Does Bigger Mean Funnier? Evaluating Humor Generation Across the Qwen3 Model Family

Jatin Agrawal
LTRC, IIITH
jatin.agrawal@research.iiit.ac.in

Radhika Mamidi
LTRC, IIITH
radhika.mamidi@iiit.ac.in

Abstract

We investigate whether scaling model parameters improves humor generation through a controlled ablation study. Using five Qwen3 variants (8B–235B, dense and MoE), we generate jokes across 50 themes. Beyond evaluating humor scaling, this work serves as an empirical study into the nature of LLM versus human evaluations on highly subjective creative tasks. While an automated judge yields a perfect monotonic ranking between parameter count and win rate, human annotators find no significant aggregate difference in humor quality. Restricting to themes where annotators agree reveals a *significant* preference for the largest model ($p = 0.039$), suggesting scaling effects exist but are masked by a "quality floor"—a baseline level of structural competence below which models rarely fall, making their outputs perceptually indistinguishable. Crucially, our analysis of bias characteristics shows that the automated judge exhibits substantial positional and length biases compared to human evaluators, further suggesting that LLMs may systematically distort quality differences on subjective tasks.

1 Introduction

Scaling laws suggest that increasing model size yields predictable improvements in language modeling (Kaplan et al., 2020). However, whether these trends extend to creative generation tasks remains unclear. Humor requires world knowledge, pragmatic inference, and timing—capabilities that may not automatically emerge with additional parameters. As LLMs are widely deployed in creative applications, understanding their ability to generate genuinely funny content is practically important, especially since state-of-the-art models often struggle with repetition and nuance (Jentzsch and Kersting, 2023; Hessel et al., 2023).

We present a dual-purpose investigation. Primarily, we conduct the first controlled, within-family ablation study of parameter count on humor generation quality. By restricting testing to the Qwen3 family, we isolate parameter count from confounds like training data and architecture, evaluating both dense and Mixture-of-Experts (MoE) models to distinguish between a model's total parameter capacity and the subset of "active" parameters it uses per token. Secondarily, we empirically examine the fundamental differences between automated LLM-as-a-judge evaluation and human annotation on highly subjective tasks.

Our contributions are:

1. A controlled within-family ablation study of humor generation across five Qwen3 models (8B–235B), combining automated pairwise judging with blinded human annotation.
2. A high-agreement subset analysis revealing that scaling effects are significant when annotators can distinguish joke quality ($p = 0.039$), but are masked in the aggregate by a "quality floor."
3. Extensive analysis showing that automated LLM judges exhibit substantial procedural artifacts—notably a 73.5% positional bias and significant length bias ($p < 0.0001$)—potentially overstating model differences that humans do not perceive.
4. An open-source pipeline, dataset, and annotation data for reproducibility, available at https://github.com/J10Official/Humor-Parameter-Ablation.

2 Related Work

Humor Generation and Scaling. ChatGPT repetitively produces the same jokes (Jentzsch and Kersting, 2023), and broad model comparisons often confound architectural benefits with training data advantages (Evstafev, 2025). While recent work focuses on generation via prompting strate-

Proceedings of the 2nd Workshop on Computational Humor (CHum 2026), pages 81–94

gies (Tikhonov and Shtykovskiy, 2024; Kim and Chilton, 2025), the underlying capability scaling of humor remains unexplored. Power-law scaling holds for language modeling loss (Kaplan et al., 2020), but whether creative humor scales smoothly, emerges suddenly (Wei et al., 2022), or plateaus remains open. Our controlled *within-family* ablation isolates the parameter count variable.

Evaluation Disconnects and Subjectivity. LLM judges exhibit vulnerabilities like positional and verbosity biases, struggling on open-ended subjective tasks (Wang et al., 2023). Our work extends these findings to creative generation, where we observe substantial procedural biases. Subjectivity in humor evaluation causes low agreement (Castro et al., 2018; Mittal et al., 2021), but as Sandri et al. (2023) argue, this disagreement is informative rather than merely noisy. We show that aggregate human disagreement conceals structured patterns, dropping sharply when models hit a "quality floor."

3 Methodology

3.1 Models

We evaluate five instruction-tuned models from the Qwen3 family (Qwen Team, 2025): three dense models (8B, 14B, 32B) and two Mixture-of-Experts (MoE) models that route to a subset of active parameters per token: 30A3 (30B total, 3B active) and 235A22 (235B total, 22B active). All models share the same training pipeline and tokenizer, differing in parameter count and architectural approach. The inclusion of both dense and MoE architectures allows us to disentangle the effects of total parameter count from active (per-token) parameter count.

3.2 Data

We curate 50 everyday humor themes (e.g., *"Autocorrect text message fails," "Airport security lines"*). Themes span a broad range of everyday situations—including technology, social interactions, workplace life, and domestic scenarios—while avoiding domain-specific knowledge requirements, culturally sensitive topics, and themes likely to produce offensive content. The full theme list is available in the repository. Each model generates one joke per theme using a standardized prompt (Appendix A.1), with `max_tokens=128` and thinking disabled.[1]

[1]The Qwen3 `/no_think` control token suppresses chain-of-thought reasoning, ensuring purely generative output.

