

000 001 002 003 004 005 006 007 008 009 010 A MIN-P BLUEPRINT FOR MORE RIGOROUS SCIENCE IN EMPIRICAL MACHINE LEARNING RESEARCH

005 **Anonymous authors**

006 Paper under double-blind review

009 ABSTRACT

011 In light of a growing crisis of rigor in empirical machine learning research, this
 012 paper provides a blueprint for conducting more meticulous science. We present a
 013 detailed case study of “Turning Up the Heat: Min-P Sampling for Creative and Co-
 014 herent LLM Outputs” (Nguyen et al., 2024), a high-visibility ICLR 2025 Oral paper
 015 that introduced a new method for sampling from language models called `min-p`.
 016 The original work claimed that `min-p` sampling achieves superior quality and
 017 diversity over established methods. However, our comprehensive re-examination
 018 of the original paper’s four main lines of evidence demonstrates that its conclusions
 019 are invalidated by its own data. Our re-analysis reveals that: (1) The original
 020 human evaluations omitted one-third of the collected data, applied statistical tests
 021 incorrectly, and inaccurately described qualitative feedback; a correct analysis
 022 shows `min-p` did not outperform baselines. (2) Extensive hyperparameter sweeps
 023 on NLP benchmarks show `min-p`’s claimed superiority vanishes when controlling
 024 for the volume of hyperparameter tuning. (3) The LLM-as-a-Judge evaluations
 025 suffered from methodological ambiguity and appear to have reported results incon-
 026 sistently, favoring `min-p`. (4) Claims of widespread community adoption were
 027 found to be unsubstantiated and were retracted. From this case study, we derive a
 028 blueprint for more rigorous research. Key lessons include the critical need to com-
 029 pare methods fairly by controlling for hyperparameter tuning, to apply statistical
 030 tests transparently and correctly (e.g., correcting for multiple comparisons), to prac-
 031 tice full data transparency, and to scrutinize qualitative summaries, methodological
 032 clarity, and potentially selective reporting. Adhering to these principles is essential
 033 for ensuring the validity of scientific claims and fostering genuine progress in the
 034 field of machine learning research.

035 1 INTRODUCTION

036 Machine learning research is currently experiencing crises on multiple fronts (Kim et al., 2025;
 037 Schaeffer et al., 2025b): The number of submissions to each conference is skyrocketing (Paper Copilot,
 038 2025), reproducibility concerns are giving rise to a standalone machine learning reproducibility
 039 conference (Pineau et al., 2017; Sinha et al., 2023), and scandals concerning prominent publications
 040 are proliferating, e.g., (Carlini, 2020; Agarwal et al., 2021; Carlini et al., 2021; 2022; Kirsch, 2022;
 041 Schaeffer et al., 2022; Orabona, 2023; Schaeffer et al., 2023; Gerstgrasser et al., 2024; Kirsch, 2024;
 042 Markov, 2024; Schaeffer et al., 2024; Chandak et al., 2025; Golechha et al., 2025; Ivanova et al.,
 043 2025; Kazdan et al., 2025; Maini & Suri, 2025; Zhang et al., 2025; Schaeffer et al., 2025a).

044 In this work, we push back against these trends by providing a blueprint for performing more
 045 meticulous and rigorous science in empirical machine learning research. Our blueprint is based on
 046 a detailed case study of a high-visibility publication: “Turning Up the Heat: Min-P Sampling for
 047 Creative and Coherent LLM Outputs” (Nguyen et al., 2024). This publication introduced a method
 048 for sampling from language models called `min-p` sampling, claiming `min-p` produces higher
 049 quality and more diverse outputs than existing sampling methods such as `top-k` (Fan et al., 2018) or
 050 `top-p` (Holtzman et al., 2020) sampling. The paper ranked as the 18th highest-scoring submission
 051 to ICLR 2025 and was selected for an Oral presentation. Given the potential impact of improving
 052 both quality *and* diversity of language model outputs, we carefully scrutinized the methodologies,
 053 data, analyses, code and conclusions presented in support of `min-p` across the authors’ four lines of
 evidence: (1) human evaluations, (2) natural language processing (NLP) benchmark evaluations, (3)

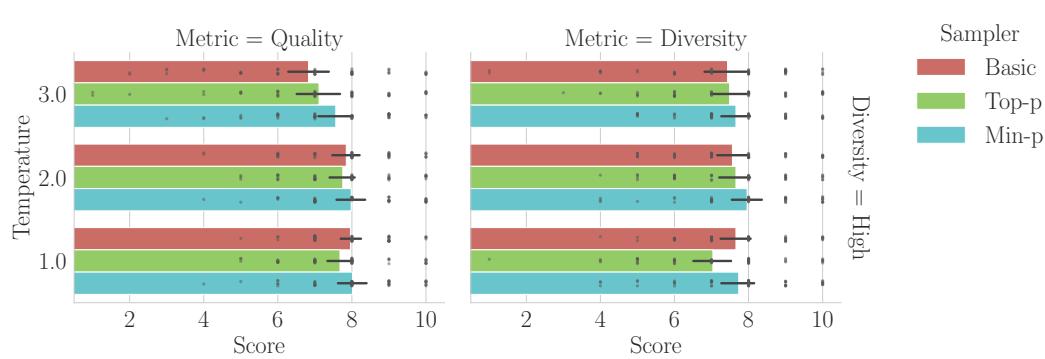


Figure 1: **Visualizing Human Evaluators’ Scores from Nguyen et al. (2024)’s Data Demonstrates Min-p Does Not “Consistently” Outperform Other Samplers.** Rather, the original paper’s data suggest min-p is largely indistinguishable from other samplers based on 95% confidence intervals.

LLM-As-A-Judge evaluations and (4) community adoption metrics. Our re-analyses of the evidence lead us to conclude that the paper’s own evidence invalidates its central claim: min-p sampling improves neither quality, nor diversity, nor the trade-off between quality and diversity.

Although Nguyen et al. (2024) presents a high-visibility example of scientific mistakes, the errors made in evaluating min-p are common in empirical machine learning research. Towards promoting rigorous science across the field, this case study offers several general lessons for researchers and reviewers alike. First, we develop a novel methodology for fairly comparing methods that require extensive hyperparameter tuning, which can also be used to detect potential cherry-picking. Second, we highlight best practices for statistical testing in human evaluations, particularly concerning the need to correct for multiple comparisons. Third, we demonstrate the critical role of data transparency and exploratory data analysis in uncovering flaws like omitted data and mischaracterized feedback. Ultimately, we argue that by establishing and adhering to robust methodologies, the field can better rein in questionable claims and ensure scientific progress advances.

2 RE-ANALYZING MIN-P’S HUMAN EVALUATIONS

We begin our case study by analyzing Nguyen et al. (2024)’s human evaluations, since human judgments are widely considered the gold standard for assessing language models (Van Der Lee et al., 2019; Roller et al., 2020; Howcroft et al., 2020; Clark et al., 2021; Liang et al., 2022; Khashabi et al., 2022; Chiang et al., 2024; Biderman et al., 2024; Schaeffer et al., 2025d). Prior work has exhaustively demonstrate human evaluations of language models frequently are laden with flaws Freitag et al. (2021); Belz et al. (2021); Thomson et al. (2024): We identified four issues.

2.1 HUMAN EVALUATORS’ SCORES FOR ONE OF TWO BASELINE SAMPLERS WERE OMITTED

Section 6 of Nguyen et al. (2024) states human participants evaluated min-p against a single baseline sampler, top-p, and their Table 4 presents results only these two samplers. However, when examining the paper’s data, we discovered that **scores for a second baseline sampler (basic sampling) were excluded from the methodology, the analysis and the results without mention or justification**. We publicly confirmed with the authors. These omitted scores comprised 1/3rd of the total collected scores. After we raised the issue, the omitted data were added to the Camera Ready’s Table 4, but the methodology, the results and the conclusions were not correspondingly updated. As we shall show, the omitted human evaluations data change the paper’s conclusions.

2.2 VISUALIZATIONS AND STATISTICAL TESTS FAIL TO SUPPORT CLAIM THAT MIN-P OUTPERFORMS OTHER SAMPLERS

Based on the human evaluators’ scores, Section 6 of Nguyen et al. (2024) concluded that min-p “consistently” outperformed top-p “across all settings”:

Metric	Alt. Hyp.	Temperature					
		$\tau = 1.0$		$\tau = 2.0$		$\tau = 3.0$	
		t	p	t	p	t	p
Quality	Min-p > Basic	0.33	.370	0.65	.260	3.13***†	.001
	Min-p > Top-p	2.05*	.023	1.18	.121	2.02*	.025
Diversity	Min-p > Basic	0.31	.378	1.86*	.034	0.85	.201
	Min-p > Top-p	2.64**	.006	1.44	.078	0.87	.195

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, † Significant after Bonferroni correction for 12 comparisons.

