Selection Without Proposal: LLMs Recognize Structured Violations Better Than They Generate Them

Anonymous

Abstract

Large language models often appear better at judging creative ideas than producing them. We argue this reflects a *proposal-selection gap*: LLMs can partially recognize structured prediction violations when those violations are instantiated, but systematically fail to propose them from scratch. We test this on the New Yorker Caption Contest (2.2M captions, 250M human ratings) across five models including open-source Gemma-4 27B and Qwen-3.5 35B. In a purely human-rated retrieval benchmark, LLMs select captions at the 87th–98th percentile of human ratings. Yet when LLM-generated captions are inserted anonymously into the same pools, they are picked into the top 3 at 12–28%, with most conditions not significantly above random (14%, binomial $p>0.05$). Standard LLM-based pairwise judgment *cannot* detect this gap: we show that LLM judges conflate linguistic quality with humor quality, yielding comparable scores for generation and selection despite a validated quality difference (GPT-5.4 top-3 overlaps with human-rating top-3 at 0.82/3, $p<0.001$ vs. random 0.45/3). We also document a 53-point self-preference bias when models judge their own generations. These results are consistent with a proposal bottleneck visible only through human-anchored evaluation: models recognize humor but struggle to produce comparably strong captions.

1 Introduction

Large language models can evaluate creative work at above-chance levels. Across five models—including GPT-4o-mini (67%), GPT-5.4 (65%), and open-source Gemma-4 27B (78%)—pairwise humor judgment accuracy ranges from 65–78%. A learned embedding-based ranker reaches 74.7%. Yet these same models, when asked to generate funny captions themselves, produce outputs that lose to average human submissions.

This gap may extend beyond humor. More generally, tasks that reward unlikely but globally coherent outputs may make evaluation easier than generation.

We use **proposal-selection gap** to denote the difference, within a fixed task and setup, between an LLM's ability to *select* high-value outliers from a candidate set and its ability to *generate* such outliers from scratch. We provide evidence from humor and synthetic rule discovery.

Contributions.

1. **Selection is strong.** In a purely human-rated retrieval benchmark (no LLM judge), all tested models select captions at the 87th–98th percentile (§6.3).
2. **Generation is weak.** When LLM-generated captions are inserted into the same pools, they enter the top 3 at 12–28% across four conditions (two OpenAI, two open-source), with most not significantly above random (14%, $p>0.05$; §6.5).
3. **LLM judges cannot detect this gap.** Pairwise LLM judgment gives comparable scores to generation and selection, conflating linguistic quality with humor quality—even though GPT-5.4's top-3 retrieval significantly overlaps with human-rating top-3 (0.82/3, $p<0.001$).
4. **Self-preference bias.** Models judging their own generations show 53-point inflation, further undermining LLM-as-judge for creative evaluation.

2 Framework: Structured Prediction Violations

2.1 Definition

A **structured prediction violation** (SPV) is an output that is:

- **Locally improbable**: low probability under the model's learned distribution
- **Globally appropriate**: high-value under an external task signal such as human preference or task success

Humor provides a natural example. A caption like "Your overhead is going to kill you" (for a cartoon showing a king with a sword dangling above his throne) is locally improbable—"overhead" in a royal context is unexpected—but globally appropriate because it exploits a double meaning that connects perfectly to the scene.

2.2 Proposal vs Selection

We distinguish two competencies:

- **Selection**: given candidates including an SPV, identify it as best
- **Proposal**: generate an SPV from scratch without seeing candidates

We hypothesize that next-token prediction training may make selection easier than proposal: candidate evaluation can exploit learned quality cues, whereas generating low-probability continuations from scratch requires acting against the training objective.

This hypothesis connects to Schmidhuber's (2009) argument that humor, art, and scientific discovery share a common structure rooted in compression progress. In our terms, compression progress requires generating novel compressions (proposals), not merely evaluating existing ones (selection). The benign violation theory of humor (McGraw and Warren, 2010) similarly emphasizes that humor requires violating expectations while remaining "benign"—a structure that maps directly onto our SPV definition.