3.3 Automated Evaluation

We perform pairwise comparisons using DeepSeek-V3.2 as a judge via OpenRouter. We deliberately select a non-Qwen judge to avoid potential family bias in evaluation.[2] For each theme, all $\binom{5}{2} = 10$ model pairs are compared in both orderings (A-first and B-first), yielding 20 comparisons per theme. This symmetric design ensures that any positional bias affects all models equally in aggregate. We aggregate results using Bradley-Terry (BT) scoring (Bradley and Terry, 1952), which estimates latent "ability" parameters from pairwise outcomes, and report win rates computed as the fraction of comparisons each model wins.

3.4 Human Evaluation

Two independent groups of six annotators participate in a blinded evaluation of three models (32B, 30A3, 235A22) across 40 themes (themes 11–50). The first 10 themes were reserved for pilot testing, and the two smaller models (8B, 14B) are excluded to keep annotation tractable. Evaluation is conducted simultaneously using a split projected screen, with full details of the annotation protocol provided in Appendix A.2.

4 Results & Discussion

We report automated judge outcomes and artifacts, followed by human ranking results and a subset analysis focusing on themes where annotators can reliably identify a winner.

4.1 Automated Judge Findings

Table 1 presents the LLM judge results across all 50 themes. Win rates increase monotonically with total parameter count, yielding a perfect Spearman correlation ($\rho = 1.0$, $p < 0.001$). The log-linear relationship (Figure 1) shows a clear positive trend, though the three mid-sized models cluster tightly between 49–55%. Furthermore, calculating a theme-level cluster bootstrap standard error (SE) of the Bradley-Terry scores—which accounts for intra-theme performance correlation—reveals that the 95% confidence intervals for these models frequently overlap. This internal variation indicates that while the LLM's aggregate point estimates scale perfectly, the underlying margin of victory

[2]One theme received 19 of 20 planned comparisons due to an API timeout, yielding 999 rather than 1000 total comparisons.

Model	Win Rate	BT Score (± SE)	Avg Rank
235A22	61.0%	0.276 ± 0.030	2.63
32B	54.6%	0.224 ± 0.024	2.84
30A3	51.6%	0.202 ± 0.020	3.18
14B	49.2%	0.187 ± 0.024	2.85
8B	33.5%	0.110 ± 0.014	3.50

Table 1: LLM judge results (50 themes, 999 comparisons). BT = Bradley-Terry score (normalized to sum 1) with cluster-level bootstrap standard error. Win rates and BT scores correlate perfectly with total parameter count ($\rho = 1.0$), though overlapping SE bounds suggest narrow actual margins of victory.

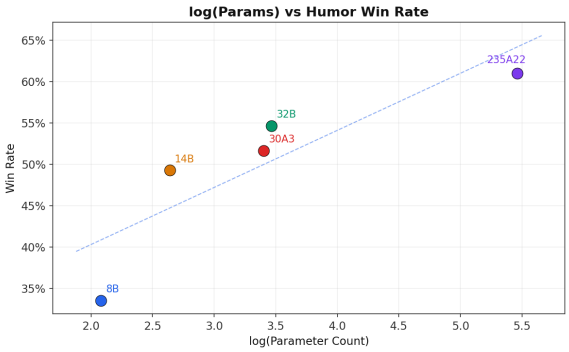

Figure 1: Log parameter count vs. LLM judge win rate. The x-axis represents the log-scaled total parameter count, and the y-axis indicates the fraction of pairwise comparisons won. Error bars denote 95% cluster-bootstrap confidence intervals. The perfect positive correlation ($\rho = 1.0$) indicates that the LLM judge systematically prefers larger models, despite overlapping confidence intervals for mid-sized models.

among consecutive models is narrow, foreshadowing the rating ambiguity observed in human judgments.

The automated judge exhibits severe procedural artifacts, notably positional and length biases. Position A (the joke presented first) wins 73.5% of comparisons ($\chi^2 = 215.9$, $p < 0.001$). While positional bias in LLM judges has been documented in factual evaluation settings (Zheng et al., 2023; Wang et al., 2023), our observed rate of 73.5% is substantially higher than the 60–65% typically reported, suggesting that subjective creative tasks may amplify this bias.

Furthermore, the judge exhibits a strong length bias: when the two jokes differ in character count, it selects the longer joke 62.2% of the time (binomial $p < 0.0001$). This contrasts sharply with our human evaluators, who show no significant preference for longer jokes (Spearman $\rho = -0.065$, $p = 0.082$). The LLM's reliance on length as a proxy for humor highlights a fundamental disconnect between automated metrics and human subjective appreciation.

Model	Mean Rank	Borda	Ranked 1st
235A22	1.900	1.100	40.8%
32B	2.033	0.967	30.4%
30A3	2.067	0.933	28.7%

Table 2: Human evaluation results (12 annotators, 40 themes, 240 total rankings). Mean rank: 1 = best, 3 = worst. Borda score: 0–2 scale. Unlike the automated judge, human evaluators show no statistically significant overall preference for larger models.

To quantify reliability, we find that only 52.7% of transposed model pairs produce consistent winners—not significantly above chance (one-sided binomial test, $p = 0.12$).

4.2 Human Evaluation Findings

Human annotators show a consistent but weaker trend than the LLM judge (Table 2). The largest model (235A22) achieves the best mean rank (1.90) and is ranked first 40.8% of the time, compared to 30.4% for 32B and 28.7% for 30A3. However, a Friedman test reveals no significant difference between models ($\chi^2 = 3.73$, $p = 0.155$). Pairwise Wilcoxon signed-rank tests confirm that no model pair differs significantly (all $p > 0.10$). The rank distribution (Appendix Figure 4) shows that 235A22 receives a disproportionate share of 1st-place rankings (41% vs. 33% expected by chance), but bootstrap 95% confidence intervals for mean rank overlap substantially (Appendix B.2, Figure 5).

