RELIABLEEVAL: A Recipe for Stochastic LLM Evaluation via Method of Moments

Gili Lior[1] **Eliya Habba**[1] **Shahar Levy**[1] **Avi Caciularu**[2] **Gabriel Stanovsky**[1]

[1]The Hebrew University of Jerusalem [2]Google Research

gili.lior@mail.huji.ac.il

Abstract

LLMs are highly sensitive to prompt phrasing, yet standard benchmarks typically report performance using a single prompt, raising concerns about the reliability of such evaluations. In this work, we argue for a stochastic method of moments evaluation over the space of meaning-preserving prompt perturbations. We introduce a formal definition of *reliable evaluation* that accounts for prompt sensitivity, and suggest RELIABLEEVAL – a method for estimating the number of prompt resamplings needed to obtain meaningful results. Using our framework, we stochastically evaluate five frontier LLMs and find that even top-performing models like GPT-4o and Claude-3.7-Sonnet exhibit substantial prompt sensitivity. Our approach is model-, task-, and metric-agnostic, offering a recipe for meaningful and robust LLM evaluation.[1]

1 Introduction

A host of recent work has noticed that LLMs are highly sensitive to seemingly arbitrary *prompt perturbations*, throwing into question many of the results reported on popular benchmarks. These perturbations span various dimensions: semantically-equivalent paraphrases of the task instructions (Mizrahi et al., 2024), changes in delimiters or whitespace (Sclar et al., 2024; Voronov et al., 2024), the order of in-context few-shot examples (Lu et al., 2022), among many others (Perlitz et al., 2024; Levy et al., 2024; Liu et al., 2024b).

While these works observed that LLMs are highly sensitive to prompt perturbations, to the best of our knowledge there is currently no prescriptive recipe for conducting meaningful evaluation which takes this sensitivity into account. Evidently, many recent evaluation efforts resort to reporting LLM performance against a single arbitrary prompt, while often acknowledging that this practice is flawed (Gu et al., 2024a,b), highlighting the need for new evaluation practices.

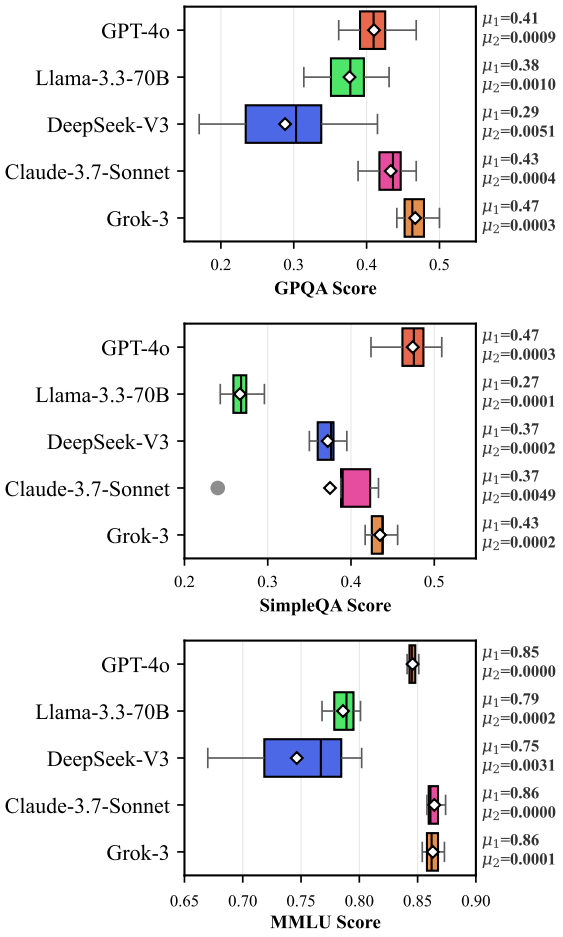


Figure 1: Evaluation of frontier LLMs on multiple meaning-preserving prompt perturbations following RELIABLEEVAL, estimating the complete prompt sample space. Models vary in both expected value and variance, highlighting the importance of stochastic evaluation.