Note: All tests were paired t-tests with $df = 52$, one-sided (alternative = "greater")

Table 1: Hypothesis Testing of Human Evaluators' Scores Fails to Support Claim that Min-p Consistently Outperforms Other Samplers. To test whether evidence supports the claim that min-p "consistently outperforms" other samplers, we conducted one-sided paired t-tests using the authors' published data. Without correcting for multiple comparisons, evidence supports min-p's superiority in 5 of 12 comparisons at $\alpha = 0.05$ and 2 of 12 comparisons at $\alpha = 0.01$. After applying a Bonferroni correcting for multiple comparisons, evidence supports min-p's superiority in 1 of 12 comparisons at $\alpha = 0.05$ and 0 of 12 comparisons at $\alpha = 0.01$. For details, see Sec. 2.3.

"Overall, min-p sampling consistently scored higher than top-p sampling across all settings [...] A paired t-test confirmed that the differences in scores between min-p and top-p sampling were statistically significant ($p < 0.05$)."

However, **both visualizations and statistical hypothesis tests of the original human evaluation data suggest min-p is indistinguishable from the baselines in almost all settings.**

To briefly explain the human evaluation methodology, three samplers (basic, top-p and min-p) were compared in six conditions: three temperatures (1.0, 2.0, 3.0) and two diversity settings ("high" and "low") corresponding to different p hyperparameters. Humans were asked to score the generated outputs under two metrics: quality and diversity. Participants were excluded if they failed attention checks. For more information, please see the original manuscript.

We focused on the "high" diversity setting for three reasons: (1) The claimed advantage of min-p is that it provides both high quality and high diversity, whereas other samplers typically trade one against the other. (2) The authors publicly told us to focus on the high diversity setting, writing, "the low [diversity] settings were quite experimental." (3) We believe top-p's p hyperparameter in the low diversity setting was poorly chosen; indeed, after we raised said concerns, the authors ran a new human evaluation (more in Sec. 2.4) which changed the low diversity top-p p from $0.1 \rightarrow 0.9$.

Using the original paper's data, Fig. 1 reveals that the three samplers provide similar quality and similar diversity, with 95% confidence intervals frequently overlapping. To more rigorously assess the claim that min-p consistently outperforms other samplers, we conducted 12 one-sided paired t-tests for each metric (quality or diversity), temperature (1.0, 2.0, 3.0) and baseline sampler (min-p versus basic, min-p versus top-p). In each test, the null hypothesis is min-p's score is less than or equal to the other sampler's score, and the alternative hypothesis is min-p's score is greater than the other sampler's score. Statistical test results are displayed in Table 1. Without correcting for multiple comparisons, we found evidence to reject the null hypotheses in 5 of 12 tests at $\alpha = 0.05$ and 2 of 12 tests at $\alpha = 0.01$. After applying a Bonferroni correction for multiple comparisons, we found evidence to reject the null hypothesis in 1 of 12 tests at $\alpha = 0.05$ and 0 of 12 tests at $\alpha = 0.01$. Furthermore, given that the original paper claims that min-p "consistently" scores higher, an Intersection-Union Test (IUT) may be the appropriate statistical test, where the alternative hypothesis is that min-p is better in all 12 comparisons and the null hypothesis is the set complement. Since the largest p-value of the 12 comparisons is 0.378, under the IUT, we again find insufficient evidence to reject the null hypothesis at both $\alpha = 0.05$ and $\alpha = 0.01$. **Based on the original paper's data, there is insufficient evidence to support the claim that min-p consistently outperforms baseline samplers across all settings.**

The original paper's statistical analysis reached an incorrect conclusion due to a combination of two factors: First, despite claiming that min-p "consistently scored higher" "across all settings"

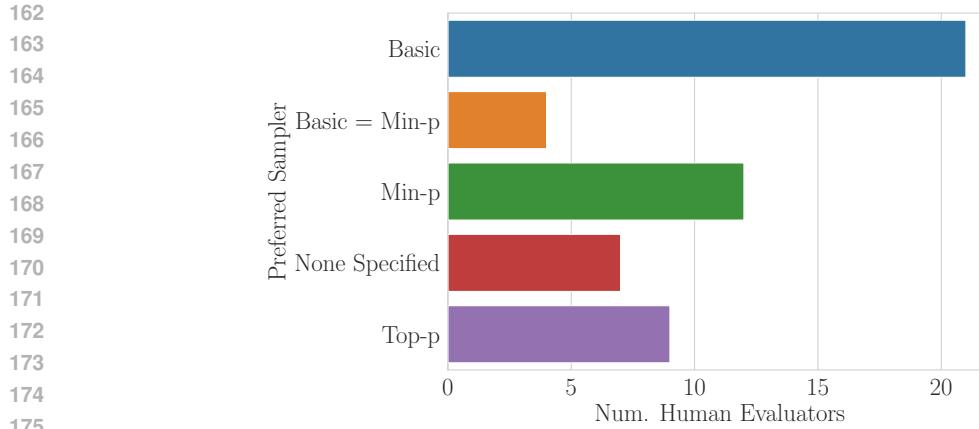


Figure 2: **Manual Annotation of Human Evaluators’ Qualitative Responses Fail to Support Claim that Min-p Was the Preferred Sampler.** We manually annotated responses from human annotators regarding their preferred sampler(s) at the end of the original paper’s study. The responses suggest min-p was not the most preferred sampler. Example responses are provided in Appendix B.

(metric, temperature, and diversity), the paper pooled data across all settings and performed a single t-test, which tests whether min-p scored higher on average. Second, pooling over all settings was misleading because top-p ’s hyperparameter p was poorly chosen in the “low” diversity condition in a way that pulled top-p down significantly; the authors said publicly to ignore this particular low diversity condition and subsequently changed p in their new human experiment (Sec. 2.4).

2.3 HUMAN EVALUATORS’ QUALITATIVE RESPONSES FAIL TO SUPPORT CLAIM THAT MIN-P IS PREFERRED OVER OTHER SAMPLERS

At the end of the human evaluation study, the original paper asked human participants to qualitatively describe which sampler(s) they preferred. The paper claimed that human evaluators’ qualitative responses support min-p over top-p :

“Participants frequently noted that outputs generated with min-p sampling were more coherent and creative, especially at higher temperatures.”

However, we believe that **the paper’s qualitative human responses contradict this**. We manually annotated the qualitative responses, then visualized our annotations of the humans’ expressed preferences (Fig. 2) and publicly posted our annotations in the same format. We found two results: (1) more human evaluators explicitly preferred **basic** sampling than preferred min-p sampling, which was not immediately obvious because the **basic** sampling data were previously excluded (Sec. 2.1), and (2) min-p was only slightly preferred over top-p . We provide quotations from human evaluators favoring **basic** sampling in Appendix B.

2.4 NEW HUMAN EVALUATION STUDY SHOWS MIN-P DOES NOT OUTPERFORM BASELINES IN QUALITY, IN DIVERSITY, OR IN A TRADEOFF BETWEEN QUALITY AND DIVERSITY

In response to our feedback, the authors conducted and added a new human evaluation study to Appendix C.2. Their new study made multiple methodological changes:

- Different sampler implementation: switched from applying temperature *after* truncation to applying temperature *before* truncation.
- Different distribution of human participants from Prolific.
- Different hyperparameters for top-p : switched from 0.1 and 0.9 to 0.9 and 0.95.
- Different hyperparameters for min-p : switched from 0.2 and 0.05 to 0.1 and 0.05.
- Different allotted reading time: increased from 30 minutes to 45 minutes.