Agreement is low (Krippendorff's $\alpha = 0.098$ (Krippendorff, 2011)), consistent with prior humor studies (Castro et al., 2018; Mittal et al., 2021) and reflecting subjectivity rather than annotation error (Sandri et al., 2023). Notably, both groups independently rank 235A22 first (Group 1: 1.88; Group 2: 1.92), suggesting the top-level trend is robust despite low per-item agreement.

Model performance varies substantially across themes (Appendix B.2, Figure 6). This observation motivates our high-agreement subset analysis (§4.3): themes where annotators *can* distinguish between models may reveal scaling effects masked in the aggregate.

4.3 High-Agreement Subset Analysis

We therefore quantify per-theme distinguishability by computing the fraction of annotator pairs that agree on which model produced the funniest joke. With 6 annotators per theme there are $\binom{6}{2} = 15$ pairs; if all 6 agree on the winner, agreement is 1.0, while under random ranking the expected agreement is 0.33. We define high-agreement themes as those with pairwise winner agreement ≥ 0.4 (above chance).[3]

Of the 40 themes, 22 (55%) meet the high-agreement threshold, with agreement scores ranging from 0.40 to 1.00.

Table 3 presents the subset analysis. On the 22 high-agreement themes, the Friedman test is **significant** ($\chi^2 = 6.47$, $p = 0.039$; Kendall's $W = 0.025$, indicating a small effect size), and a pairwise Wilcoxon test shows that 235A22 significantly outperforms 30A3 ($W = 3404$, $p = 0.020$). On the 18 low-agreement themes, the Friedman test shows no effect whatsoever ($\chi^2 = 0.35$, $p = 0.839$), and all three models are statistically indistinguishable (all mean ranks within 0.07 of 2.0).

The subset-restricted figures confirm this pattern visually. Appendix Figure 7 shows that on the high-agreement subset, 235A22's confidence interval separates clearly from 30A3, and Appendix Figure 8 shows a pronounced shift in first-place rankings toward 235A22. The per-theme heatmap restricted to high-agreement themes (Appendix Figure 9) reveals sharper color contrasts between models compared to the all-theme version (Appendix B.2), confirming that model differences are concentrated in this subset.

4.4 Architectural Efficiency: Dense vs. MoE

Comparing the 32B (dense) and 30A3 (MoE) models reveals a sharp divergence in how judges evaluate architectural efficiency. To the LLM judge, the 30A3 (51.6% win rate) performs remarkably close to the 32B dense model (54.6%) despite utilizing just 3B active parameters per token, implying total parameter capacity suffices for structural competence. Conversely, human evaluators show a directional preference for dense compute: in the high-agreement subset, the 32B dense model (mean rank 2.023) outperforms the 30A3 MoE (2.144), though this difference is not statistically significant (Wilcoxon $W = 3916.5$, $p = 0.259$). This suggests that while sparse routing maps adequate world knowledge to formulate a coherent joke, the advantage of dense per-token reasoning for humor remains suggestive rather than conclusive.

Subset	Mean Rank			Friedman p
	235A22	32B	30A3	
High-agree (22)	**1.833**	2.023	2.144	**0.039***
Low-agree (18)	1.981	2.046	1.972	0.839
All (40)	1.900	2.033	2.067	0.155

Table 3: Subset analysis by inter-annotator agreement. On high-agreement themes, the Friedman test reaches significance ($p < 0.05$), and 235A22 significantly outperforms 30A3 (Wilcoxon $p = 0.020$). Low-agreement themes show no model differences, confirming that scaling benefits are only perceptible when evaluating distinguishable topics.

4.5 Qualitative Analysis: Why Scaling Helps

To connect our quantitative findings to humor mechanisms, we examine representative outputs through the lens of the Incongruity-Resolution framework (Suls, 1972), which posits that humor arises when an expectation is established and then subverted by an incongruous but interpretable punchline.

On the highest-agreement theme (*Astrological signs*, agreement = 1.0, 235A22 wins unanimously), the models produce strikingly different outputs:

- **8B:** *"I asked my friend if she believed in horoscopes, and she said, 'I don't believe in them, but I do believe in the power of a good reading to make you feel seen."'* — No incongruity or punchline; the response reads as a sincere comment.
- **32B:** *"I checked my horoscope this morning and it said I'd have a lucky day—turns out the 'lucky' part was the only part they got right."* — Weak incongruity; the subversion ("lucky" being the only accurate part) is vague and lacks a concrete comic image.
- **235A22:** *"I checked my horoscope, and it said today is a great day to reconnect with someone from your past—so I texted my ex. Turned out, the universe meant your* dentist." — Classic incongruity-resolution: the setup establishes a

[3]The result is robust to threshold choice: at 0.35, the same 22 themes qualify ($p = 0.039$); at 0.50, only 6 themes qualify but the rank ordering (235A22 > 32B > 30A3) is preserved ($p = 0.08$, non-significant due to reduced power). The gap between 0.40 and 0.45 is sharp: no themes have agreement in $(0.40, 0.467)$, so 0.40 is a natural boundary.

romantic expectation (texting an ex), then the punchline subverts it with a mundane, dreaded alternative (the dentist). The delayed reveal and italicized final word demonstrate precise comedic timing.