In this work, we argue that the evaluation of such sensitive LLMs requires *stochastic evaluation* over the spectrum of perturbations via a method of moments analysis (expected value, variance, etc.). To estimate moments over the combinatorially large

[1]Code and data available at https://github.com/SLAB-NLP/Reliable-Eval

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 11146–11153

perturbation sample space, we define the notion of *reliable evaluation*, which bounds the probability that a sample of prompt perturbations is representative of the entire sample space. Further, we formulate RELIABLEEVAL – a simple recipe for estimating the number of samples needed to achieve reliable evaluation per dataset.

Using our recipe, we perform stochastic evaluation of five frontier models, as well as leading open-source models, on three popular benchmarks. Our findings, shown in Figure 1, reveal the statistical differences between models, highlighting the need for stochastic evaluation. Moreover, we show that the number of resamplings required to reliably estimate model performance varies depending on both the model and the dataset being evaluated.

We hope that our recommendations will be adopted to achieve meaningful and reliable reporting of LLM performance.

2 Stochastic Evaluation of LLMs: Desiderata and Approximation

Here we propose a set of desired metrics for LLM evaluation in light of their observed sensitivity (§2.1). Since computing these metrics directly is infeasible, we also describe the desired statistical properties of a reliable approximation (§2.2). In the following sections, we will operationalize these concepts (§3), and use this approach to evaluate frontier LLMs (§4).

2.1 Characterizing LLM Performance Using Distributional Analysis

We formulate the behavior a model M as a random variable with respect to a deterministic evaluation metric ε:

$$\varepsilon_M : S_D \mapsto \mathbb{R}_+ \tag{1}$$

Where D denotes an evaluation dataset (e.g., MMLU), the sample space S_D denotes the space of all meaning-preserving prompt perturbations of D (e.g., different instruction paraphrases, different answer enumerators, addition or removal of whitespace), and $\varepsilon_M(s)$ denotes the performance of model M on a single prompt $s \in S_D$ according to metric ε. For example, $\varepsilon_M(s) \in [0, 1]$ can denote the exact-match accuracy of Llama (M) on a single MMLU instance under prompt s. Using this notation, the limitations of current evaluations are evident – they report the values of ε_M on arbitrary samples from S_D, while aiming to make claims about the entire sample space S_D.

A statistically-meaningful evaluation of LLMs. This stochastic formulation of LLM performance gives rise to a *method of moments* analysis of its behavior (Casella and Berger, 2024). In particular, we treat $s \in S_D$, as i.i.d. resulting from uniform sampling over S_D. I.e., since we focus on meaning-preserving prompt perturbations, they are considered to be equally likely. We further focus on the first and second moments of ε_M.

The first moment μ_1 denotes the model's *expected value* over the space of all meaning-preserving prompt perturbations:

$$\begin{aligned}\mu_1(M, S_D) &= \mathop{\mathbb{E}}_{s \overset{\text{i.i.d.}}{\sim} S_D}[\varepsilon_M] \\ &= \sum_{s \in S_D} \varepsilon_M(s) \cdot P(S = s) \\ &\overset{\text{uniform i.i.d.}}{=} \frac{1}{|S_D|} \sum_{s \in S_D} \varepsilon_M(s)\end{aligned} \tag{2}$$

Similarly, the second moment μ_2, i.e., *variance*, is given by:

$$\begin{aligned}\mu_2(M, S_D) &= \mathbb{E}_{s \overset{\text{i.i.d.}}{\sim} S_D}[{\varepsilon_M}^2] \\ &= \mathbb{E}_{s \overset{\text{i.i.d.}}{\sim} S_D}[(\varepsilon_M(s) - \mu_1)^2] \\ &\overset{\text{uniform i.i.d.}}{=} \frac{1}{|S_D|} \sum_{s \in S_D} (\varepsilon_M(s) - \mu_1)^2\end{aligned} \tag{3}$$