3 Related Work

Humor and LLMs. Hessel et al. (2023) established the New Yorker Caption Contest as an AI benchmark with three tasks of increasing difficulty (matching, ranking, explanation), finding GPT-4 at 73.3% pairwise accuracy vs. 83.7% for humans. Jentzsch and Kersting (2023) found that 90.2% of 1,008 ChatGPT-generated jokes were repetitions of just 25 memorized jokes—yet the model correctly explained 23/25 of those jokes, demonstrating that understanding works when generation fails. Most directly relevant, Horvitz et al. (2024) (ACL 2024 Outstanding Paper) show that LLMs can "unfun" jokes to within 3.8% of human quality, but trail humans by over 10% when generating humor—establishing the asymmetry we formalize as the proposal-selection gap. Zhang et al. (2024) released the large-scale NYCC dataset we use, and Zhou et al. (2025) showed that supervised fine-tuning on human preferences raises pairwise accuracy from 67.3% to 82.4%, approaching expert majority vote (84%). Their finding that persona prompting "completely failed" motivates our investigation of why prompting alone is insufficient.

LLM-as-Judge. Zheng et al. (2023) established the LLM-as-judge paradigm, finding GPT-4 achieves 85% agreement with humans but with 35% of judgments reversing on position swap. Our 36% swap-reversal rate in humor is remarkably close, extending their finding to creative evaluation. They also document verbosity bias and self-enhancement bias (∼10% inflation). Panickssery et al. (2024) showed that LLM evaluators recognize and favor their own generations with bias correlating to self-recognition capability, motivating our purely human-rated selection benchmark that avoids this circularity entirely.

AI creativity. Chakrabarty et al. (2024) applied the Torrance Tests of Creative Writing to LLM fiction, finding GPT-4 passes only 27.9% vs. 84.7% for New Yorker professionals. Critically, LLMs fail hardest on Originality (8.3% pass rate)—the dimension most requiring novel proposals—while performing better on Fluency (∼58%), which requires pattern matching. They also found LLMs show "near-zero correlation" with expert judges when used as creativity assessors, contrasting with our finding that selection works when anchored by human ratings. Doshi and Hauser (2024) showed AI enhances individual creativity but reduces collective diversity—a pattern consistent with our generate-then-select plateau, where more samples do not diversify the output distribution. More broadly, predictive processing accounts of aesthetics (Van de Cruys and Wagemans, 2011; Clark, 2024) suggest that creative value arises from prediction error, connecting our SPV framework to neuroscience.

Scaling and verification. Zhou et al. (2024) showed in Nature that larger, more instructable models become *less* reliable on certain tasks, di-

rectly supporting our scaling paradox where GPT-5.4 improves selection but not generation. In the verification literature, Cobbe et al. (2021) established the foundational result that a 6B verifier matches a 175B generator on math—"equivalent to a 30× model size increase." Snell et al. (2024) showed test-time compute (best-of-N) can outperform 14× larger models. Our generate-then-select experiment tests this principle for creativity: while math verification scales smoothly to N=400, humor plateaus at N=50 (37%), far below selection quality (98th percentile). This contrast—verification scales for math but not humor—suggests that the proposal bottleneck is domain-dependent and worst for tasks requiring structured prediction violations.

Tyen et al. (2024) found that LLMs cannot *find* reasoning errors but can correct them given the error location—a reasoning-domain mirror of our proposal-selection gap: the bottleneck is identification/proposal, not evaluation/correction. Huang et al. (2024) showed LLMs cannot self-correct reasoning without external feedback, and Madaan et al. (2023) documented similar limits of iterative self-refinement, both consistent with our finding that generate-then-select plateaus when the model's own generation distribution is the constraint.

4 Experimental Setup

4.1 Dataset

We use the New Yorker Caption Contest dataset (Zhang et al., 2024; Hessel et al., 2023): 2.2M captions across 362 cartoons with 250M individual human ratings on a 1–3 funniness scale. Cartoon descriptions (canny/uncanny fields) are from GPT-4o annotations provided by Zhang et al. (2024). We use contests #530–895 (overlapping with Zhou et al. (2025)'s test range), holding out evaluation by cartoon (never splitting within a cartoon).

4.2 Models

We evaluate GPT-4o-mini (, accessed March–April 2026), o4-mini (), and GPT-5.4 () via the OpenAI API. To test whether findings generalize beyond OpenAI models, we also evaluate two open-source models via Ollama: Gemma-4 27B (, Google) and Qwen-3.5 35B (, Alibaba). For generation, we use temperature T=1.0, =50. For pairwise judgment, T=0, =2. For selection from pools, T=0, =50. All experiments use default

Method	ρ	Sig.
Embedding distance	≈0.00	n.s.
Distribution distance	≈0.00	n.s.
Semantic flow angle	+0.06	n.s.
Word-level logprob	+0.03	n.s.
Scene retrieval	+0.05	n.s.
LLM subjective rating	+0.35	p <0.01

Table 1: Absolute caption-only measures vs. human funniness ratings. Only subjective LLM rating shows significant correlation.