This pattern recurs across high-agreement themes. On *Getting older/Turning 30* (agreement = 0.667, 32B wins), the 32B model produces a dark, unexpected punchline (*"half my friends are dead and the other half are on LinkedIn"*), while 30A3 delivers an extended metaphor that dilutes the comic punch. On *Coffee shop baristas* (agreement = 0.667, 235A22 wins), 235A22's punchline (*"she looked at me like I'd asked for a time machine"*) delivers a sharp, concrete visual incongruity, whereas 30A3's version (*"a side of judgment?"*) relies on a generic quip.

Conversely, on low-agreement themes like *Dealing with landlords* (agreement = 0.2), all three models produce structurally competent but interchangeable jokes—each relies on a familiar complaint-as-metaphor structure without genuine incongruity, explaining why annotators cannot distinguish them.

These examples suggest that scaling benefits manifest specifically as improved *incongruity construction*: larger models more reliably establish an expectation and subvert it with a concrete, surprising punchline, whereas smaller models tend toward generic observations or over-extended metaphors that lack a clear comic turn.

4.6 The Human-LLM Disconnect

The divergence between automated and human evaluation is not merely a matter of degree but of kind. While human evaluators correctly judge structurally competent but equally unfunny jokes as ties, the LLM relies on surface-level proxies (fluency, exact length) that correlate with model size to force a monotonic ranking despite overlapping Bradley-Terry confidence intervals. This is starkly illustrated by the low-agreement themes (§4.3): on themes like *Dealing with landlords*, where all models produce indistinguishable outputs, human annotators appropriately assign near-uniform ranks (all within 0.07 of 2.0), whereas the LLM judge still reports clear preferences—preferences that, as our bias analysis shows, track joke length and presentation order rather than humor.

5 Conclusion

Through our controlled ablation study across the Qwen3 family, we find that scaling parameter count yields a consistent but *non-significant* trend in human-judged humor quality overall. However, restricting analysis to high-agreement themes reveals a *significant* preference for the largest model ($p = 0.039$), indicating that scaling effects are real but frequently masked by a theme-dependent "quality floor."

Crucially, our study highlights the potential danger of using LLM evaluators for creative generation tasks. In stark contrast to human annotators, the automated judge reports a perfect scaling correlation ($\rho = 1.0$) by systematically relying on substantial procedural artifacts, exhibiting 73.5% positional bias and 62.2% length bias. Automated judges may distort subjective quality differences and can be unreliable as standalone tools for assessing creativity.

Limitations

Our study has several limitations that highlight directions for future work. First, regarding evaluation robustness, each model generated only a single joke per theme, precluding the assessment of within-model variance. Furthermore, our dataset is relatively small, and our automated evaluation relies on a single LLM judge (DeepSeek-V3.2), leaving inter-judge reliability untested. Second, the human evaluation was constrained by a small sample size and overall low annotator agreement, and manual ratings were only obtained for three of the five evaluated models. Third, by using the `/no_think` token, we explicitly disabled model reasoning to isolate generative capability. However, recent findings suggest that deliberative inference significantly improves an LLM's humor competence, a factor our design excludes (Narad et al., 2025). Fourth, it remains unclear whether larger models generate novel jokes or retrieve memorized examples (Amir, 2025; Jentzsch and Kersting, 2023). Although our judging prompt includes originality as a criterion, the pipeline parses only the final `WINNER: A/B` token rather than per-criterion scores, so we cannot assess how the judge weighs originality against structural proxies. Finally, while we ground our qualitative analysis in the Incongruity-Resolution framework (§4.5), extending this to additional humor theories (e.g., benign violation, superiority) remains future work.

References

O. Amir. 2025. Are AI-generated jokes truly original? Charting the "joke space.". In *Proceedings of the 16th International Conference on Computational Creativity (ICCC)*.

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

Santiago Castro, Luis Chiruzzo, Aiala Rosá, Diego Garat, and Guillermo Moncecchi. 2018. A crowd-annotated Spanish corpus for humor analysis. In *Proceedings of the 6th International Workshop on Natural Language Processing for Social Media*, pages 7–11.

Iurii Evstafev. 2025. Optimizing humor generation in LLMs: Temperature configurations and architectural trade-offs. *arXiv preprint arXiv:2504.02858*.

Jack Hessel, Ana Marasović, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Ali Farhadi, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from The New Yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 688–714.

Sophie Jentzsch and Kristian Kersting. 2023. ChatGPT is fun, but it is not funny! humor is still challenging for large language models. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*, pages 325–340. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Hyunwoo Kim and Lydia B. Chilton. 2025. AI humor generation: Cognitive, social and creative skills for effective humor. *arXiv preprint arXiv:2502.07981*.

Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability. *Departmental Papers (ASC)*.

Anirudh Mittal and 1 others. 2021. So you think you're funny?: Rating the humour quotient in standup comedy. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

R. Narad, S. Suresh, J. Chen, P. S. Dysart-Bricken, B. Mankoff, R. Nowak, and 1 others. 2025. Which LLMs get the joke? probing non-stem reasoning abilities with HumorBench. *arXiv preprint arXiv:2507.21476*.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Elisa Sandri, Elisa Leonardelli, and Sara Tonelli. 2023. Why don't you do it right? Analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 2428–2440.