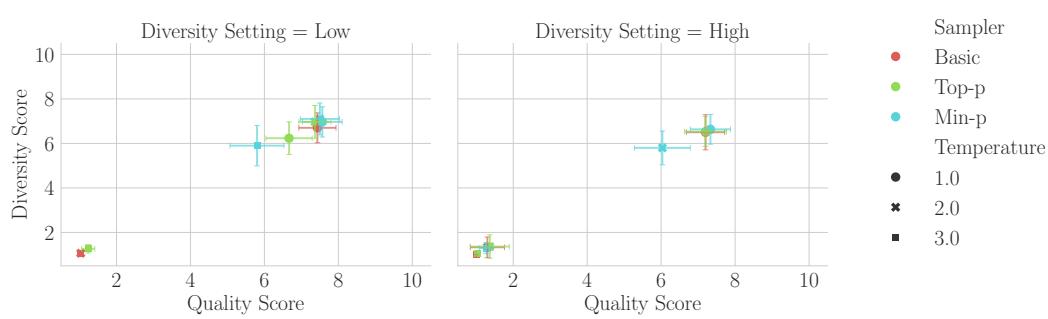


Figure 3: **New Human Evaluation Study Suggests Min-p Does Not Outperform Baselines in Quality, in Diversity or in a Pareto-Optimal Tradeoff Between Quality and Diversity.** Visualization of scores from Nguyen et al. (2024)’s second human experiment. For researchers or practitioners seeking maximal quality and maximal diversity, these human evaluation results suggest that min-p offers no apparent advantage over `basic` sampling or `top-p` sampling.

- Different sampled text: 3 short paragraphs were replaced with a single complete story.
- Different rubric for human participants to evaluate sampled outputs.

Regarding the new human evaluation data and results, we share two discoveries here: First, more generally, the data show again that min-p does not outperform baseline sampling methods in quality, in diversity or in a favorable tradeoff between quality and diversity (Fig. 3). Second, we believe one value is incorrectly reported: in Nguyen et al. (2024)’s Table 15, the average score of min-p at $p = 0.05$ and temperature $T = 2$ is reported as 7.80, but based on the authors’ publicly posted data, we believe the correct numerical value should be 5.80. **For anyone seeking higher quality or diversity, min-p offers no apparent advantage over previously existing samplers.**

3 EXTENDING $\text{MIN-P}'$ S NLP BENCHMARK EVALUATIONS

We next turned to the original paper’s NLP benchmark evaluations of several models on GSM8K with Chain-of-Thought (Cobbe et al., 2021) and GPQA (5-shot) (Rein et al., 2023), which concluded that:

“ Min-p sampling achieves superior performance across benchmarks and temperatures.”

3.1 THOROUGH HYPERPARAMETER SWEEP ON GSM8K CONTRADICTS CLAIM OF $\text{MIN-P}'$ S SUPERIORITY

Verifying whether min-p achieves superior performance was difficult because it was not clear whether sampling methods were compared equally. To test this, we used the authors’ code to conduct an extensive sweep on GSM8K CoT over the following models, samplers and hyperparameters:

- **9 Models:** Qwen 2.5 (Qwen et al., 2025) 0.5B, 1.5B, 3B and 7B; Mistral 7Bv0.1 (Jiang et al., 2023); Llama (Grattafiori et al., 2024) 3.1 8B and 3.2 3B; Gemma 2 (Team et al., 2024) 2B and 9B.
- **2 Model Stages:** Pre-trained (“Base”) and Post-Trained (“Instruct”).
- **4 Samplers:** `basic`, `top-p`, `top-k`, `min-p`.
- **31 Temperatures:** 0.0 (“greedy”) to 3.0 in increments of 0.1.
- **6 Hyperparameters Per Sampler:** We chose 6 hyperparameters per sampler, except for `basic` which has no hyperparameter beyond temperature. The values were taken from the original paper; some were lightly edited to make them more evenly distributed:
 - `basic`: No hyperparameters other than temperature.
 - `top-k`: $k \in \{10, 30, 50, 100, 150, 200\}$.
 - `top-p`: $p \in \{0.99, 0.98, 0.95, 0.9, 0.8, 0.7\}$.
 - `min-p`: $p \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.3\}$.

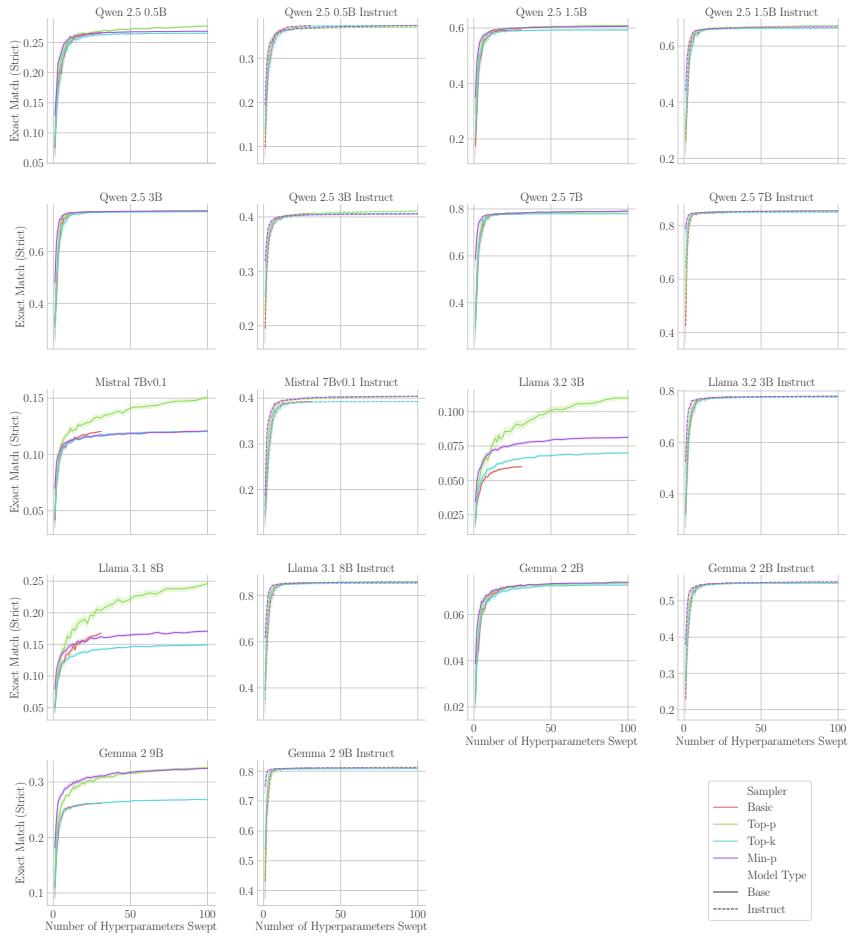


Figure 4: Min-P Does Not Consistently Outperform Other Samplers on GSM8K When Controlling For Hyperparameter Volume. In our first analysis, we measured how the maximum Exact Match (Strict) for each sampler improves as the number of hyperparameters increases. Basic sampling has only a temperature hyperparameter, and is thus swept to a lesser extent.

- **3 Random Seeds for Sampling:** $\{0, 1, 2\}$

Due to our compute budget, we only evaluated GSM8K CoT (albeit under two prompt formats, for reasons explained below). This sweep and the sweep below required ~ 6000 Nvidia A100-hours.

To evaluate how performant each sampler is, we first averaged over the three sampling seeds and then conducted two complementary analyses:

1. For each sampler, we subsampled an equal number of hyperparameters ranging from $N = 1$ to $N = 100$ and computed the maximum Exact Match (Strict) score achieved by the sampled subset of size N . We repeated this process 150 times, averaging over the subsampled subsets' scores. This "Best-of- N " analysis (Nakano et al., 2021; Stiennon et al., 2020; Hughes et al., 2024; Schaeffer et al., 2025c) tells us the best possible performance each sampler will likely obtain as its hyperparameter space increases.
2. For $N = 1$ to $N = 100$, we subsampled N hyperparameters per sampler and computed the difference of the maximum Exact Match (Strict) score achieved by min-p minus the maximum score achieved by any other sampler. We repeated this process 150 times, averaging over the subsampled subsets. This tells us by how much min-p outperforms all other samplers, controlling for the size of hyperparameter space of each sampler.