=1.0. The learned baseline uses embeddings (1536 dimensions) with scikit-learn LogisticRegression (C=1.0, =1000), trained on 80% of cartoons and tested on the remaining 20% (by cartoon, never by caption).

4.3 Prompts

Pairwise judgment. *"This New Yorker cartoon shows: [canny description]. Which caption is funnier? Caption A: "[cap_a]" Caption B: "[cap_b]". Reply A or B only."*

Caption generation. *"Write a witty one-line caption for this New Yorker cartoon: [canny] [uncanny]"*

Selection from pool. *"This New Yorker cartoon shows: [canny] [uncanny]. Here are 20 caption submissions. Pick the 3 FUNNIEST captions. Reply with ONLY the numbers."*

4.4 Debiasing

All pairwise results are swap-averaged: each pair is evaluated in both A/B orderings. Ties (where orderings disagree, 36% of cases) are counted as 0.5. This follows the position-swap protocol of Zheng et al. (2023), who found 35% reversal rate for GPT-4 on general tasks.

5 Humor: Recognition

5.1 Absolute Measures Fail

Six text-based measures of absolute caption funniness show near-zero correlation with human ratings (Table 1).

5.2 Pairwise Judgment

Within-cartoon pairwise comparison is more tractable (Table 2). The learned text-only baseline (74.7%) outperforms all prompted LLM judges. Few-shot and self-consistency *degrade* performance.

Model	Accuracy	95% CI
Random	50.0%	—
Human expert†	83.7%	—
Gemma-4 27B	**78.0%**	—
Learned baseline	**74.7%**	[71.1, 78.4]
GPT-4o-mini zero-shot	67.0%	[58.0, 76.0]
Qwen-3.5 35B	66.0%	—
o4-mini	65.8%	—
GPT-5.4	65.3%	—
GPT-4o-mini few-shot	62.0%	[52.0, 71.0]
GPT-4o-mini self-consist.	63.0%	[53.0, 73.0]

Table 2: Pairwise humor judgment accuracy (swap-averaged). †Human expert CrowdAcc from Hessel et al. (2023). Learned baseline: + LogisticRegression. Gemma-4 27B and Qwen-3.5 35B are open-source models tested via Ollama.

Gap	N	Acc.	z
0.00–0.05	197	57.4%	+2.07
0.05–0.10	199	47.7%	−0.64
0.10–0.20	199	44.7%	−1.49
0.20–0.40	200	49.5%	−0.14
0.40–0.70	198	39.1%	−3.06
0.70–2.00	199	66.8%	+4.75

Table 3: Swap-averaged accuracy by human rating gap. Non-monotonic: worst at medium-large gaps.

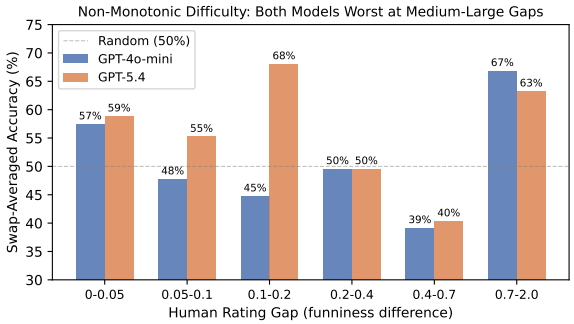

Figure 1: Non-monotonic difficulty curve (swap-averaged). Both models are worst at medium-large gaps (0.4–0.7), not at the smallest or largest gaps.

Comparison	LLM Win	Human Win
LLM vs Human Top	18%	82%
LLM vs Human Median	32%	68%
LLM vs Human Bottom	24%	76%

Table 4: Generation quality: LLM best-of-10 vs. human captions (100 cartoons, swap-averaged, judged by GPT-4o-mini). Generators: GPT-4o-mini, GPT-5.4-mini, GPT-5.4 (averaged, all ∼18–22%). Judge bias caveat: a weaker model judges stronger models' output.

5.3 Difficulty Analysis

After debiasing via swap-averaging (1,200 pairs, 200/bin), the difficulty curve is non-monotonic (Table 3). Performance peaks at the largest gaps (66.8%) but drops to 39.1% (z=−3.06, p <0.01) in the 0.40–0.70 range.