Jerry M. Suls. 1972. A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis. In Jeffrey H. Goldstein and Paul E. McGhee, editors, *The Psychology of Humor: Theoretical Perspectives and Empirical Issues*, pages 81–100. Academic Press.

Alexey Tikhonov and Pavel Shtykovskiy. 2024. Humor mechanics: Advancing humor generation with multistep reasoning. In *Proceedings of the 15th International Conference on Computational Creativity (ICCC)*.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, and 1 others. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, and 1 others. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS)*.

A Experimental Setup

A.1 Prompts

Joke generation prompt. The following system prompt is used for all five Qwen3 models, with `max_tokens=128` and thinking disabled (`/no_think`):

> *"You are a stand-up comedian known for sharp, original observational humor. When given a theme, write one short joke about it — either a classic setup/punchline or a punchy one-liner. The joke must be self-contained, land clearly, and end on the punchline. Output only the joke text, nothing else — no title, no explanation, no commentary."*

The user message is simply the theme text (e.g., "Autocorrect text message fails").

Judging prompt. The DeepSeek-V3.2 judge receives the following system prompt:

> *"You are an expert comedy critic with a sharp, discerning sense of humor. Your task is to decide which of two jokes is funnier."*

The user message follows this template:

Theme: "{theme}"

Joke A:
{joke_a}

Joke B:
{joke_b}

Which joke is funnier? Consider:
- Comedic timing and structure
- Originality of the idea
- Clarity and strength of the punchline
- How well it fits the theme

First, write 2–3 sentences explaining your reasoning. Then on the final line write ONLY:
WINNER: A
or
WINNER: B

A.2 Annotation Protocol & Interface

Annotation protocol. Annotators are split into two independent groups of six. Evaluation is conducted simultaneously using a split projected screen that displays different themes to each group (Figure 2). For each theme, three anonymized jokes are presented in a random order for 60 seconds while annotators independently rank them from funniest to least funny.

B Additional Results & Visualizations

B.1 LLM Judge Per-Theme Heatmap

B.2 Human Evaluation: All 40 Themes

Figures 4, 5, and 6 show the full aggregate results across all 40 themes, including the 18 low-agreement themes where models produce indistinguishable outputs. These figures complement the high-agreement subset analysis presented in the main text (§4.3).

B.3 High-Agreement Subset Figures

Figures 7, 8, and 9 provide visualization for the high-agreement subset analysis presented in the main text (§4.3).

C Qualitative Examples & Data

C.1 Example Jokes

Table 4 shows example model outputs for selected themes, illustrating cases where models are clearly distinguishable (high agreement) and cases where outputs converge (low agreement).

C.2 High-Agreement Theme List

The 22 themes meeting the high-agreement threshold (pairwise winner agreement ≥ 0.4) are listed below with their agreement scores and per-model mean ranks.

Figure 2: The split-screen annotation interface projected in the classroom. The left half displays a theme for Group 1 (themes 11–30), while the right half simultaneously displays a different theme for Group 2 (themes 31–50). This layout allows 12 annotators (two groups of six) to evaluate simultaneously without seeing each other's responses. Each set of three randomized, anonymized jokes (A, B, C) is displayed for 60 seconds, a duration determined via pilot testing, while annotators independently rank them from funniest to least funny.

Theme	**32B**	**30A3 (MoE)**	**235A22 (MoE)**
High agreement — Astrological signs (agreement = 1.0, 235A22 wins unanimously)			
Astrological signs	I checked my horoscope this morning and it said I'd have a lucky day—turns out the "lucky" part was the only part they got right.	I told my zodiac sign I was breaking up, and it said, "But we're meant to be!" I said, "Yeah, I know, that's why I'm going to find someone who's not on my level."	I checked my horoscope, and it said today is a great day to reconnect with someone from your past—so I texted my ex. Turned out, the universe meant your *dentist*.
Low agreement — Dealing with landlords (agreement = 0.2, three-way tie at rank 2.0)			
Dealing with landlords	I told my landlord his rent increase was like a horror movie—it had a slow build, a shocking revelation, and left me screaming into the void.	I told my landlord I was going to start a band, he said, "Great, but can you at least pay rent on time?"	I asked my landlord if the apartment included a gym, and he said, "No, but the stairs are free."

Table 4: Example themes illustrating high- and low-agreement cases. Full generated outputs for all 50 themes across all 5 models are available in the repository. The high-agreement examples show clear qualitative differences in punchline strength, whereas low-agreement jokes demonstrate similar structural competency but equal lack of comedic effect.