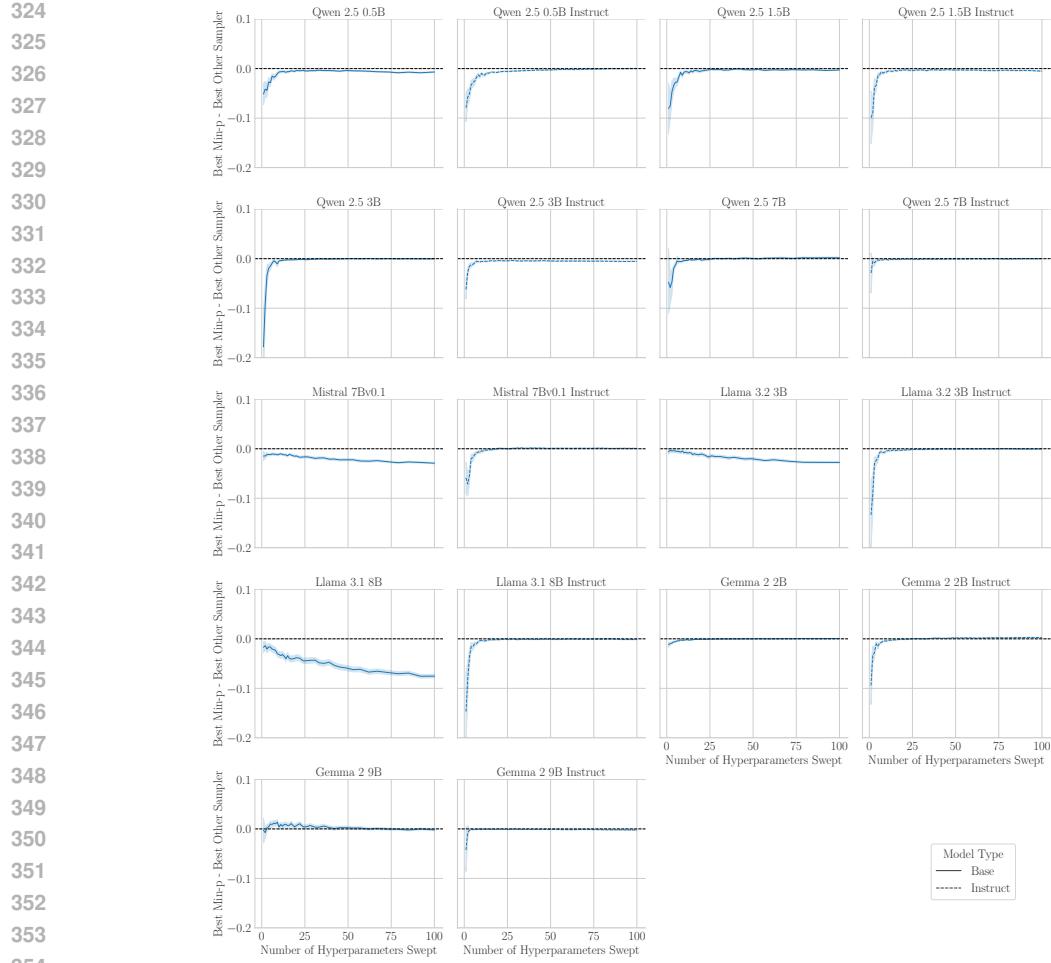


Figure 5: Min-p Does Not Consistently Outperform Other Samplers on GSM8K When Controlling For Hyperparameter Volume. In our second analysis, we measured how the difference between min-p’s highest score and the best non-min-p sampler’s highest score changes as the number of swept hyperparameters increases. Min-p matches or underperforms other samplers.

Both analyses reached consistent results: min-p does not outperform other samplers when equalizing the volume of hyperparameter space. Fig. 4 and Fig. 5 respectively demonstrate that min-p is largely indistinguishable from other samplers. After we showed these results to the authors, they informed us that their code default used the incorrect benchmark prompt formatting. We reran the experiments using standard formatting of GSM8K CoT prompts. The results were nearly identical (Appendix C), with one small difference: min-p does produce higher scores for 2 of 12 language models. Min-p does not outperform other samplers when controlling for hyperparameter volume.

4 INVESTIGATING MIN-P’S LLM-AS-A-JUDGE EVALUATIONS

We next turned to the original paper’s LLM-as-a-Judge evaluations (Zheng et al., 2023), specifically AlpacaEval creative writing evaluations (Dubois et al., 2023).

4.1 UNDER-SPECIFIED AND INDIRECT METHODOLOGY HINDERS REPRODUCTION AND INTERPRETATION

The methodology in the manuscript was under-specified in several ways: There is no mention which model(s) were sampled from, which model(s) served as the judge(s), or how hyperparameters were chosen or swept. Additionally, there is no description of uncertainty for reported win rates, meaning

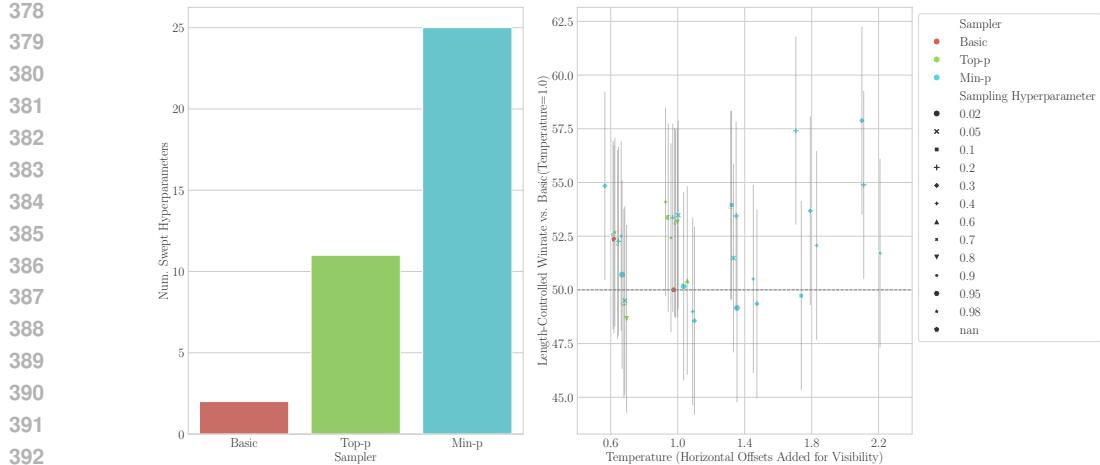


Figure 6: Nguyen et al. (2024)’s LLM-As-A-Judge Evaluations Suggest Min-p Typically Matches Other Samplers Despite 2 \times to 10 \times More Hyperparameter Tuning. Left: Nguyen et al. (2024) swept min-p with more than twice as many hyperparameters as top-p and more than ten times as many hyperparameters as basic. Right: Pairwise comparisons show min-p typically performs on-par with other samplers. Data were obtained from a public GitHub repository.

win rates may be indistinguishable from chance (50.00%). Furthermore, the experiment seems designed in a manner that introduces a confounder. As context, AlpacaEval reports win rates between pairwise comparisons. Instead of directly comparing min-p against other samplers, the authors compared each sampler against a common fixed sampler: basic($\tau = 1.0$). This comparison strategy is indirect; comparing directly against min-p would offer a clearer assessment of its superiority while using the same number of comparisons. The experimental design is additionally concerning because LLM-judge preferences are not transitive, as shown by recent research (Xu et al., 2025). Therefore, comparing all methods to basic($\tau = 1.0$) provides no reliable inference about min-p’s performance relative to top-p or basic at other temperatures. **These under-specified aspects of the methodology and the indirect experimental design make drawing conclusions difficult.**

4.2 MIN-P RECEIVED MORE HYPERPARAMETER TUNING YET FREQUENTLY FAILS TO WIN

Closely scrutinizing (ongoing work to publish) the data revealed two more insights: (1) min-p received $\sim 2\times$ more hyperparameter tuning than top-p sampling and $\sim 10\times$ more tuning than basic sampling (Fig. 6, left), potentially tilting the scales in its favor. (2) the win-rates show that min-p frequently fails to outperform top-p and basic sampling, especially when accounting for confidence intervals; we visualized the new data with 95% confidence intervals (with horizontal offsets added for visibility) (Fig. 6, right).

4.3 TABLE 3(B) REPORTED THE HIGHER OF TWO SCORES FOR MIN-P BUT THE LOWER OF TWO SCORES FOR TOP-P

As evidence for the LLM-As-A-Judge evaluation scores in the original paper’s Table 3(b), the first author publicly shared a Telegram link that showed the higher of two scores was reported for min-p (the reported win rate of 52.01 corresponds to $p = 0.05$, but $p = 0.01$ yields a lower win rate of 50.14) but the lower of two score was reported for top-p (the reported win rate of 50.07 corresponds to $p = 0.9$, but $p = 0.98$ yields a higher win rate of 50.43).

5 MIN-P’S CLAIMED GITHUB REPOSITORIES & STARS WERE UNSUBSTANTIATED AND RETRACTED

Nguyen et al. (2024) included specific claims about min-p’s widespread adoption:

432
433
434
435

“Community Adoption: Min-p sampling has been rapidly adopted by the open-source community, with over 54,000 GitHub repositories using it, amassing a cumulative 1.1 million stars across these projects.”