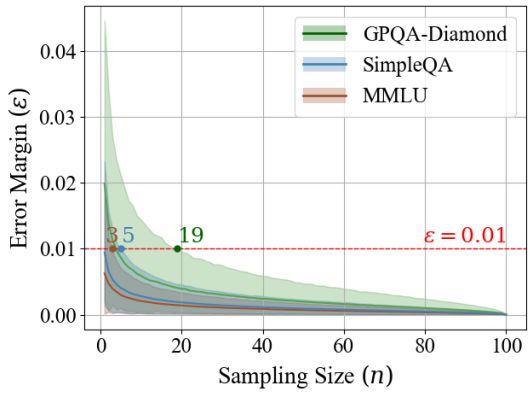
This framework allows future work to extend the analysis to additional moments and other distributions beyond uniform i.i.d (Siska et al., 2024).

2.2 Reliable Estimation of Distributional Analysis

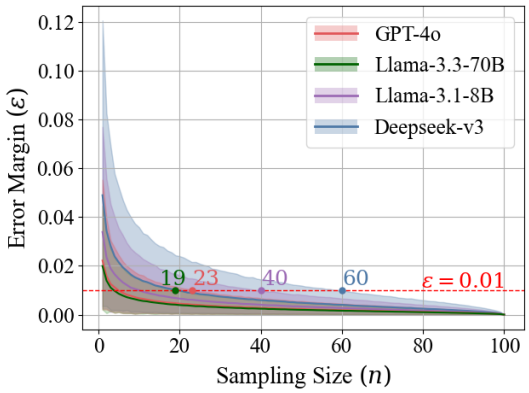
Note that explicitly computing the moments in Equations 2 and 3 is infeasible, as it requires knowing the entire space of meaning-preserving prompt perturbations, which explodes combinatorially (e.g., for all of the permutations of few shot examples) and is even hard to enumerate (e.g., such is the case for the space of all instruction paraphrases). Instead, we aim to estimate these moments using a random sample $S' \subset S_D$, relying on the linearity of expectations, as is similarly done in stochastic gradient descent.

Below we define dataset-specific requirements to make sure that S' is large enough to enable reliable estimation of the true moments.

Definition 1 (Reliable evaluation)**.** Given an error margin ϵ and confidence level δ, let S_D be the space



(a) Llama-3.3-70B convergence on different benchmarks.



(b) Different models convergence on GPQA-Diamond.

Figure 2: Convergence of the deviation from the true mean accuracy with increasing resampling size. Round markers indicate n^*, the min. resamplings as defined in Eq. 6, shown per benchmark in (a) and per model in (b).

of all meaning-preserving prompt perturbations of dataset D, and let $S' \subset S_D$ be a random subset of size n. Then, we say that n samples yield a *reliable evaluation* if for every moment μ_i (expected value and variance), it holds that:

$$\mathop{\mathbf{P}}_{\substack{S' \subset S_D \\ |S'|=n}} \left[\left| \mu_i(M, S') - \mu_i(M, S_D) \right| > \epsilon \right] < \delta \quad (4)$$

In other words, an evaluation based on n resamplings of $S' \subset S_D$ with $|S'| = n$ is considered *reliable* if the probability that the empirical momentum of the sample S' deviates from the momentum over the entire distribution by more than ϵ is bounded by δ. In section 3 we propose a method for estimating the required n, by constructing a confidence interval around this deviation.

We can then perform stochastic evaluation over this reduced resampling space, reporting empirical moments which are expected to yield with high probability a good estimation of the true moments over the entire sample space.

3 RELIABLEEVAL: Recipe for Stochastic Evaluation

In this section, we present a practical recipe for conducting a reliable stochastic evaluation of LLMs.

The recipe assumes a scenario aiming to evaluate a set of models $M_1, \ldots, M_k$ on a dataset D, while accounting for LLMs' sensitivity to meaning-preserving prompt perturbations.

