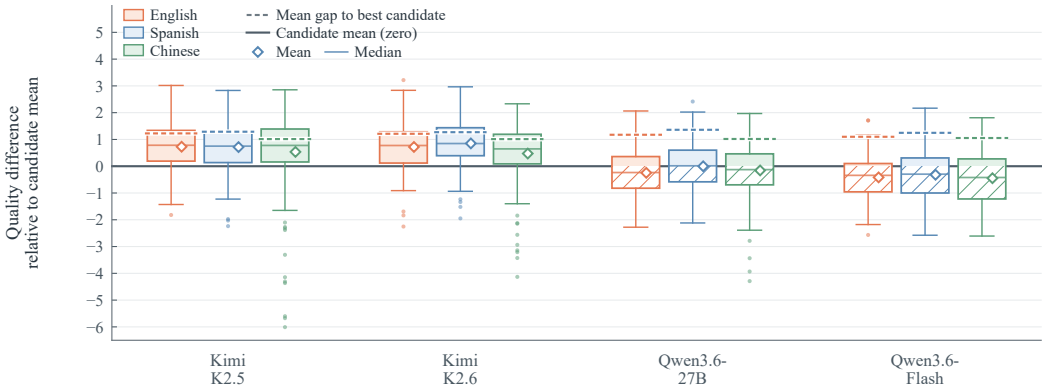


Figure 4: **Initial-creation quality relative to the candidate mean.** Each observation is the original creation's score minus the mean candidate score for one headline. Boxes span the interquartile range; solid lines and diamonds mark medians and means, respectively. Hatching marks negative differences; dashed crossbars show the mean gap between the highest-scoring candidate and the candidate mean for each model–language group.

and generation gains, while distinguishing improvements over actual selections from gains beyond the available candidate pool.

5.3 Does Initial Creation Outperform Candidates?

This subsection presents the results of Experiment B. We evaluate jokes produced in sampled initial CREATE decisions by Kimi-K2.5, Kimi-K2.6, Qwen3.6-27B, and Qwen3.6-Flash, all of which chose to create more often than the models evaluated in Experiment A. Figure 4 shows per-headline quality differences relative to the candidate mean; full numerical results appear in Appendix Table 4.

Both Kimi models have positive mean differences of 0.47–0.85 points across all three languages, with medians and interquartile ranges also above zero. Their creations exceed the candidate mean on 76.0%–86.5% of evaluated headlines, although some Chinese headlines show large negative outliers. In contrast, Qwen3.6-Flash has negative mean differences in every language, while Qwen3.6-27B has negative mean differences in English and Chinese and a near-zero difference in Spanish.

Against the highest-scoring candidate, however, all four models have negative mean differences in every language (Appendix Table 4). Average shortfalls range from 0.42 to 0.57 points for the Kimi models and from 1.18 to 1.57 points for the Qwen models. Initial creations surpass this reference on some headlines, but no model does so on a majority in any language. The highest observed win rate is 40.0% (Kimi-K2.5, Chinese), whereas both Qwen models have win rates of at most 9.2%.

Taken together, these results show that **initial creation can exceed the candidate average, but this advantage varies across models**. Moreover, **none of the four evaluated models surpasses the highest-scoring candidates on average in any language**. Under our evaluation protocol, choosing to create therefore does not guarantee a quality gain beyond what the existing candidate pool offers.

5.4 How Robust Are the Quality Comparisons?

Under the scoring rule in Equation 1, we exclude evaluators from the generator's model family separately for each joke. A generated joke and its candidate references may therefore be scored by different evaluator subsets; we refer to the resulting estimates as *source-filtered*. To assess sensitivity to evaluator composition, we recompute the quality differences using a *shared-panel* that evaluates each generated joke and all four candidates. When evaluating DeepSeek, Kimi, or Qwen models, we use GPT-5.5 and Gemini 3.1 Pro Preview as the shared-panel; when evaluating GPT-5.4, we use Gemini 3.1 Pro Preview alone, so that no GPT evaluator scores GPT-generated text. We recompute the candidate mean and highest candidate score from the shared-panel scores. Appendix A reports the source-filtered and shared-panel estimates side by side in Table 2.

The main findings remain unchanged under shared-panel scoring. Forced generation remains below the initially selected candidate on average for DeepSeek-V4-Flash and Qwen3-Max in each language. Jokes initially created by both Kimi models remain above the candidate mean on average in each language. Across both experiments, generated jokes score below the highest-scoring candidate on average for all eight evaluated models in each of the three languages under both scoring approaches. Every corresponding pointwise 95% bootstrap interval also lies entirely below zero.