436
437
438
439
440
441
442
443
444
445
446
447
448

We attempted to verify these numbers through analysis of major GitHub language modeling repositories. Per our calculations, the combined GitHub stars of leading LM repositories (transformers, ollama, llama.cpp, vLLM, Unslot, mamba, SGLang, llama-cpp-python) sum to 453k stars as of March 2025, less than half the 1.1M stars claimed by min-p alone. We could not substantiate either 49k GitHub repositories or 1.1M GitHub stars. When we inquired how these numbers were calculated, the authors publicly stated that GitHub was searched for “min-p”, which yields many false positives. **The authors retracted both the 54k GitHub repository claim and the 1.1M GitHub stars claim from the ICLR 2025 Camera Ready manuscript.** We highlight this point because 3 of 4 ICLR 2025 reviewers and the Area Chair identified these retracted community adoption numbers as the main justification for their strong endorsement. The ICLR 2025 Camera Ready manuscript has a different statement of community adoption, which we believe remains misleading.

6 DISCUSSION AND LIMITATIONS

449
450
451
452
453
454

Scientific Conclusions: Our case study led us to conclude that the four lines of evidence presented by Nguyen et al. (2024) – (1) human evaluations, (2) NLP benchmark evaluations, (3) LLM-as-a-Judge evaluations, (4) community adoption – do not support min-p’s claimed superiority. While min-p is useful as another method to try, the original paper’s data and our extensions of the original paper’s data suggest that samplers perform approximately equally if given equal hyperparameter tuning.

455
456
457

Key Limitation: Our manuscript re-analyzes the evidence presented by Nguyen et al. (2024) and additional evidence created using the original paper’s code. *Conclusions here are based on that evidence.* We emphasize that new evidence might lead to different conclusions.

458
459

General Lessons for Reviewers and Researchers: This case study of min-p reveals several general lessons for more rigorous science in empirical machine learning research:

460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485

1. **Compare methods fairly by controlling for hyperparameter volume.** As demonstrated in the NLP benchmark analysis, a method’s advantage may disappear when the volume of hyperparameter space being searched is equalized across all contenders. Our “Best-of-N” analysis is an effective way to control for this and detect potential cherry-picking.
2. **Apply statistical tests rigorously and transparently.** The re-analysis of human evaluation data reveals how incorrect statistical practices can lead to false conclusions. This includes inappropriately pooling data across different experimental conditions, and, failing to correct for multiple comparisons when testing multiple hypotheses. Visualizing data with appropriate uncertainty estimates is another crucial step to prevent misinterpretation.
3. **Demand and practice full data transparency.** A key finding was that the original study omitted one-third of its human evaluation data without justification, and the inclusion of this data changed the paper’s conclusions. To allow for independent verification, researchers should release all collected data, annotations, and analysis code.
4. **Scrutinize qualitative summaries and unrealistic claims.** The original paper’s summary of qualitative feedback did not align with a direct reading of the evaluators’ responses. Furthermore, bold claims regarding community adoption were unsubstantiated, yet they heavily influenced the original reviewers’ assessments. Such claims must be carefully verified.
5. **Ensure methodological clarity for full reproducibility.** The LLM-as-a-Judge evaluations lacked crucial details hindering interpretation and reproduction. Empirical papers must provide enough detail for other researchers to faithfully replicate the work.
6. **Watch for inconsistent or selective reporting.** In the LLM-as-a-Judge results, the higher of two scores was reported for min-p, while the lower of two was reported for a baseline method. This selective reporting creates a misleading picture of the method’s performance. All results must be reported using a consistent methodology to avoid bias.

486 REFERENCES
487

488 Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare.
489 Deep reinforcement learning at the edge of the statistical precipice. *Advances in Neural Information
490 Processing Systems*, 34, 2021.

491 Anya Belz, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. The ReproGen shared task on
492 reproducibility of human evaluations in NLG: Overview and results. In Anya Belz, Angela Fan,
493 Ehud Reiter, and Yaji Sripada (eds.), *Proceedings of the 14th International Conference on Natural
494 Language Generation*, pp. 249–258, Aberdeen, Scotland, UK, August 2021. Association for
495 Computational Linguistics. doi: 10.18653/v1/2021.inlg-1.24. URL <https://aclanthology.org/2021.inlg-1.24/>.

496

497 Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi,
498 Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi,
499 Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal,
500 Wilson Y. Lee, Haonan Li, Charles Lovering, Niklas Muennighoff, Ellie Pavlick, Jason Phang,
501 Aviya Skowron, Samson Tan, Xiangru Tang, Kevin A. Wang, Genta Indra Winata, François Yvon,
502 and Andy Zou. Lessons from the trenches on reproducible evaluation of language models, 2024.
503 URL <https://arxiv.org/abs/2405.14782>.

504 Nicholas Carlini. Instahide disappointingly wins bell labs prize, 2nd place. Blog
505 post, dec 2020. URL <https://nicholas.carlini.com/writing/2020/instahide-disappointingly-wins-bell-labs-prize.html>. Accessed on
506 May 21, 2025.

507

508 Nicholas Carlini, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, and Florian
509 Tramer. Neuracrypt is not private, 2021. URL <https://arxiv.org/abs/2108.07256>.

510

511 Nicholas Carlini, Vitaly Feldman, and Milad Nasr. No free lunch in "privacy for free: How does
512 dataset condensation help privacy". *arXiv preprint arXiv:2209.14987*, 2022. URL <https://arxiv.org/abs/2209.14987>.

512

513 Nikhil Chandak, Shashwat Goel, and Ameya Prabhu. Incorrect baseline evaluations
514 call into question recent llm-rl claims. <https://safe-lip-9a8.notion.site/Incorrect-Baseline-Evaluations-Call-into-Question-Recent-LLM-RL-Claims-2012f1fbf0e0pvs=4>. Notion Blog.

515

516 Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng
517 Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena:
518 An open platform for evaluating llms by human preference, 2024. URL <https://arxiv.org/abs/2403.04132>.

519

520 Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith.
521 All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the
522 59th Annual Meeting of the Association for Computational Linguistics and the 11th International
523 Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7282–7296, 2021.

524

525 Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser,
526 Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John
527 Schulman. Training verifiers to solve math word problems, 2021. URL <https://arxiv.org/abs/2110.14168>.

528

529 Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos
530 Guestrin, Percy S Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for
531 methods that learn from human feedback. *Advances in Neural Information Processing Systems*,
532 36:30039–30069, 2023.

533

534 Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. *arXiv preprint
535 arXiv:1805.04833*, 2018.

536

537 Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey.
538 Experts, errors, and context: A large-scale study of human evaluation for machine translation.
539 *Transactions of the Association for Computational Linguistics*, 9:1460–1474, 2021.

540 Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes,
 541 Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang,
 542 David L. Donoho, and Sanmi Koyejo. Is model collapse inevitable? breaking the curse of
 543 recursion by accumulating real and synthetic data. *arXiv preprint arXiv:2404.01413*, 2024. URL
 544 <https://arxiv.org/abs/2404.01413>.

545 Satvik Golechha, Lucius Bushnaq, Euan Ong, Neeraj Kayal, and Nandi Schoots. Intricacies of feature
 546 geometry in large language models. In *The Fourth Blogpost Track at ICLR 2025*, 2025. URL
 547 <https://openreview.net/forum?id=Ut3ml7Hdwx>.

548 Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad
 549 Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan,
 550 Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev,
 551 Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru,
 552 Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak,
 553 Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu,
 554 Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle
 555 Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego
 556 Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova,
 557 Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel
 558 Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon,
 559 Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan
 560 Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet,
 561 Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde,
 562 Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie
 563 Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua
 564 Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak,
 565 Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley
 566 Chiu, Kunal Bhalla, Kushal Lakhota, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence
 567 Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas
 568 Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri,
 569 Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie
 570 Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes
 571 Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne,
 572 Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal
 573 Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong,
 574 Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic,
 575 Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie
 576 Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana
 577 Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie,
 578 Sharan Narang, Sharath Raparth, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon
 579 Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan,
 580 Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas
 581 Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami,
 582 Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti,
 583 Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier
 584 Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao
 585 Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song,
 586 Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe
 587 Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya
 588 Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei
 589 Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu,
 590 Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit
 591 Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury,
 592 Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer,
 593 Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu,
 Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido,
 Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu
 Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer,
 Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu,

594 Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc
 595 Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily
 596 Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers,
 597 Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank
 598 Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee,
 599 Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan,
 600 Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph,
 601 Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog,
 602 Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James
 603 Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny
 604 Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings,
 605 Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai
 606 Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik
 607 Veeraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle
 608 Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng
 609 Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madien Khabsa, Manav Avalani, Manish
 610 Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim
 611 Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle
 612 Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang,
 613 Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandini Santhanam,
 614 Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier,
 615 Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia
 616 Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro
 617 Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani,
 618 Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy,
 619 Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin
 620 Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu,
 621 Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh
 622 Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay,
 623 Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang,
 624 Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie
 625 Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta,
 626 Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman,
 627 Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun
 628 Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria
 629 Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru,
 630 Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz,
 631 Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv
 632 Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi,
 633 Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait,
 634 Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The
 635 llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.