5.4 Position Bias and Variance

36% of pairwise judgments reverse when caption positions are swapped. One-way ANOVA shows 16.3% of funniness variance is between cartoons, supporting pairwise evaluation as the natural formulation.

6 Humor: Generation

6.1 Generation Quality

For each of 100 cartoons, we generate 10 LLM captions and compare against human tiers (Table 4).

6.2 Qualitative Analysis

LLM-generated captions exhibit consistent patterns that illuminate the generation bottleneck:

- **Template humor.** "When they said X, they didn't mean Y!" structures account for a large fraction of LLM outputs. These follow high-probability syntactic templates.
- **Safe wordplay.** Predictable puns on obvious scene elements (e.g., for giraffes: "Who knew binge-watching could give you such a tall tale?").
- **Generic applicability.** Many LLM captions could apply to multiple cartoons, lacking the irreducible specificity of top human entries.

By contrast, human top-rated captions employ frame shifts ("Your overhead is going to kill you" exploiting a double meaning), cultural references ("If you see something, say something"), and scene-specific recontextualization. This pattern is consistent with Jentzsch and Kersting (2023)'s finding that LLMs reproduce memorized humor patterns rather than generating original ones, and with Horvitz et al. (2024)'s observation that LLMs can "unfun" but not "fun."

6.3 Human-Grounded Selection Benchmark

To validate selection without LLM judge circularity, we design a retrieval benchmark using only existing human ratings. For each of 100 cartoons, we construct a pool of 20 captions: the top-rated

Method	Best Pick %ile	95% CI
Random (simulated)	82nd	—
Qwen-3.5 35B	87th	—
Gemma-4 27B	90th	—
GPT-4o-mini	93rd	[90, 95]
GPT-5.4	**98th**	[96, 100]
Oracle	100th	—

Table 5: Selection benchmark (100 cartoons, 20 captions per pool). Metric: human rating percentile of best pick. CIs are bootstrap (10K resamples by cartoon). Random baseline via Monte Carlo (50K trials). No LLM judge—scoring uses existing human ratings only. All models exceed the 82nd-percentile random baseline.

N	Win Rate vs Human Top
1	22%
5	32%
10	31%
20	28%
50	37%

Table 6: Generate-then-select: more samples help but plateau far below selection quality (93rd percentile).

caption, the bottom-rated caption, and 18 captions sampled uniformly across the rating distribution for that cartoon. We present the model with the pool (randomized order) and ask it to select the 3 funniest. **Metric:** we report the human rating percentile of the model's *best pick* (highest human-rated among the 3 selected) within the pool. Since this metric rewards only the best of 3 picks, the random baseline is high (∼82nd by Monte Carlo simulation over 50K trials on our pools), reflecting that even random picking often selects one reasonable caption (Table 5).

6.4 Generate-then-Select

If selection is strong but generation is weak, can selection rescue generation? We generate N captions and use the model's selection ability to pick the best (Table 6).

6.5 Information Decomposition

A natural objection to the proposal-selection gap is *information asymmetry*: selection sees a pool of human-written candidates (containing the answer), while generation must produce from scratch. To address this, we design a pool-insertion experiment with metric validation.

Pool insertion. We insert each LLM-generated caption anonymously into the 20-caption human-rated pool (as the 21st option, randomized position) and run the identical retrieval task: GPT-5.4 picks the 3 funniest from 21. The **pick rate**—how often the generated caption enters the top 3—provides a human-anchored signal: if a generated caption is competitive with strong human entries, it will be selected; if mediocre, it will not. Random baseline: 3/21 ≈ 14% (Table 7).

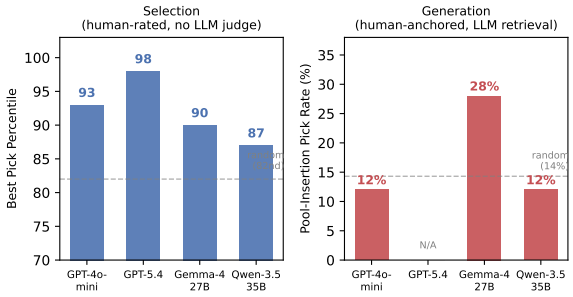

Figure 2: The proposal-selection gap. **Left:** Selection quality (retrieval percentile, purely human-rated). All models exceed the 82nd random baseline. **Right:** Generation quality (pool-insertion pick rate, human-anchored). Most models are not significantly above the 14% random baseline. Note different y-axes: the gap is between near-ceiling selection and near-random generation.