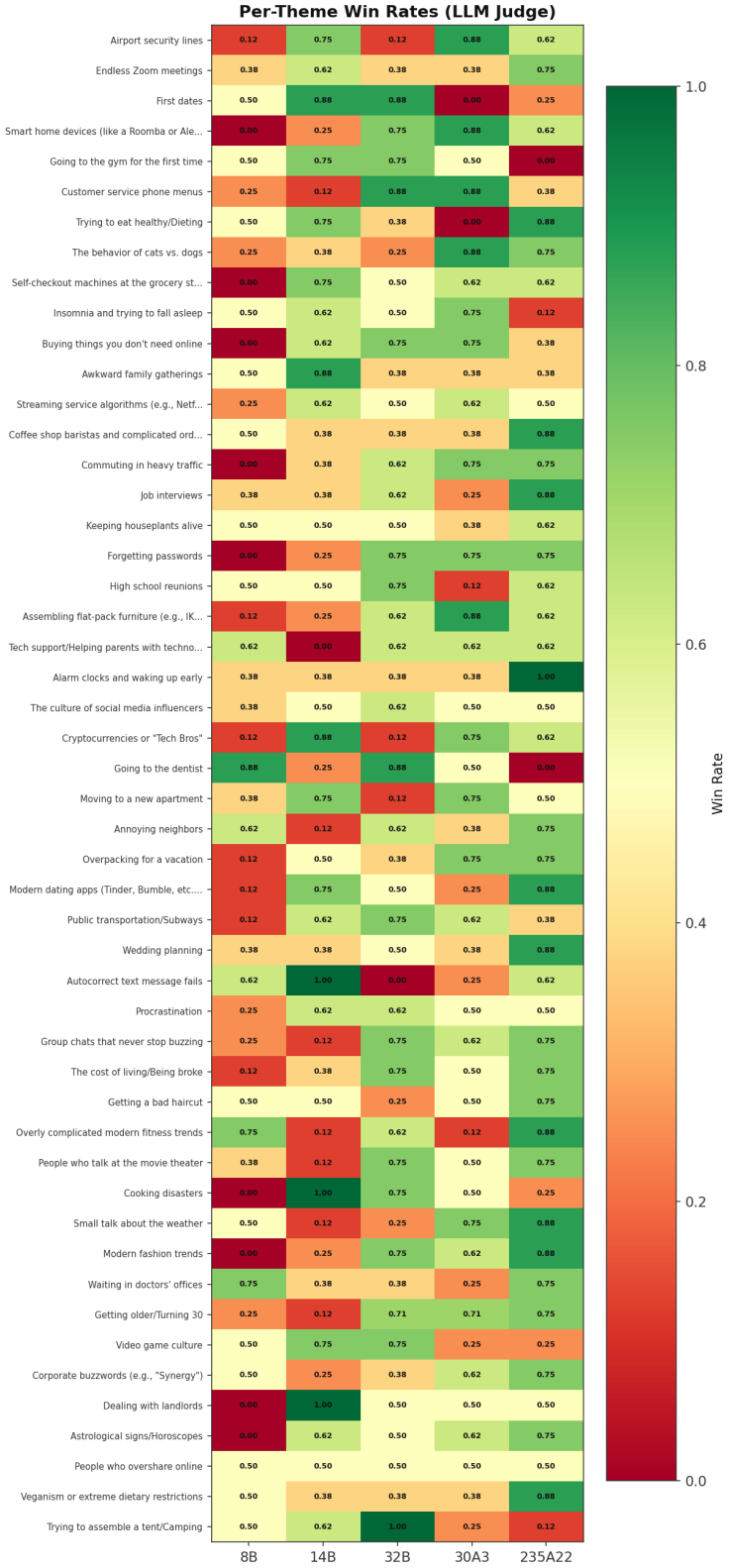

Figure 3: LLM judge per-theme win rates across all 5 models (50 themes). Warmer colors (green) indicate higher win rates. The 235A22 model shows consistently high win rates while 8B shows consistently low win rates, but per-theme variation is substantial. Reading horizontally reveals a given model's performance variance across topics, while reading columns vertically illustrates the judge's size preference on specific topics.

Theme	Winner	Agreement	32B	30A3	235A22
Astrological signs/Horoscopes	235A22	1.000	3.00	2.00	**1.00**
Trying to assemble a tent/Camping	32B	1.000	**1.00**	2.67	2.33
Getting a bad haircut	235A22	0.667	2.00	2.83	**1.17**
People who talk at the movie theater	30A3	0.667	2.00	**1.33**	2.67
Getting older/Turning 30	32B	0.667	**1.17**	2.17	2.67
Coffee shop baristas	235A22	0.667	2.00	2.83	**1.17**
Procrastination	30A3	0.467	2.83	**1.33**	1.83
The cost of living/Being broke	30A3	0.467	1.83	**1.50**	2.67
Veganism/dietary restrictions	235A22	0.467	1.83	2.83	**1.33**
Awkward family gatherings	235A22	0.467	2.50	2.00	**1.50**
Group chats	235A22	0.400	2.17	2.50	**1.33**
Modern fitness trends	235A22	0.400	2.17	2.50	**1.33**
Cooking disasters	30A3	0.400	2.00	**1.67**	2.33
Modern fashion trends	32B	0.400	**1.50**	1.83	2.67
Waiting in doctors' offices	32B	0.400	**1.50**	2.17	2.33
People who overshare online	235A22	0.400	2.17	2.33	**1.50**
Buying things online	235A22	0.400	2.33	2.33	**1.33**
Streaming service algorithms	32B	0.400	**1.67**	2.50	1.83
Job interviews	235A22	0.400	1.83	2.33	**1.83**
Forgetting passwords	30A3/235A22	0.400	2.67	**1.67**	**1.67**
Cryptocurrencies/"Tech Bros"	30A3	0.400	2.67	**1.50**	1.83
Annoying neighbors	32B	0.400	**1.67**	2.33	2.00

Table 5: High-agreement themes (≥ 0.4 pairwise winner agreement). Bold indicates best (lowest) mean rank. 235A22 wins 9 of 22 themes, 32B wins 8, 30A3 wins 4, and 1 is tied.