Shared-panel scoring affects some comparisons close to zero. The mean differences change sign for DeepSeek-V4-Pro relative to its selection in Chinese and for Qwen3.6-27B relative to the candidate mean in Spanish and Chinese; all three shared-panel intervals include zero. GPT-5.4 retains positive mean differences relative to its selections in every language, although the English interval includes zero under both scoring approaches and the Chinese interval does so under the shared-panel. **Nevertheless, in each of these cases, the mean difference relative to the highest-scoring candidate remains negative, with its interval entirely below zero.**

6 Discussion

Selection-creation preferences characterize models' decisions when selection and creation are both available, but do not by themselves indicate how effectively the models use the candidate pool. Our results expose two distinct gaps: the selection-preferring models evaluated here often miss the highest-scoring candidate, while generated outputs may outperform the selected candidate or candidate mean on average yet remain below the highest-scoring candidate. This shortfall relative to the highest-scoring candidate persists under shared-panel scoring, whereas some comparisons close to zero are sensitive to evaluator composition. The findings therefore motivate evaluating behavioral preference, candidate discrimination, and generation quality as distinct but related aspects of performance.

The interpretation of these findings is shaped by the task's acceptance criterion, quality measure, and candidate pool. In the select-or-create prompt, models are asked to select the best candidate if any is sufficiently good and to create a new joke otherwise. This qualitative criterion has no numerical cutoff and does not require a prediction about the quality of a later generation. A forced generation that scores above the initially selected joke therefore does not by itself establish that the model should have chosen to create at the outset. Our quality comparisons rely on weighted LLM-judge scores. The highest-scoring candidate under this measure may not match human preferences, and some observed mean differences are small on this scale, leaving their perceptual significance unclear. Finally, each headline has four model-generated candidates, so the findings may depend on the composition of the pool. Given prior evidence that LLM and expert assessments can diverge in creative writing (Chakrabarty et al., 2024), human judgments are needed to determine whether the score differences in our humor task matter to intended audiences. Evaluations with other candidate pools would further test the scope of the findings.

7 Conclusion

In this work, we investigate LLMs' selection-creation preferences and their relationship to output quality using an evaluation framework and JEST, a multilingual humor dataset. Across 11 SOTA models, preferences range from near-universal selection to more frequent creation than selection; model rankings remain highly consistent across languages despite differences in absolute selection rates. In our LLM-based quality evaluation, four selection-preferring models identify a highest-scoring candidate on 47.0%–61.8% of sampled initial SELECT instances pooled across languages. Forced generation does not consistently improve upon initial selections; GPT-5.4's forced generations score above its selections but below the highest-scoring candidates on average. Among the four models evaluated on initial CREATE instances, both Kimi models' creations exceed the candidate mean on average in every language, yet all four models' creations fall below the highest-scoring candidates on average in every language. Overall, models can miss the highest-scoring candidate when selecting; generated outputs can also score higher than the selected candidate or the candidate mean on average while still scoring below the highest-scoring candidate. In select-or-create tasks, we should therefore evaluate how well models identify high-quality candidates and generate new alternatives, as well as how they choose between selection and creation.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic candidate jokes for the JEST dataset and to score joke quality with LLM evaluators. Additionally, we used generative AI tools to draft the paper's framework figure. We have reviewed all AI-assisted work: the generated jokes, evaluator scores, and framework figure were checked by the authors against the experimental records. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Alibaba Cloud. qwen3.6-flash. https://www.alibabacloud.com/help/en/model-studio/qwen3-6-flash, September 2026.

Santiago Castro, Luis Chiruzzo, Santiago Góngora, Naihao Deng, Salar Rahili, Ignacio Sastre, Aiala Rosá, Victoria Amoroso, Guillermo Rey, Guillermo Moncecchi, et al. Semeval-2026 task 1: Mwahaha, models write automatic humor and humans annotate. In *Proceedings of the 20th International Workshop on Semantic Evaluation (2026)*, pp. 3797–3822, 2026.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–34, 2024.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. Universal self-consistency for large language model generation. *arXiv preprint arXiv:2311.17311*, 2023.

Yixiang Chen, Tianshi Zheng, Shijue Huang, Zhitao He, and Yi R Fung. Self-redraft: Eliciting intrinsic exploration-exploitation balance in test-time scaling for code generation. *arXiv preprint arXiv:2511.02854*, 2025a.