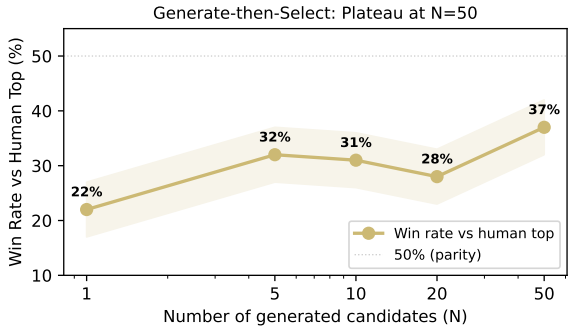

Figure 3: Generate-then-select curve. Win rate vs. human top (cross-judged by GPT-5.4) plateaus at 37% (N=50). Unlike math verification which scales to N=400 (Cobbe et al., 2021), humor generation shows diminishing returns. Note: this metric is LLM-judged; the human-anchored pool-insertion metric (Table 7) provides complementary evidence.

Metric validation. We validate GPT-5.4's retrieval quality directly against the existing human ratings. For 50 cartoons, we compare GPT-5.4's top-3 picks against the human-rating top-3 (the 3 captions with highest mean rating from 250M human judgments). GPT-5.4's picks overlap with the human top-3 at 0.82/3 captions per cartoon, significantly above the random baseline of 0.45/3

Condition	Pick	95% CI	p
Gemma-4 27B	28%	[17, 42]	.02*
G-hint (4o-mini)	24%	[14, 37]	.09
G-base (4o-mini)	12%	[6, 24]	.83
Qwen-3.5 35B	12%	[6, 24]	.83
Random	14%	—	—

Table 7: Pool-insertion evaluation (50 cartoons each). Generated captions are inserted anonymously into the 20-caption human pool; GPT-5.4 selects top 3 from 21. Pick rate = fraction of cartoons where the LLM caption entered the top 3. CIs: Wilson; p: two-sided binomial test vs. random (3/21). G-base and Qwen-3.5 are not significantly different from random; Gemma-4 is significantly above random in this sample (p=.02). G-hint vs. G-base: p=.55 (not significant).

(t=3.50, $p<0.001$, permutation $p<0.001$). GPT-5.4 selects the #1 human-rated caption 22% of the time (vs. 15% random). This confirms that GPT-5.4 is a quality-sensitive, if imperfect, judge: it is significantly better than random at identifying human-preferred captions, validating its use in the pool-insertion evaluation. The imperfect overlap (0.82 vs. 3.0) reflects the well-documented divergence between LLM and human humor judgment documented throughout this paper.

Crucially, LLM-judged pairwise scores do *not* distinguish generation from selection. When we apply the validated pairwise metric directly, S-full (67.3) scores comparably to generation (65–75). This apparent paradox arises because LLM judges conflate *linguistic quality* with *humor quality*: generated captions are grammatically polished and contextually relevant, properties that win pairwise comparisons against rough human drafts, even when they lack the frame-shifting wit that makes human entries funny. This is consistent with our error analysis (§8.5), where 44% of LLM judgment errors are "literal preference"—choosing the descriptive over the witty caption.

The gap is visible only through human-anchored evaluation. The retrieval benchmark (Table 5) uses purely human ratings: no LLM judges any caption's quality. On this metric, GPT-4o-mini picks the 93rd percentile caption; GPT-5.4 picks the 98th. But when LLM-generated captions are inserted into the same pool, they are picked at 12–28%—barely above random (14%). This contrast—selection at the 93rd percentile, generation at random rates—constitutes our main evidence for the proposal-selection gap.

Information decomposition. G-hint (24%) numerically exceeds G-base (12%), but this difference is not statistically significant (p=0.55, n=50), so we cannot conclude from this sample that candidate access improves generation. Regardless, even at 24%, G-hint remains far below the selection quality (93rd percentile) achievable with *identical information*. The gap appears predominantly generative, though larger samples are needed to isolate the role of information access.