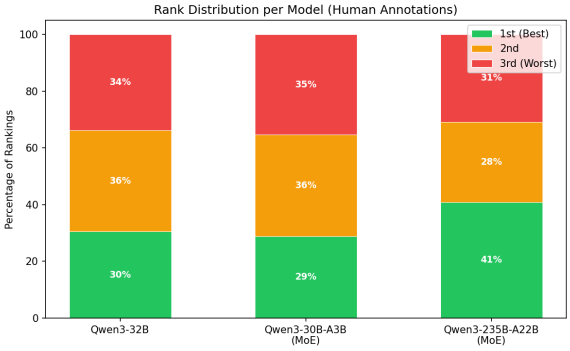

Figure 4: Distribution of 1st, 2nd, and 3rd place rankings across all 40 themes. Bar segments denote the percentage of times each model received a specific rank from human evaluators. While 235A22 receives 41% of first-place votes vs. 33% expected by chance, the overall Friedman test is non-significant ($p = 0.155$).

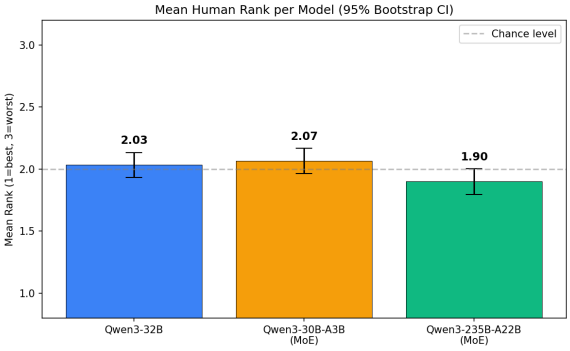

Figure 5: Mean human rank per model with 95% bootstrap CIs across all 40 themes. Points denote the average human rank (lower is better), and vertical lines show 95% confidence intervals. All confidence intervals overlap and include the chance level (2.0). This visualizes the aggregate "quality floor," showing no significant separation between models.

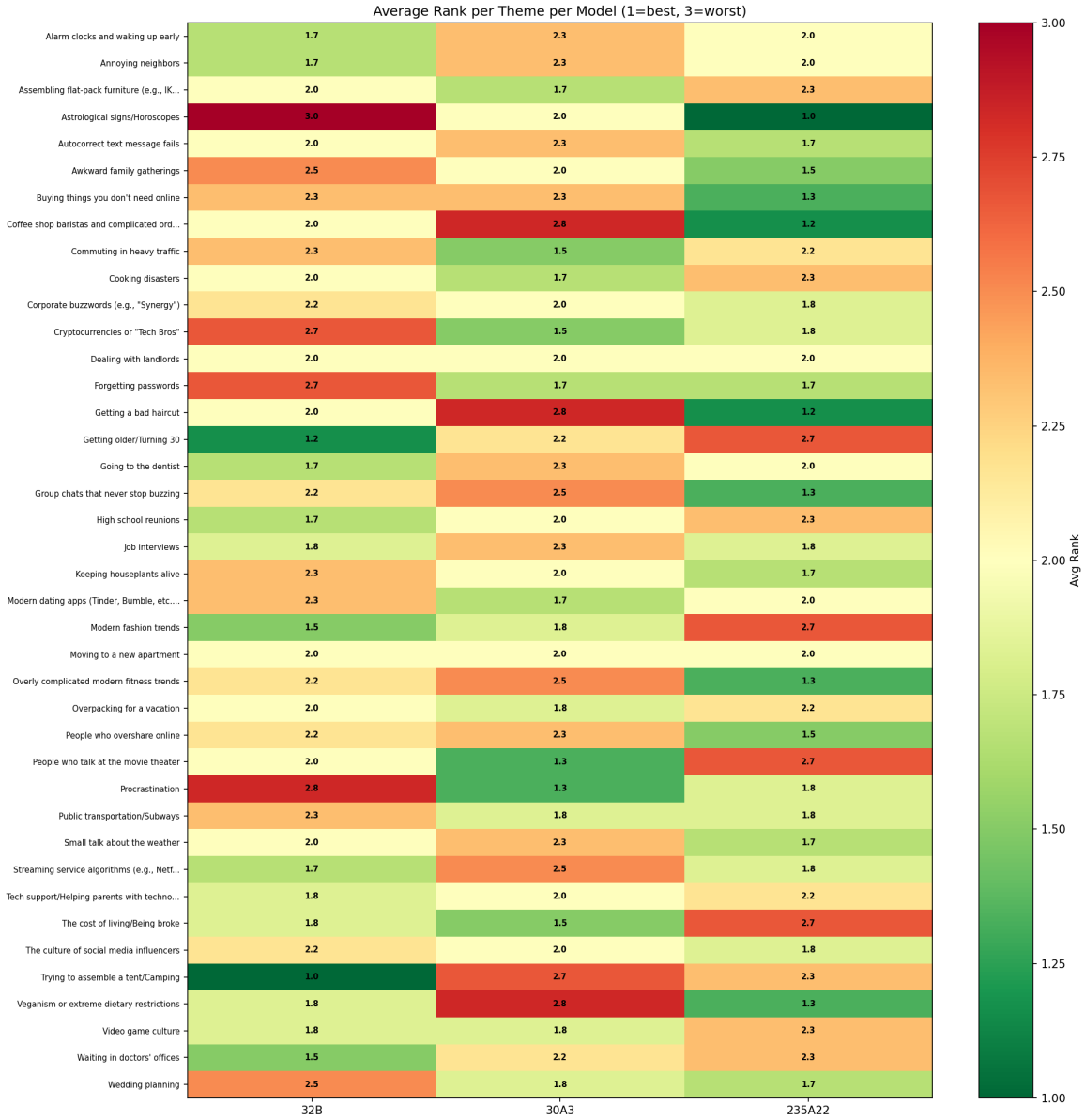

Figure 6: Per-theme average human rank across all 40 themes (1 = best, 3 = worst). Some themes show strong differentiation (dark green for the winner), while many cluster near 2.0 (yellow), indicating that annotators could not distinguish the models. The predominance of yellow cells highlights how frequently models produced indistinguishable outputs.