Zhi-Yuan Chen, Hao Wang, Xinyu Zhang, Enrui Hu, and Yankai Lin. Beyond the surface: Measuring self-preference in llm judgments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1653–1672, 2025b.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Xiyan Fu and Wei Liu. How reliable is multilingual llm-as-a-judge? In *EMNLP (Findings)*, pp. 11040–11053, 2025.

Carlos Gómez-Rodríguez and Paul Williams. A confederacy of models: A comprehensive evaluation of llms on creative writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14504–14528, 2023.

Google DeepMind. Gemini 3.1 pro model card. https://deepmind.google/models/model-cards/gemini-3-1-pro, February 2026.

Drew Gorenz and Norbert Schwarz. How funny is chatgpt? a comparison of human-and ai-produced jokes. *Plos one*, 19(7):e0305364, 2024.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yujiu Yang, Nan Duan, Weizhu Chen, et al. Critic: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*, volume 2024, pp. 57734–57811, 2024.

He He, Nanyun Peng, and Percy Liang. Pun generation with surprise. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)*, pp. 1734–1744, 2019.

Nabil Hossain, John Krumm, and Michael Gamon. "president vows to cut< taxes> hair": Dataset and analysis of creative text editing for humorous headlines. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)*, pp. 133–142, 2019.

Nabil Hossain, John Krumm, Tanvir Sajed, and Henry Kautz. Stimulating creativity with funlines: A case study of humor generation in headlines. In *Proceedings of the 58th annual meeting of the association for computational linguistics: System demonstrations*, pp. 256–262, 2020.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International conference on learning representations*, volume 2024, pp. 32808–32824, 2024.

Sophie Jentzsch and Kristian Kersting. Chatgpt is fun, but it is not funny! humor is still challenging large language models. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pp. 325–340, 2023.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14165–14178, 2023.

Dongwei Jiang, Jingyu Zhang, Orion Weller, Nathaniel Weir, Benjamin Van Durme, and Daniel Khashabi. Self-[in] correct: Llms struggle with discriminating self-generated responses. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 24266–24275, 2025.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4334–4353, 2024.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. Training language models to self-correct via reinforcement learning. In *International Conference on Learning Representations*, volume 2025, pp. 54523–54549, 2025.

Mina Lee, Percy Liang, and Qian Yang. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pp. 1–19, 2022.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Piotr Mirowski, Juliette Love, Kory Mathewson, and Shakir Mohamed. A robot walks into a bar: Can language models serve as creativity supporttools for comedy? an evaluation of llms' humour alignment with comedians. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1622–1636, 2024.

Moonshot AI. Kimi K2.6: Advancing open-source coding. `https://www.kimi.ai/blog/kimi-k2-6`, April 2026.

OpenAI. Update to GPT-5 system card: GPT-5.2. `https://openai.com/index/gpt-5-system-card-update-gpt-5-2/`, December 2025.

OpenAI. GPT-5.4 thinking system card. `https://openai.com/index/gpt-5-4-thinking-system-card/`, March 2026a.

OpenAI. GPT-5.5 system card. https://openai.com/index/gpt-5-5-system-card/, April 2026b.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802, 2024.

Brian Porter and Edouard Machery. Ai-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Scientific Reports*, 14(1):26133, 2024.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, et al. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1504–1518, 2024.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026a. URL https://qwen.ai/blog?id=qwen3.6-27b.

Qwen Team. Qwen3.7: The agent frontier, May 2026b. URL https://qwen.ai/blog?id=qwen3.7.

Mark A Runco and Garrett J Jaeger. The standard definition of creativity. *Creativity research journal*, 24(1):92–96, 2012.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in llm-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 292–314, 2025.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, volume 2025, pp. 10131–10165, 2025.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*, 2024.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Ziwei Chai, Y Charles, HS Che, Cheng Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Qwen Team. Qwen3-max: Just scale it, September 2025.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Y Zou. Mixture-of-agents enhances large language model capabilities. In *International Conference on Learning Representations*, volume 2025, pp. 33944–33963, 2025.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, et al. Large language models are not fair evaluators. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9440–9450, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. Generating sequences by learning to self-correct. *arXiv preprint arXiv:2211.00053*, 2022.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Jifan Zhang, Lalit Jain, Yang Guo, Jiayi Chen, Kuan L Zhou, Siddharth Suresh, Andrew Wagenmaker, Scott Sievert, Timothy Rogers, Kevin Jamieson, et al. Humor in ai: Massive scale crowdsourced preferences and benchmarks for cartoon captioning. *Advances in Neural Information Processing Systems*, 37:125264–125286, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Kuan Lok Zhou, Jiayi Chen, Siddharth Suresh, Reuben Narad, Timothy T Rogers, Lalit K Jain, Robert D Nowak, Bob Mankoff, and Jifan Zhang. Bridging the creativity understanding gap: Small-scale human alignment enables expert-level humor ranking in llms. In *EMNLP (Findings)*, pp. 16273–16287, 2025.