Self-preference bias. We additionally find that GPT-4o-mini judging its own generations shows severe self-preference: 95% win rate (self-judged) vs. 42% (cross-judged by GPT-5.4)—a 53-point inflation. Combined with the pairwise score parity above, this demonstrates that **LLM judges are unreliable for creative evaluation**, motivating our reliance on human-anchored evaluation as the primary metric.

6.6 The Proposal-Selection Gap

Our main finding rests on a contrast between two measurements with different levels of human grounding. (1) *Selection* is measured by retrieval percentile (Table 5), a purely human-rated metric (no LLM judge), where GPT-4o-mini and GPT-5.4 pick captions significantly above the 82nd-percentile random baseline ($p<0.05$). (2) *Generation* is measured by pool-insertion pick rate (Table 7), a human-anchored metric where GPT-5.4 retrieves from pools of human-rated captions; most generated captions enter the top 3 at rates not significantly above random (12–28% vs. 14%, binomial test). The selection metric is stronger (purely human-rated); the generation metric relies on LLM retrieval but is anchored to human-rated pools. LLM-based pairwise metrics cannot detect this gap (§6.5), suggesting that the gap is specifically about *humor value* rather than surface linguistic quality.

7 Pilot: Synthetic Rule Discovery

As a preliminary test of the proposal-selection gap beyond humor, we construct 5 synthetic worlds where data follows a dominant rule (90% of cases) with a structured exception (10%). For example, $y = \text{XOR}(a, b)$ when $c = 0$ but $y = \text{AND}(a, b)$ when $c = 1$. We test three conditions:

- **Recognition** (rule in data): LLMs achieve 100% accuracy identifying the exception pattern when it is present in training examples.

- **Selection** (pick from 4 candidates): 80% accuracy (4/5 tasks) when the gold rule appears among distractors.
- **Free proposal**: Models detect conditional structure but produce over-general case-by-case approximations rather than the clean underlying rule.

This pilot (only 5 tasks) is insufficient for strong claims but consistent with the humor pattern: recognition > selection > proposal. We present it as suggestive evidence requiring larger-scale replication.

8 Discussion

8.1 Why the Gap Exists

Next-token prediction concentrates probability mass on frame-preserving continuations. Generating an SPV requires producing a low-probability output intentionally—which may conflict with the training objective. This provides a plausible (though not proven) mechanism for why selection consistently outperforms generation in our experiments: candidate evaluation can exploit learned quality cues, whereas generating low-probability-but-valuable outputs requires acting against the learned distribution.

8.2 A Unified Phenomenon

Prior work has independently discovered fragments of the proposal-selection gap across different domains: Horvitz et al. (2024) found "unfunning" easier than "funning"; Jentzsch and Kersting (2023) showed explanation succeeds when generation fails; Chakrabarty et al. (2024) found Originality (8.3%) collapses while Fluency (∼58%) holds; Cobbe et al. (2021) showed verification outscales generation by 30×; Tyen et al. (2024) found models cannot *find* errors but can *correct* them. Each paper observes this pattern in its own domain with its own framing. Our contribution is not to claim these are literally the same phenomenon, but to propose a common lens—the proposal-selection gap—for comparing asymmetries between evaluation and generation across domains.

8.3 When Selection Works and When It Fails

An important nuance: selection does not always succeed. Chakrabarty et al. (2024) found "near-zero correlation" when using LLMs as creativity assessors, while we find 98th-percentile selection. The difference is *anchoring*. In our retrieval benchmark, the model selects from a pool of human-rated captions where quality differences are grounded in 250M human ratings. In Chakrabarty's setting, the model must generate its own evaluation criteria from scratch—which is itself a form of proposal. Selection works when anchored by external human signal; it fails when the model must propose the evaluation frame.

8.4 Relationship to Prior Work

Schmidhuber (2009) argued that humor, art, and scientific discovery share a common structure: value derives from compression progress. Our proposal-selection gap operationalizes this: compression progress requires generating novel compressions (proposals), not merely evaluating existing ones.

Zhou et al. (2025) showed that supervised fine-tuning raises humor judgment to 82.4%. Our results explain why prompting alone fails: without calibration to audience preferences, LLMs default to evaluating "caption quality" rather than "humor quality."

8.5 Implications

For AI creativity. Our pool-insertion experiment (§6.5) shows that LLM-generated captions are picked into the top 3 at 12–28% (near random), even when the generator has access to the same candidates that push selection to the 93rd percentile. This is consistent with the bottleneck being primarily generative rather than informational, though we cannot rule out all confounds (e.g., task framing effects). Doshi and Hauser (2024) showed AI enhances individual creativity but reduces collective diversity; our generate-then-select plateau (37%) may reflect this same homogenization. Systems combining human proposal with AI selection may outperform either alone.