636 Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text
 637 degeneration. In *International Conference on Learning Representations*, 2020.

638 David M Howcroft, Anya Belz, Miruna Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood,
 639 Simon Mille, Emiel Van Miltenburg, Sashank Santhanam, and Verena Rieser. Twenty years of
 640 confusion in human evaluation: Nlg needs evaluation sheets and standardised definitions. In *13th
 International Conference on Natural Language Generation 2020*, pp. 169–182. Association for
 641 Computational Linguistics, 2020.

642 John Hughes, Sara Price, Aengus Lynch, Rylan Schaeffer, Fazl Barez, Sanmi Koyejo, Henry Sleight,
 643 Erik Jones, Ethan Perez, and Mrinank Sharma. Best-of-n jailbreaking, 2024. URL <https://arxiv.org/abs/2412.03556>.

644 Desi R Ivanova, Ilija Ilievski, and Momchil Konstantinov. Towards more rigorous evaluations
 645 of language models. In *ICLR Blogposts 2025*, 2025. URL <https://iclr-blogposts.github.io/2025/blog/towards-more-rigorous-llm-evals/>. <https://iclr-blogposts.github.io/2025/blog/towards-more-rigorous-llm-evals/>.

648 Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot,
 649 Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier,
 650 Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas
 651 Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL <https://arxiv.org/abs/2310.06825>.

652

653 Joshua Kazdan, Rylan Schaeffer, Apratim Dey, Matthias Gerstgrasser, Rafael Rafailev, David L.
 654 Donoho, and Sanmi Koyejo. Collapse or thrive? perils and promises of synthetic data in a self-
 655 generating world. *arXiv preprint arXiv:2410.16713*, 2025. URL <https://arxiv.org/abs/2410.16713>.

656

657

658 Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi,
 659 Noah A. Smith, and Daniel S. Weld. Genie: Toward reproducible and standardized human
 660 evaluation for text generation, 2022. URL <https://arxiv.org/abs/2101.06561>.

661

662 Jaeho Kim, Yunseok Lee, and Seulki Lee. Position: The AI conference peer review crisis de-
 663 mands author feedback and reviewer rewards. In *Forty-second International Conference on*
 664 *Machine Learning Position Paper Track*, 2025. URL <https://openreview.net/forum?id=18QemUZaIA>.

665

666 Andreas Kirsch. Paper review: Bayesian model selection, the marginal like-
 667 lihood, and generalization. <https://blog.blackhc.net/2022/06/bayesian-model-selection-marginal-likehood-generalization/>, June
 668 2022. Accessed on May 21, 2025.

669

670

671 Andreas Kirsch. Important Prior Work Attribution: RhoLoss and Rho-1. Public comment on OpenRe-
 672 view, dec 2024. URL <https://openreview.net/forum?id=0NMzBwqaAJ¬eId=dZcOZBIIVe>. Accessed: 2025-05-22. Comment on OpenReview forum ID 0NMzBwqaAJ, note
 673 ID dZcOZBIIVe.

674

675 Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian
 676 Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language
 677 models. *arXiv preprint arXiv:2211.09110*, 2022.

678

679 Pratyush Maini and Anshuman Suri. Reassessing emnlp 2024's best paper: Does
 680 divergence-based calibration for membership inference attacks hold up? In *ICLR*
 681 *Blogposts 2025*, 2025. URL <https://iclr-blogposts.github.io/2025/blog/calibrated-mia/>.

682

683

684 Igor L. Markov. The false dawn: Reevaluating google's reinforcement learning for chip macro
 685 placement, 2024. URL <https://arxiv.org/abs/2306.09633>.

686

687 Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher
 688 Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted
 689 question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

690

691 Minh Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-
 692 Ziv. Turning up the heat: Min-p sampling for creative and coherent llm outputs. *arXiv preprint*
arXiv:2407.01082, 2024.

693

694 Francesco Orabona. Yet another ICML award fiasco. [https://parameterfree.com/](https://parameterfree.com/yet-another-icml-award-fiasco/)
 695 yet-another-icml-award-fiasco/, August 2023. Blog post on *Parameter-free Learn-
 696 ing and Optimization Algorithms*.

697

698 Paper Copilot. Neurips statistics. <https://papercopilot.com/statistics/neurips-statistics/>, 2025. Accessed: 2025-09-24.

699

700 Joelle Pineau, Genevieve Fried, Rosemary Nan Ke, and Hugo Larochelle. ICLR
 701 2018 Reproducibility Challenge. <https://www.cs.mcgill.ca/~jpineau/ICLR2018-ReproducibilityChallenge.html>, 2017. Accessed on May 20, 2025.

702 Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan
 703 Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang,
 704 Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin
 705 Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi
 706 Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan,
 707 Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL
 708 <https://arxiv.org/abs/2412.15115>.

709 David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani,
 710 Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark,
 711 2023. URL <https://arxiv.org/abs/2311.12022>.

712 Stephen Roller, Y-Lan Boureau, Jason Weston, Antoine Bordes, Emily Dinan, Angela Fan, David
 713 Gunning, Da Ju, Margaret Li, Spencer Poff, Pratik Ringshia, Kurt Shuster, Eric Michael Smith,
 714 Arthur Szlam, Jack Urbanek, and Mary Williamson. Open-domain conversational agents: Current
 715 progress, open problems, and future directions, 2020. URL <https://arxiv.org/abs/2006.12442>.

716 Rylan Schaeffer, Mikail Khona, and Ila Fiete. No free lunch from deep learning in neuro-
 717 science: A case study through models of the entorhinal-hippocampal circuit. In S. Koyejo,
 718 S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neu-
 719 ral Information Processing Systems*, volume 35, pp. 16052–16067. Curran Associates, Inc.,
 720 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/66808849a9f5d8e2d00dbdc844de6333-Paper-Conference.pdf.

721 Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language
 722 models a mirage? In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.),
 723 *Advances in Neural Information Processing Systems*, volume 36, pp. 55565–55581. Curran Asso-
 724 ciates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/adc98a266f45005c403b8311ca7e8bd7-Paper-Conference.pdf.

725 Rylan Schaeffer, Mikail Khona, Sarthak Chandra, Mitchell Ostrow, Brando Miranda, and Sanmi
 726 Koyejo. Does maximizing neural regression scores teach us about the brain? In *UniReps: 2nd
 727 Edition of the Workshop on Unifying Representations in Neural Models*, 2024. URL <https://openreview.net/forum?id=vbtj05J68r>.

728 Rylan Schaeffer, Joshua Kazdan, Alvan Caleb Arulandu, and Sanmi Koyejo. Position: Model collapse
 729 does not mean what you think, 2025a. URL <https://arxiv.org/abs/2503.03150>.

730 Rylan Schaeffer, Joshua Kazdan, Yegor Denisov-Blanch, Brando Miranda, Matthias Gerstgrasser, Su-
 731 san Zhang, Andreas Haupt, Isha Gupta, Elyas Obbad, Jesse Dodge, Jessica Zosa Forde, Francesco
 732 Orabona, Sanmi Koyejo, and David Donoho. Position: Machine learning conferences should
 733 establish a "refutations and critiques" track, 2025b. URL <https://arxiv.org/abs/2506.19882>.

734 Rylan Schaeffer, Joshua Kazdan, John Hughes, Jordan Juravsky, Sara Price, Aengus Lynch, Erik
 735 Jones, Robert Kirk, Azalia Mirhoseini, and Sanmi Koyejo. How do large language monkeys get
 736 their power (laws)?, 2025c. URL <https://arxiv.org/abs/2502.17578>.