A Robustness of Quality Evaluation

This appendix details the shared-panel analysis summarized in Section 5.4. The analysis reuses the recorded evaluator scores and retains the dimension weights specified in Equation 1.

Scoring configurations. The scoring rule, labeled ***Source-filtered*** in Table 2, excludes evaluators from the model family of each scored text's generator. Thus, Qwen-generated candidates are scored without Qwen evaluators, GLM-generated candidates without GLM evaluators, and each newly generated joke without evaluators from its generator's family. Because this exclusion is applied separately to each text, a generated joke and its candidate references may be scored by different evaluator sets. By contrast, ***Shared-panel*** scoring uses the same evaluators for each generated joke and all four candidates within a comparison. When evaluating DeepSeek, Kimi, or Qwen models, the panel comprises GPT-5.5 and Gemini 3.1 Pro Preview; when evaluating GPT-5.4, it comprises Gemini 3.1 Pro Preview alone, so that no GPT evaluator scores GPT-generated text. Candidate means and highest candidate scores are recomputed from the shared-panel scores. In Experiment A, the initially selected candidate is held fixed, but its score is recomputed under the shared-panel.

Sampling and uncertainty. For each model and language, we sample up to 200 instances with the relevant initial decision; for Qwen3.6-Flash in Chinese, all 131 eligible CREATE instances are included. In Experiment A, we average the scores of three forced generations per headline before computing differences; in Experiment B, we use the score of the joke produced in the initial CREATE decision. We form pointwise 95% percentile bootstrap intervals using 10,000 headline resamples. Each resample draws headlines within a model and language, keeping the generated and candidate scores for each sampled headline together. We recompute the mean paired difference for each resample and use the 2.5th and 97.5th percentiles of these means as the interval bounds. These intervals reflect uncertainty from sampling headlines, conditional on the recorded outputs and evaluator scores.

Results. Table 2 presents the results of Experiments A and B under source-filtered and shared-panel scoring. In Experiment A, the shared-panel mean differences relative to the highest-scoring candidate move closer to zero for both DeepSeek models and Qwen3-Max; for Qwen3-Max in English, the estimate changes from -1.121 to -0.727. For GPT-5.4, the difference is nearly unchanged in English but becomes more negative in Spanish and Chinese, moving from -0.364 to -0.674 in Chinese. In Experiment B, the shortfall relative to the highest-scoring candidate narrows for all four models; for Qwen3.6-Flash in Chinese, it changes from -1.501 to -1.042. Both Kimi models also remain above the candidate mean in every language; for Kimi-K2.6 in Spanish, the mean difference falls from $+0.852$ to $+0.673$, with both intervals entirely above zero. Some comparisons close to zero are more sensitive: Qwen3.6-27B's Spanish difference relative to the candidate mean changes from -0.002 to $+0.116$, with both intervals containing zero. **The central finding remains unchanged despite these shifts**: under both scoring approaches, generated jokes fall below the highest-scoring candidate on average for every model in each language, with every corresponding pointwise 95% interval entirely below zero.

Table 2: **Source-filtered scoring versus a shared evaluator panel.** Mean score differences (generated output minus reference) with pointwise 95% bootstrap intervals. Pref.: Select = selection-preferring models (forced generation); Create = creation-preferring models (initial joke). Ref.: S = initially selected candidate; M = mean of four candidates; B = highest-scoring candidate. DS denotes DeepSeek; EN, ES, and ZH denote English, Spanish, and Chinese. Select rows average three forced-generation scores per headline. Analyses use up to 200 headlines per model and language; Qwen3.6-Flash ZH has 131 eligible cases. Shaded cells indicate intervals that include zero.