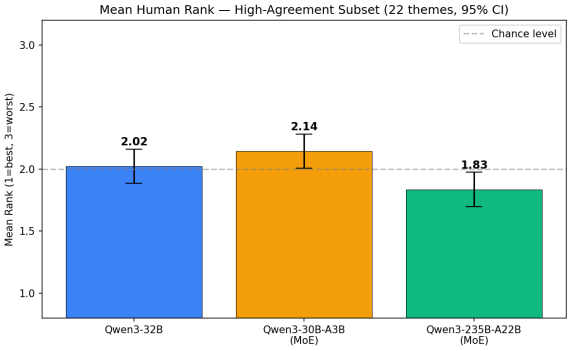

Figure 7: Mean human rank on the 22 high-agreement themes (95% bootstrap CI). Points denote the average human rank on this restricted subset, and vertical lines show 95% confidence intervals. The clear separation of 235A22's interval from 30A3's visually confirms the significant statistical preference ($p = 0.039$) found in this subset.

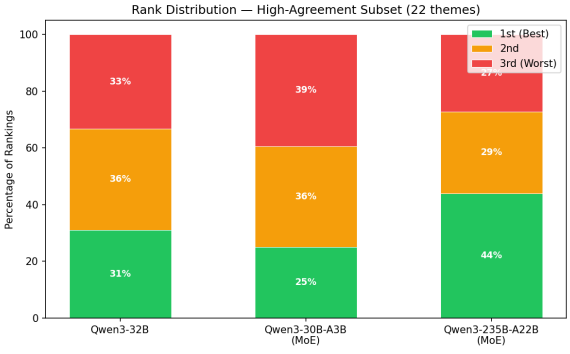

Figure 8: Rank distribution on the 22 high-agreement themes. Bar segments denote the percentage of times each model was ranked 1st, 2nd, or 3rd within this high-agreement subset. Compared to the aggregate distribution, 235A22 secures a noticeably larger majority of first-place votes here.

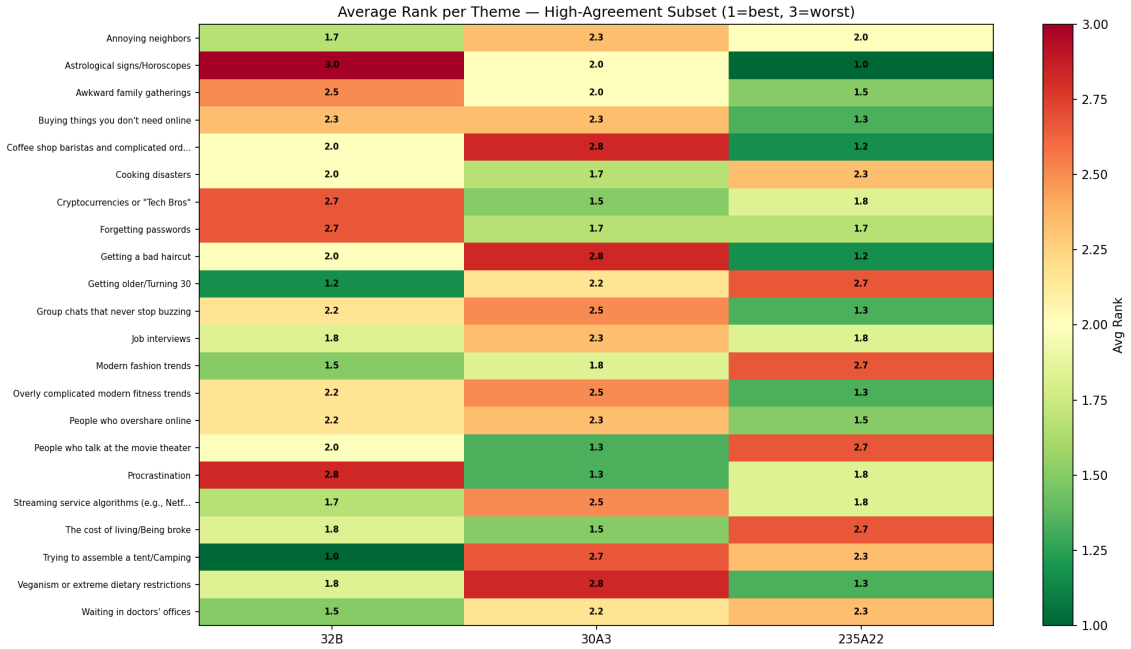

Figure 9: Per-theme average human rank on the 22 high-agreement themes (1 = best, 3 = worst). Dark green indicates better (lower) mean ranks, while light yellow indicates chance-level performance. Sharper color contrasts indicate clearer model differentiation compared to the full 40-theme heatmap. The absence of yellow tie-columns emphasizes that these are themes where distinct, perceptible quality differences emerged.