737 Rylan Schaeffer, Punit Singh Koura, Binh Tang, Ranjan Subramanian, Aaditya K Singh, Todor
 738 Mihaylov, Prajwal Bhargava, Lovish Madaan, Niladri S. Chatterji, Vedanuj Goswami, Sergey
 739 Edunov, Diewuke Hupkes, Sanmi Koyejo, and Sharan Narang. Correlating and predicting human
 740 evaluations of language models from natural language processing benchmarks, 2025d. URL
<https://arxiv.org/abs/2502.18339>.

741 Koustuv Sinha, Jessica Zosa Forde, Mandana Samiei, Arna Ghosh, Lintang Sutawika, and
 742 Siba Smarak Panigrahi (eds.). *Machine Learning Reproducibility Challenge 2023 Proceedings*,
 743 2023. ReproML. URL <https://reproml.org/proceedings/mlrc2023/>. Proceed-
 744 ings published in 2024 for the 2023 challenge.

745 Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford,
 746 Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in
 747 neural information processing systems*, 33:3008–3021, 2020.

756 Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya
 757 Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan
 758 Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar,
 759 Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin,
 760 Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur,
 761 Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison,
 762 Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia
 763 Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris
 764 Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger,
 765 Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric
 766 Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary
 767 Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra,
 768 Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha
 769 Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost
 770 van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed,
 771 Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia,
 772 Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago,
 773 Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel
 774 Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow,
 775 Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan,
 776 Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad
 777 Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda,
 778 Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep
 779 Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh
 780 Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien
 781 M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan
 782 Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kociský, Tulsee Doshi,
 783 Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei,
 784 Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei,
 785 Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins,
 786 Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav
 787 Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena
 788 Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi,
 789 and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024. URL
 790 <https://arxiv.org/abs/2408.00118>.

791 Craig Thomson, Ehud Reiter, and Anya Belz. Common flaws in running human evaluation experiments in NLP. *Computational Linguistics*, 50(2):795–805, June 2024. doi: 10.1162/coli_a_00508.
 792 URL <https://aclanthology.org/2024.cl-2.9/>.

793 Chris Van Der Lee, Albert Gatt, Emiel Van Miltenburg, Sander Wubben, and Emiel Krahmer. Best
 794 practices for the human evaluation of automatically generated text. In *Proceedings of the 12th
 795 International Conference on Natural Language Generation*, pp. 355–368, 2019.

796 Yi Xu, Laura Ruis, Tim Rocktäschel, and Robert Kirk. Investigating non-transitivity in llm-as-a-judge,
 797 2025. URL <https://arxiv.org/abs/2502.14074>.

798 Jie Zhang, Christian Schlarbmann, Kristina Nikolić, Nicholas Carlini, Francesco Croce, Matthias Hein,
 799 and Florian Tramèr. Evaluating the robustness of the "ensemble everything everywhere" defense,
 800 2025. URL <https://arxiv.org/abs/2411.14834>.

801 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang,
 802 Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and
 803 chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

804

805

806

807

808

809

810 **A LANGUAGE MODEL USAGE**
811

812 Language models were used by the authors to aid or polish the writing of the paper. Authors take full
813 responsibility for the content.
814

815

816

817

818

819

820

821

822

823

824

825

826

827

828

829

830

831

832

833

834

835

836

837

838

839

840

841

842

843

844

845

846

847

848

849

850

851

852

853

854

855

856

857

858

859

860

861

862

863

864
 865 **B EXAMPLES OF HUMAN QUALITATIVE RESPONSES FAVORING BASIC**
 866 **SAMPLING OVER MIN-P SAMPLING**

867 In Section 2.3, we described how qualitative responses from many human participants in the original
 868 paper’s study favored **basic** sampling. Direct quotes from human evaluators favoring **basic**
 869 sampling are provided below. In the study, **basic** sampling was called “Model A”; for clarity, we
 870 substituted the pseudonyms for the actual sampling methods):

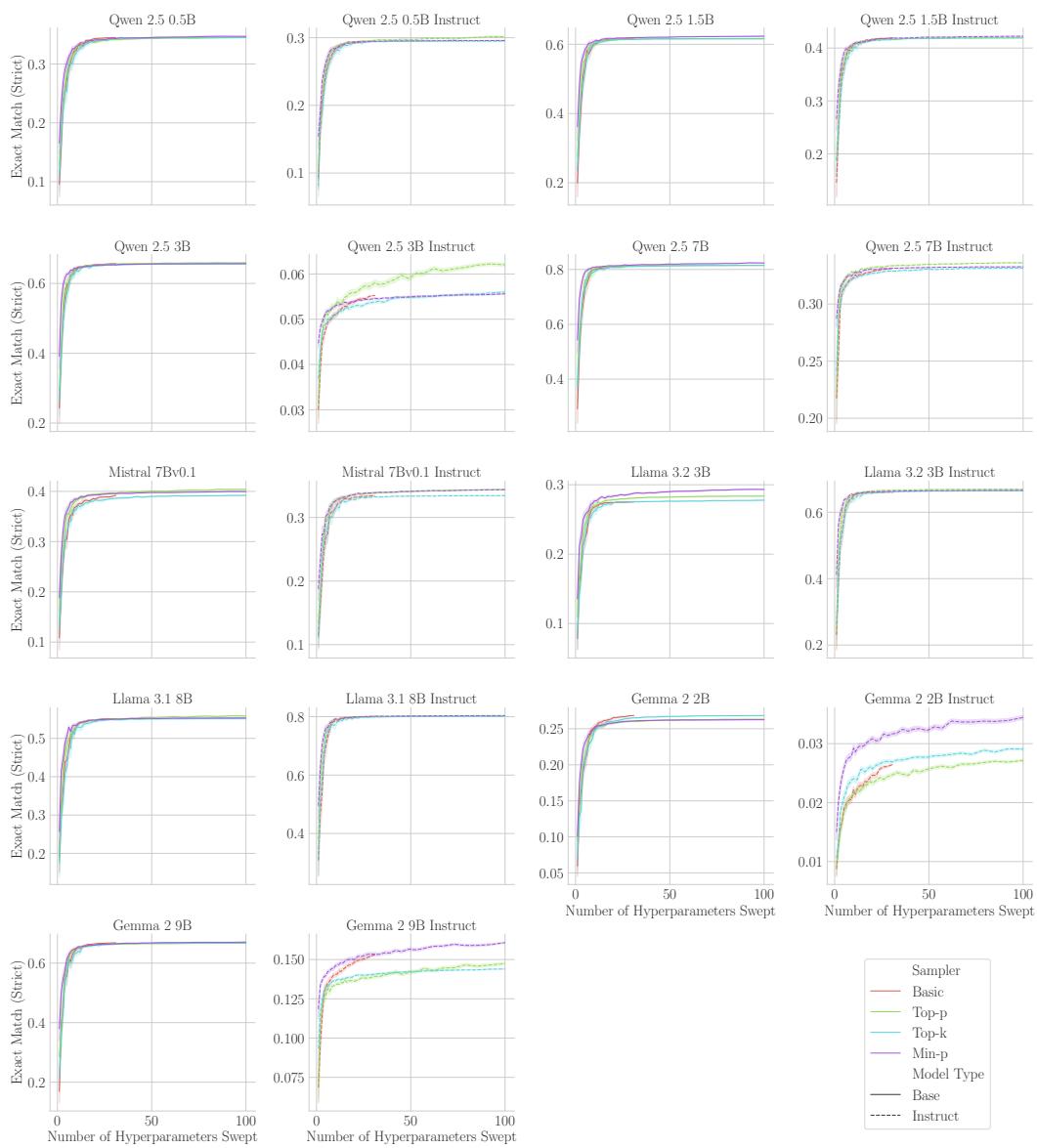
- 872 • “[**basic** sampling] on Temp 3.0 - High Diversity setting. The stories where [sic] more
 873 intereting [sic], felt more different compared to the others, which felt like the same ideia
 874 [sic] just in a different format.”
- 875 • “I felt like [**basic** sampling] was most diverse and most interesting with it’s [sic] descrip-
 876 tions of the characters and the setting. It appealed to me most and seemed to have less
 877 ‘broken’ sentences that didn’t make sense. Descriptions were painterly [sic] and elaborate.”
- 878 • “[**basic** sampling] was more engaging, it aroused my curiosity.”
- 879 • “[**basic** sampling] provided more depth and easy to read for me and there was more
 880 diversity.”
- 881 • “[**basic** sampling], they presented creative storytelling”
- 882 • “[**basic** sampling]. From the very beginning the verbiage and descriptions were very
 883 creative and vivid. And each story was unique”
- 884 • “I believe that [**basic** sampling] has provided stories with more differentiation overall than
 885 the other two models. From the point of view of creativity, all three models are more or less
 886 equivalent as they almost always talk about stories set in extraterrestrial worlds both from a
 887 physical and mental (dreams) point of view”
- 888 • “[**Basic** sampling]: Sample 2: Temperature Setting F (Temp 3.0 - High Diversity). The
 889 story was captivating, it took inside the mystical land and walked you right besides all
 890 the characters, you can even draw the characters from just th descriptions provided by the
 891 prompt. you Could even smell them, smell the setting and be at one with the setting.”
- 892 • “I personally preferred [**basic** sampling] on the setting of creative, descriptive storytelling.
 893 I enjoyed how the writing was creative, showing imagination and a strong use of language.
 894 The stories were quite evocative, with intriguing settings and characters that helped to draw
 895 the reader in. I also appreciated the diversity of themes that were explored, from night
 896 weavers to dream manipulation and mysterious libraries, which kept the stories engaging
 897 and interesting.”
- 898 • “Temporature setting C on [**basic** sampling] was the best. The story was fascinating and
 899 very engaging. I wanted to read more.”
- 900 • “I prefered the first [**basic** sampling]. Tho [**basic** sampling] and C seem to be very head
 901 to head. But something about [**basic** sampling] seemed different in quality about it to me.”