Pref.	Model	Lang.	Ref.	Source-filtered	Shared-panel
Select	DS-V4-Flash	EN	S	-0.739 $[-0.865, -0.611]$	-0.468 $[-0.580, -0.351]$
			B	-1.120 $[-1.218, -1.019]$	-0.843 $[-0.927, -0.756]$
		ES	S	-0.705 $[-0.843, -0.558]$	-0.531 $[-0.655, -0.399]$
			B	-1.152 $[-1.247, -1.057]$	-0.959 $[-1.052, -0.866]$
		ZH	S	-0.605 $[-0.718, -0.487]$	-0.452 $[-0.557, -0.343]$
			B	-0.958 $[-1.045, -0.867]$	-0.803 $[-0.888, -0.717]$
Select	DS-V4-Pro	EN	S	-0.372 $[-0.498, -0.249]$	-0.121 $[-0.237, -0.006]$
			B	-0.702 $[-0.804, -0.603]$	-0.469 $[-0.556, -0.381]$
		ES	S	-0.172 $[-0.303, -0.041]$	-0.037 $[-0.164, 0.091]$
			B	-0.511 $[-0.608, -0.414]$	-0.357 $[-0.447, -0.267]$
		ZH	S	-0.070 $[-0.184, 0.043]$	$+0.058$ $[-0.049, 0.165]$
			B	-0.383 $[-0.469, -0.299]$	-0.302 $[-0.382, -0.223]$
Select	Qwen3-Max	EN	S	-0.664 $[-0.796, -0.527]$	-0.274 $[-0.399, -0.146]$
			B	-1.121 $[-1.218, -1.025]$	-0.727 $[-0.814, -0.639]$
		ES	S	-0.558 $[-0.700, -0.415]$	-0.289 $[-0.411, -0.164]$
			B	-0.947 $[-1.052, -0.844]$	-0.637 $[-0.733, -0.542]$
		ZH	S	-0.587 $[-0.709, -0.462]$	-0.236 $[-0.347, -0.120]$
			B	-0.923 $[-1.017, -0.831]$	-0.540 $[-0.626, -0.454]$
Select	GPT-5.4	EN	S	$+0.057$ $[-0.106, 0.222]$	$+0.103$ $[-0.078, 0.286]$
			B	-0.680 $[-0.786, -0.576]$	-0.686 $[-0.810, -0.565]$
		ES	S	$+0.168$ $[0.005, 0.329]$	$+0.217$ $[0.020, 0.408]$
			B	-0.416 $[-0.529, -0.300]$	-0.519 $[-0.660, -0.380]$
		ZH	S	$+0.264$ $[0.114, 0.421]$	$+0.092$ $[-0.088, 0.279]$
			B	-0.364 $[-0.467, -0.261]$	-0.674 $[-0.815, -0.535]$
Create	Kimi-K2.5	EN	M	$+0.728$ $[0.612, 0.843]$	$+0.756$ $[0.646, 0.864]$
			B	-0.498 $[-0.625, -0.375]$	-0.303 $[-0.419, -0.189]$
		ES	M	$+0.715$ $[0.596, 0.833]$	$+0.586$ $[0.471, 0.698]$
			B	-0.574 $[-0.706, -0.439]$	-0.503 $[-0.633, -0.372]$
		ZH	M	$+0.536$ $[0.340, 0.719]$	$+0.463$ $[0.281, 0.631]$
			B	-0.479 $[-0.672, -0.295]$	-0.358 $[-0.536, -0.188]$
Create	Kimi-K2.6	EN	M	$+0.720$ $[0.604, 0.837]$	$+0.684$ $[0.569, 0.797]$
			B	-0.492 $[-0.623, -0.359]$	-0.344 $[-0.460, -0.228]$
		ES	M	$+0.852$ $[0.736, 0.965]$	$+0.673$ $[0.547, 0.794]$
			B	-0.419 $[-0.530, -0.309]$	-0.383 $[-0.505, -0.261]$
		ZH	M	$+0.474$ $[0.321, 0.618]$	$+0.351$ $[0.206, 0.496]$
			B	-0.539 $[-0.695, -0.393]$	-0.464 $[-0.609, -0.321]$
Create	Qwen3.6-27B	EN	M	-0.247 $[-0.360, -0.127]$	-0.056 $[-0.171, 0.063]$
			B	-1.422 $[-1.542, -1.299]$	-1.033 $[-1.148, -0.914]$
		ES	M	-0.002 $[-0.124, 0.119]$	$+0.116$ $[-0.005, 0.234]$
			B	-1.362 $[-1.496, -1.225]$	-1.065 $[-1.197, -0.934]$
		ZH	M	-0.157 $[-0.297, -0.026]$	$+0.054$ $[-0.072, 0.172]$
			B	-1.175 $[-1.320, -1.035]$	-0.766 $[-0.894, -0.644]$
Create	Qwen3.6-Flash	EN	M	-0.417 $[-0.528, -0.308]$	-0.254 $[-0.370, -0.140]$
			B	-1.516 $[-1.649, -1.388]$	-1.181 $[-1.312, -1.055]$
		ES	M	-0.319 $[-0.448, -0.192]$	-0.158 $[-0.280, -0.035]$
			B	-1.566 $[-1.702, -1.431]$	-1.214 $[-1.339, -1.086]$
		ZH	M	-0.448 $[-0.611, -0.281]$	-0.180 $[-0.337, -0.019]$
			B	-1.501 $[-1.682, -1.315]$	-1.042 $[-1.213, -0.868]$