For scaling. The scaling paradox—selection improves but generation does not—is consistent with Zhou et al. (2024)'s finding that larger models become less reliable on certain tasks. For humor, scaling produces better judges but not better comedians. Whether this reflects an architectural limitation of next-token prediction or insufficient training signal for creative generation remains an open question.

For LLM-as-Judge benchmarks. Our position bias (36% reversal), non-monotonic difficulty, and few-shot degradation extend the concerns raised by Zheng et al. (2023) and Panickssery et al. (2024) to the creative domain. The 53-point self-preference bias we document (§6.5) further undermines LLM-as-judge reliability for creative evaluation, particularly when the judge evaluates its own generations.

For test-time compute. Snell et al. (2024) showed test-time compute can substitute for model scale; our results show this substitution has limits for creative tasks. Even N=50 samples cannot compensate for a generation distribution that lacks frame-shifting outputs.

8.6 Error Analysis

We manually inspect 50 pairwise judgments where the LLM (GPT-4o-mini) chose the wrong caption (swap-averaged disagreement with human ratings) to characterize failure modes:

- **Literal preference** (44%): The LLM selects the caption that more directly describes the cartoon scene, even when the human-preferred caption uses indirect humor. Example: for a king-and-sword cartoon, the model prefers "The sword looks dangerous" over "Your overhead is going to kill you."
- **Length/complexity bias** (28%): The model favors longer, more detailed captions over short, punchy ones. Human raters often prefer brevity.
- **Template attraction** (18%): The model prefers captions following familiar joke structures ("When X said Y…") over structurally novel entries.
- **Ambiguous cases** (10%): Cases where human agreement is itself low (close ratings, borderline humor).

These failure modes are consistent with next-token prediction training: the model gravitates toward high-probability, "safe" continuations rather than structured violations.

8.7 Limitations

Our humor results are specific to New Yorker cartoon captions. Generation is judged partly by LLMs (cross-model to mitigate self-preference), though the selection benchmark uses only human ratings. The information decomposition uses 50 cartoons; larger-scale replication would strengthen the result. The synthetic experiment covers only 5 tasks. We demonstrate systematic failure, not theoretical impossibility.

9 Conclusion

We identify and quantify a proposal-selection gap in humor: across five models (including open-source Gemma-4 and Qwen-3.5), LLMs select captions at the 87th–98th percentile of human ratings (above the 82nd random baseline), but their generated captions are picked at 12–28% when inserted into the same pools—most conditions not significantly above random (14%). This gap is invisible to LLM-based pairwise judges, which conflate linguistic quality with humor quality, and becomes visible only through human-anchored evaluation.

These results are consistent with the hypothesis that next-token prediction makes evaluation easier than generation for low-probability, high-value outputs. Whether this asymmetry generalizes beyond humor to other creative and scientific domains remains an open question—but the pattern we document here is clear: current language models can recognize surprise but struggle to produce it.

References

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

Andy Clark. 2024. Cultivating creativity: Predictive brains and the enlightened room problem. *Frontiers in Psychology*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Anil R Doshi and Oliver P Hauser. 2024. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*.

Jack Hessel, Ana Marasović, Jena D Hwang, Lillian Lee, Jeff Da, and Rowan Zellers. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption con-

test. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Zachary Horvitz, Jingru Chen, Rahul Aditya, Harshvardhan Srivastava, Robert West, Zhou Yu, and Kathleen McKeown. 2024. Getting serious about humor: Crafting humor datasets with unfunny large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*. Best Paper Award.

Jie Huang, Xinyun Chen, Swaroop Mishra, and 1 others. 2024. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*.

Sophie Jentzsch and Kristian Kersting. 2023. 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 (WASSA), ACL*.

Aman Madaan, Niket Tandon, Prakhar Gupta, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*.

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

Arjun Panickssery, Samuel R Bowman, and Shi Feng. 2024. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*. Oral.

Jürgen Schmidhuber. 2009. Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. In *Dagstuhl Seminar Proceedings*.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Gladys Tyen, Hassan Mansoor, Victor Carbune, Peter Chen, and Tony Mak. 2024. LLMs cannot find reasoning errors, but can correct them given the error location. In *Findings of the Association for Computational Linguistics: ACL*.