903 More quotes are in the original paper’s data. We urge readers to draw their own conclusions.

905
 906
 907
 908
 909
 910
 911
 912
 913
 914
 915
 916
 917

918 C GSM8K CHAIN-OF-THOUGHT SCORES WITH “STANDARD” FORMATTING
919

920 At the request of Nguyen et al. (2024), we reran our GSM8K Chain-of-Thought sweeps using “stan-
921 dard” formatting instead of “Llama” formatting. **Both analyses reached consistent results: min-p**
922 **does not consistently outperform other samplers when controlling the volume of hyperparame-
923 ter space.**



924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971

Figure 7: **Min-p Does Not Consistently Outperform Other Samplers on GSM8K When Controlling For Hyperparameter Volume.** We reran our GSM8K sweep using “standard” formatting rather than “Llama” formatting and observed qualitatively similar data.

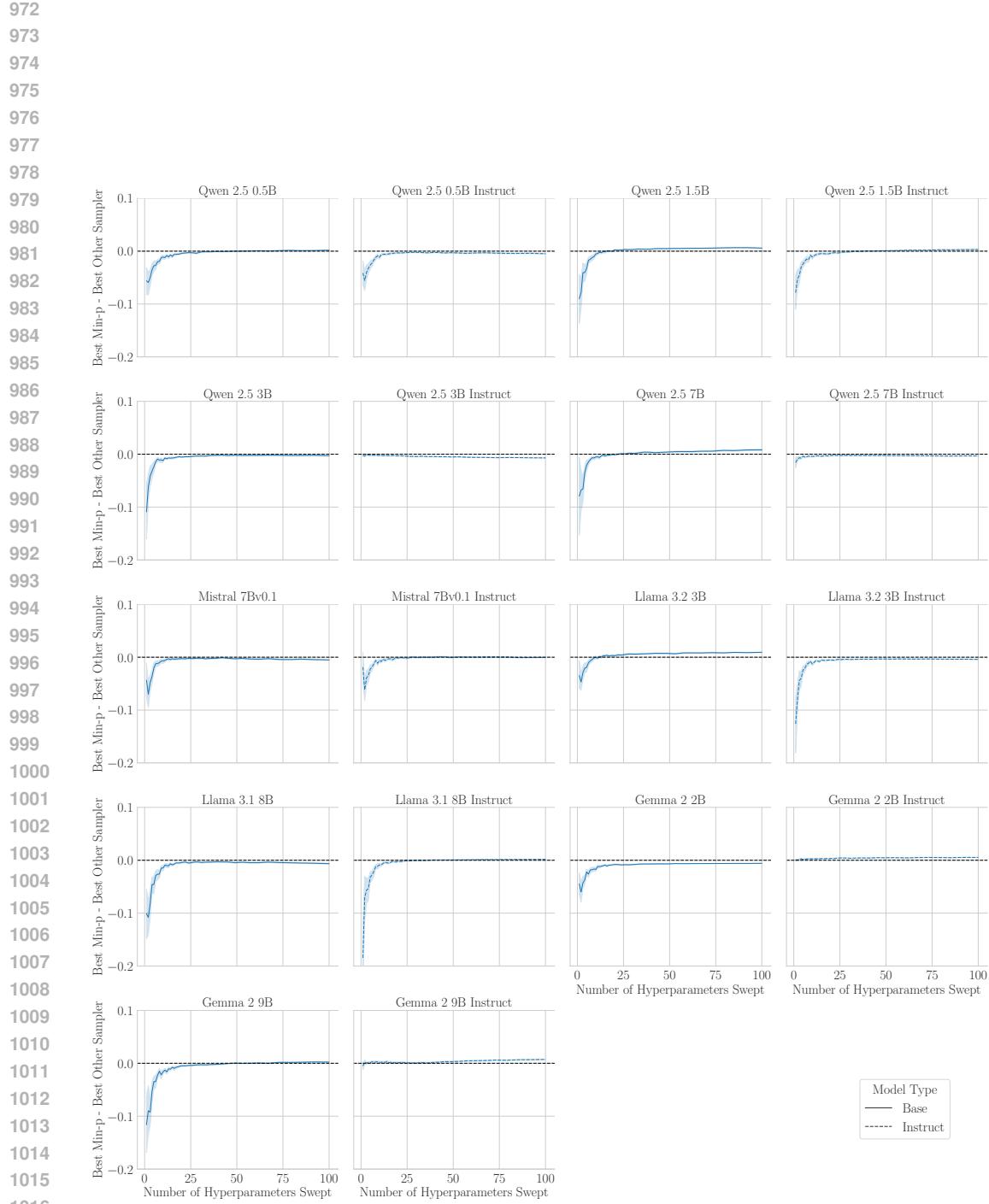
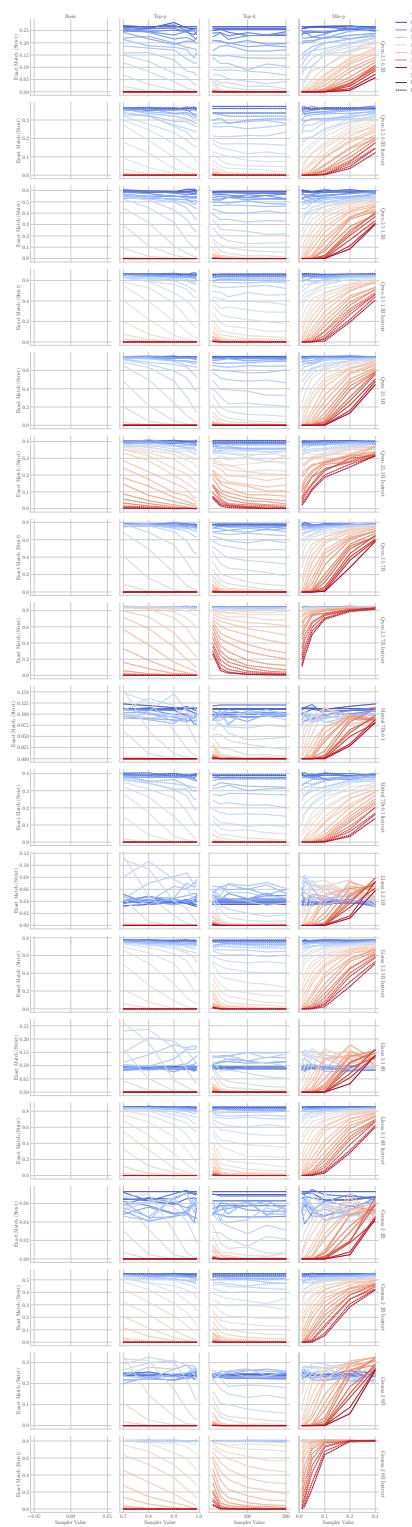


Figure 8: **Min-p Does Not Consistently Outperform Other Samplers on GSM8K When Controlling For Hyperparameter Volume.** We reran our GSM8K sweep using “standard” formatting rather than “Llama” formatting and observed qualitatively similar data.

1026 **D GSM8K SCORES BY MODEL, SAMPLER AND HYPERPARAMETERS**
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
10781079 **Figure 9: GSM8K Scores By Model, Sampler and Sampler Hyperparameters.** Many models
achieve their highest scores at low temperatures across samplers.