B Detailed Quality Comparisons for Experiments A and B

Table 3: **Experiment A: forced generation after selection.** Mean paired score differences (generation minus reference) with pointwise 95% bootstrap CIs from 10,000 headline resamples. Scores for three forced generations are averaged per headline; $n = 200$ per model and language. Scoring is source-filtered. Win rates count strictly positive differences; ties count as non-wins. Blue/orange shading indicates CIs entirely above/below zero; white indicates CIs containing zero. EN, ES, and ZH denote English, Spanish, and Chinese.

		vs. selected candidate		vs. highest-scoring candidate	
Model	**Lang.**	**Mean Δ [95% CI]**	**Win (%)**	**Mean Δ [95% CI]**	**Win (%)**
DeepSeek-V4-Flash	EN	$\mathbf{-0.739}$ [−0.865, −0.611]	18.5	$\mathbf{-1.120}$ [−1.218, −1.019]	6.5
	ES	$\mathbf{-0.705}$ [−0.843, −0.558]	20.5	$\mathbf{-1.152}$ [−1.247, −1.057]	3.5
	ZH	$\mathbf{-0.605}$ [−0.718, −0.487]	21.0	$\mathbf{-0.958}$ [−1.045, −0.867]	4.5
DeepSeek-V4-Pro	EN	$\mathbf{-0.372}$ [−0.498, −0.249]	34.5	$\mathbf{-0.702}$ [−0.804, −0.603]	14.5
	ES	$\mathbf{-0.172}$ [−0.303, −0.041]	35.5	$\mathbf{-0.511}$ [−0.608, −0.414]	22.5
	ZH	$\mathbf{-0.070}$ [−0.184, +0.043]	44.5	$\mathbf{-0.383}$ [−0.469, −0.299]	25.5
Qwen3-Max	EN	$\mathbf{-0.664}$ [−0.796, −0.527]	22.0	$\mathbf{-1.121}$ [−1.218, −1.025]	5.5
	ES	$\mathbf{-0.558}$ [−0.700, −0.415]	27.0	$\mathbf{-0.947}$ [−1.052, −0.844]	8.0
	ZH	$\mathbf{-0.587}$ [−0.709, −0.462]	23.5	$\mathbf{-0.923}$ [−1.017, −0.831]	9.0
GPT-5.4	EN	$\mathbf{+0.057}$ [−0.106, +0.222]	48.5	$\mathbf{-0.680}$ [−0.786, −0.576]	18.0
	ES	$\mathbf{+0.168}$ [+0.005, +0.329]	52.0	$\mathbf{-0.416}$ [−0.529, −0.300]	29.0
	ZH	$\mathbf{+0.264}$ [+0.114, +0.421]	55.5	$\mathbf{-0.364}$ [−0.467, −0.261]	32.5

Table 4: **Experiment B: initial creation versus candidates.** Original CREATE outputs are compared with the mean and highest score of four candidates per headline. Each model–language condition has $n = 200$, except Qwen3.6-Flash ZH ($n = 131$, all eligible cases). Scoring, bootstrap CIs, shading, and win-rate definitions follow Table 3.