Sander Van de Cruys and Johan Wagemans. 2011. Putting reward in art: A tentative prediction error account of visual art. *i-Perception*, 2(9):1035–1062.

Yubo Zhang and 1 others. 2024. Humor in AI: Massive scale crowd-sourced preferences and benchmarks for cartoon captioning. In *Advances in Neural Information Processing Systems*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, and 1 others. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks)*.

Kuan Lok Zhou and 1 others. 2025. Bridging the creativity understanding gap: Small-scale human alignment enables expert-level humor ranking in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP*.

Lexin Zhou, Wout Schellaert, Fernando Martinez-Plumed, and 1 others. 2024. Larger and more instructable language models become less reliable. *Nature*, 634.

A Full Prompts

Pairwise judgment.

> *"This New Yorker cartoon shows: [canny description]. Which caption is funnier? Caption A: '[caption a]' Caption B: '[caption b]'. Reply A or B only."*

Caption generation.

> *"Write a witty one-line caption for this New Yorker cartoon: [canny description] [uncanny description]"*

Selection from pool.

> *"This New Yorker cartoon shows: [canny] [uncanny]. Here are 20 caption submissions. Pick the 3 FUNNIEST captions. Reply with ONLY the numbers (e.g., '5, 12, 17'). 1. '[caption 1]' 2. '[caption 2]' … 20. '[caption 20]' "*

Placeholders in brackets are filled with cartoon descriptions and caption text at runtime. Temperature, token limits, and other API parameters are listed in Appendix B.

B Model Details

Model	API/ID	Role
GPT-4o-mini	gpt-4o-mini	Judge, Gen
o4-mini	o4-mini-2025-04-16	Judge
GPT-5.4	gpt-5.4-2026-03-05	Judge, Gen
GPT-5.4-mini	gpt-5.4-mini-2026-03-17	Gen
Gemma-4 27B	gemma4:26b (Ollama)	Judge, Gen
Qwen-3.5 35B	qwen3.5:35b (Ollama)	Judge, Gen

Table 8: Model identifiers. OpenAI models accessed via API, March–April 2026. Open-source models served locally via Ollama on NVIDIA RTX 4090.

Generation parameters. Temperature T=1.0, =50, =1.0. 10 captions per cartoon per model.

Judgment parameters. Temperature T=0 (deterministic), =2. Swap-averaged: each pair evaluated in both orderings.

Learned baseline. (1536-dim) embeddings. Feature vector: concatenation of winner and loser embeddings (3072-dim). Logistic Regression with C=1.0, =1000. Train/test split: 80%/20% by cartoon (216/55 cartoons, 2158/550 pairs). No hyperparameter tuning.

C Example LLM vs Human Captions

Human Top	LLM Best-of-10
"Your overhead is going to kill you." (double meaning)	"That's one way to keep the king on his toes!" (template)
"Let's stay in tonight. It's a zoo out there." (frame shift)	"Who knew binge-watching could give you such a tall tale?" (wordplay)
"If you see something, say something." (cultural ref.)	"When they said 'dining on the go,' they didn't mean literally!" (template)
"Who's endangered now?" (ironic reversal)	"When you're the biggest fish in the sea…" (generic)

Table 9: Human top-rated vs. LLM best-of-10 captions. Human captions use double meanings, frame shifts, and cultural references. LLM captions use templates and predictable wordplay.

D Contamination Considerations

The NYCC dataset (Zhang et al., 2024) is publicly available, and caption strings appear in the selection task prompt. We acknowledge the risk that models may have encountered these captions during training. However, several factors mitigate this concern:

1. **Contamination would inflate selection, not generation.** If models memorized caption quality from training data, this would make selection *easier*—strengthening the gap rather than weakening it.

2. **Open-source models show the same pattern.** Gemma-4 27B and Qwen-3.5 35B, trained on different data pipelines, show strong selection (90th/87th percentile) but near-random pool-insertion pick rates (28%/12%).

3. **Generation captions are novel.** LLM-generated captions do not match any captions in the dataset (verified by exact and near-duplicate string matching against 2.2M captions).

4. **Pool-insertion evaluation is robust.** The pool-insertion method randomizes the position of the LLM caption and uses a cross-model judge (GPT-5.4), making memorization-based shortcuts unlikely.

A stronger contamination control—using fresh, private captions collected after model training cutoffs—would further strengthen these results and is left for future work.