		vs. candidate mean		vs. highest-scoring candidate	
Model	**Lang.**	**Mean Δ [95% CI]**	**Win (%)**	**Mean Δ [95% CI]**	**Win (%)**
Kimi-K2.5	EN	$\mathbf{+0.728}$ [+0.612, +0.843]	84.0	$\mathbf{-0.498}$ [−0.625, −0.375]	32.0
	ES	$\mathbf{+0.715}$ [+0.596, +0.833]	80.0	$\mathbf{-0.574}$ [−0.706, −0.439]	29.0
	ZH	$\mathbf{+0.536}$ [+0.340, +0.719]	79.0	$\mathbf{-0.479}$ [−0.672, −0.295]	40.0
Kimi-K2.6	EN	$\mathbf{+0.720}$ [+0.604, +0.837]	79.0	$\mathbf{-0.492}$ [−0.623, −0.359]	29.0
	ES	$\mathbf{+0.852}$ [+0.736, +0.965]	86.5	$\mathbf{-0.419}$ [−0.530, −0.309]	33.5
	ZH	$\mathbf{+0.474}$ [+0.321, +0.618]	76.0	$\mathbf{-0.539}$ [−0.695, −0.393]	33.0
Qwen3.6-27B	EN	$\mathbf{-0.247}$ [−0.360, −0.127]	38.5	$\mathbf{-1.422}$ [−1.542, −1.299]	5.0
	ES	$\mathbf{-0.002}$ [−0.124, +0.119]	51.5	$\mathbf{-1.362}$ [−1.496, −1.225]	8.0
	ZH	$\mathbf{-0.157}$ [−0.297, −0.026]	47.5	$\mathbf{-1.175}$ [−1.320, −1.035]	8.0
Qwen3.6-Flash	EN	$\mathbf{-0.417}$ [−0.528, −0.308]	29.0	$\mathbf{-1.516}$ [−1.649, −1.388]	4.0
	ES	$\mathbf{-0.319}$ [−0.448, −0.192]	33.5	$\mathbf{-1.566}$ [−1.702, −1.431]	6.0
	ZH	$\mathbf{-0.448}$ [−0.611, −0.281]	35.1	$\mathbf{-1.501}$ [−1.682, −1.315]	9.2

C PROMPT TEMPLATES

The prompt templates below follow the study's workflow, from constructing JEST to evaluating output quality. Prompt 1 asks each of the four generator models to produce one joke per headline. Prompt 2 elicits the initial SELECT/CREATE decisions; Experiment B evaluates the jokes produced when models initially chose CREATE. In Experiment A, Prompt 3 elicits forced generations for sampled instances on which models initially chose SELECT. Prompt 4 is used to score candidate jokes and newly generated jokes in both experiments. The English, Spanish, and Chinese versions convey equivalent task instructions and specify their respective output languages. Bracketed text denotes instance-specific inputs; each candidate list contains four numbered jokes.

Prompt 1: Candidate Joke Generation

English

Create a concise, creative, and genuinely funny joke based on the news headline. Return only the joke.

News headline: [headline]

Output language: English.

Spanish

Crea un chiste breve, creativo y realmente gracioso a partir del titular de la noticia. Devuelve solo el chiste.

Titular de la noticia: [titular]

Idioma de salida: español.

Chinese

根据新闻标题创作一个简洁、有创意且真正好笑的笑话。只输出笑话。

新闻标题：[标题]

输出语言：中文。

Prompt 2: Select-or-Create Decision

English

Evaluate the candidate jokes. If any candidate is sufficiently good— funny, creative, and relevant to the news—select the best one. If none is good enough, create a better joke. Briefly explain your decision. Return only JSON.

News headline: [headline]

Candidate jokes: [numbered candidate list]

Output language: English.

Spanish

Evalúa los chistes propuestos. Si alguno es lo bastante bueno —es decir, gracioso, creativo y relacionado con la noticia—, elige el mejor. Si ninguno lo es, crea uno mejor. Explica brevemente tu decisión. Devuelve solo JSON.

Titular de la noticia: [titular]

Chistes propuestos: [lista numerada de chistes]

Idioma de salida: español.

Chinese

评估候选笑话。如果存在足够好的候选，即有趣、有创意且与新闻相关，选择其中最优者；如果所有候选都不够好，创作一个更好的笑话。简要说明你的决策理由。只输出JSON。

新闻标题：[标题]

候选笑话：[带编号的候选列表]

输出语言：中文。

Prompt 3: Forced Generation

English

Evaluate the candidate jokes. A sufficiently good candidate is funny, creative, and relevant to the news. For this task, create a better joke based on the news headline whether or not any candidate is good enough. Do not select a candidate. Briefly explain how your new joke differs from the candidates. Return only JSON with the fields `joke` and `reason`.

News headline: [headline]

Candidate jokes: [numbered candidate list]

Output language: English.

Spanish

Evalúa los chistes propuestos. Para ser lo bastante bueno, un chiste debe ser gracioso, creativo y estar relacionado con la noticia. En esta tarea, crea uno mejor a partir del titular, haya o no algún chiste lo bastante bueno. No elijas ninguno de los propuestos. Explica brevemente en qué se diferencia tu nuevo chiste de los demás. Devuelve solo JSON con los campos `joke` y `reason`.

Titular de la noticia: [titular]

Chistes propuestos: [lista numerada de chistes]

Idioma de salida: español.

Chinese

评估候选笑话。足够好的候选应当有趣、有创意，并且贴合新闻。无论候选中有没有足够好的笑话，本次都要根据新闻标题另写一个更好的笑话，不能选择现有候选。简要说明新笑话与候选笑话有何不同。只输出包含`joke`和`reason`字段的JSON。

新闻标题：[标题]

候选笑话：[带编号的候选列表]

输出语言：中文。

Prompt 4: Quality Evaluation

English

Be an extremely strict joke quality critic. Independently rate four dimensions from 0 to 10: news relevance, humor, creativity, and naturalness and fluency as a combined dimension; decimals are allowed. Use the full scale when warranted and follow these anchors strictly (scores 0/1/3/5/7/9, respectively):

News relevance: unrelated / keywords only / weak and clumsy / surface-level / good use of news / tightly interwoven.

Humor: no humor / failed attempt / barely a joke / mild smile / fairly amusing / very funny.

Creativity: pure copy / obvious template / cliché variation / somewhat original / interesting approach / highly novel.

Naturalness and fluency: incoherent / nonsensical / manufactured / acceptable but crafted / fairly natural / extremely smooth.

Do not guess the joke's source. Return only JSON with numeric scores under the fields `news_relevance`, `humor`, `creativity`, and `naturalness` (the combined score for naturalness and fluency).

News headline: [headline]

Joke: [joke]

Output language: English.

Spanish

Actúa como un evaluador de humor extremadamente exigente. Puntúa por separado cuatro dimensiones en una escala de 0 a 10: la relación con la noticia, el humor, la creatividad y, como una sola dimensión, la naturalidad y la fluidez; puedes usar decimales. Usa toda la escala cuando corresponda y respeta estrictamente estos criterios (puntuaciones 0/1/3/5/7/9, en ese orden):

Relación con la noticia: sin relación / solo palabras clave / relación débil y forzada / superficial / buen uso de la noticia / estrechamente entrelazado con la noticia.

Humor: sin humor / intento fallido de hacer reír / apenas un chiste / sonrisa leve / bastante divertido / muy gracioso.

Creatividad: copia pura / plantilla evidente / variación de un cliché / algo original / enfoque interesante / muy novedoso.

Naturalidad y fluidez: incoherente / sin sentido / artificioso / aceptable, aunque algo elaborado / bastante natural / extremadamente fluido.

No intentes adivinar el origen del chiste. Devuelve únicamente JSON con puntuaciones numéricas en los campos `news_relevance`, `humor`, `creativity` y `naturalness` (la puntuación conjunta de naturalidad y fluidez).

Titular de la noticia: [titular]

Chiste: [chiste]

Idioma de salida: español.

Chinese

请以极其严格的标准，从新闻相关性、幽默程度、创意程度和自然流畅度四个方面分别为笑话打分。评分范围为0–10 分，可以使用小数。评分时应严格参照以下锚点，必要时使用完整量程（依次对应0/1/3/5/7/9 分）：

新闻相关性：完全无关/ 只提及关键词/ 关联薄弱且生硬/ 表面相关/ 较好地利用新闻/ 与新闻紧密交织。

幽默程度：毫无幽默感/ 试图搞笑但失败/ 勉强算个笑话/ 略能让人微笑/ 颇为有趣/ 非常好笑。

创意程度：纯粹复制/ 明显套用模板/ 陈词滥调的变体/ 有一定新意/ 有趣的构思角度/ 高度新颖。

自然流畅度：不连贯/ 语义不通/ 刻意造作/ 尚可但有雕琢痕迹/ 较为自然/ 极其流畅。

不要猜测笑话来源。只返回包含`news_relevance`、`humor`、`creativity` 和`naturalness`（自然流畅度）字段及数值评分的JSON。

新闻标题：[标题]

笑话：[笑话内容]

输出语言：中文。