Step 1: Specify evaluation parameters ϵ and δ. Set the acceptable deviation ϵ between the empirical value of the i-th moment over a sample $S' \subset S_D$ and the corresponding moment over the full distribution S_D, as well as the confidence level δ with which this guarantee should hold, as defined in Equation 4. In particular, we propose to set $\epsilon = 0.01$ and $\delta = 0.1$, i.e., that evaluation should be considered reliable if it deviates from true distribution by no more than 0.01 with probability of at least 0.9. This can critically examine claims of state of the art performance, which typically revolve around a difference of a few performance points between models (Liu et al., 2024a).

Step 2: Define the sample space of meaning-preserving paraphrases S_D. Identify dimensions of meaning-preserving prompt perturbations that may influence model performance – such as instruction phrasing, output format, or few-shot examples. We recommend leveraging existing work aligned with the task type. For instance, for multiple-choice QA datasets, the framework by Habba et al. (2025) can be used to generate the prompt perturbation space S_D. Their approach builds on the Unitxt framework for structured data preparation, which can also be extended to generate prompt perturbations for other task types (Bandel et al., 2024). Notably, our proposed method is flexible and not restricted to any predefined set of meaning-preserving prompt perturbations, and other paraphrases can be used to construct the sample space S_D.

Step 3: Estimate the minimal reliable sample size n^*. Our goal here is to identify the smallest sample size n which satisfies the reliability condition in Definition 1. This is challenging since it requires computing true moments over the entire distribution. To estimate this, we propose to choose

a reference model $\hat{M}$ and compute its empirical moments over large N as proxy for true moments. In the following section, we will show that choosing a relatively cheap model gives empirically good estimates, which hold across models. For each candidate sample size $n = 1, 2, \ldots, N$, compute the set of deviations between the empirical value of the i-th moment over each subset $S' \subset S_D$ of size n, and the i-th moment computed over N samples:

$$\Delta(n) = \left\{ \left| \mu_i(M, S') - \mu_i(M, S_D) \right| : |S'| = n \right\} \tag{5}$$

Next, construct the δ-level confidence interval (CI) over $\Delta(n)$, which filters $\Delta(n)$ to the range between the $\delta/2$ and $1 - \delta/2$ percentiles. For instance, if $\delta = 0.1$, the corresponding $\mathbf{CI}_{0.1}(\Delta(n))$ includes all values of $\Delta(n)$ which lie between the 5th and 95th percentiles. Then, define n^* as the smallest n for which ϵ is larger than the maximum of this confidence interval:

$$n^* = \min \{ n \in [1, N] \mid \epsilon \geq \max \mathbf{CI}_\delta(\Delta(n)) \} \tag{6}$$

We note that in some scenarios, such as when the focus is on evaluating a single model or when the variations between models is large, it may be preferable to use a reference dataset instead of a reference model. For example, if we want to evaluate model M on multiple datasets, we can choose a reference dataset D', compute its empirical moments over large N as a proxy for the true moments.

Step 4: Report empirical distribution analysis. Finally, sample a subset of perturbations $S' \subset S_D$ of size $|S'| = n^*$ uniformly at random. Then, evaluate each model $M_1, \ldots, M_k$ on all prompt variations $s \in S'$, and report empirical moment analysis. In particular, we recommend reporting box plot showing median and interquartile range of observed performance, as can be seen in Figure 1.

4 Reliable Stochastic Evaluation of Frontier Models

In this section, we present a reliable stochastic evaluation of five state-of-the-art LLMs, including both open-source and proprietary models, across three widely used benchmarks.

4.1 Experimental Setup

We run RELIABLEEVAL on MMLU (Hendrycks et al., 2021), GPQA-Diamond (Rein et al., 2024), and SimpleQA (Wei et al., 2024), which are all widely-used English benchmarks. The curation of the meaning-preserving prompt perturbations space is done by leveraging unitxt (Bandel et al., 2024) and Dove (Habba et al., 2025).We evaluate five LLMs: Llama-3.3-70B (Grattafiori et al., 2024), Deepseek-v3 (Liu et al., 2024a), GPT-4o (Hurst et al., 2024), Claude-3.7-Sonnet (Anthropic, 2025), and Grok-3 (xAI, 2025). As defined in Section 2, we set the following parameters to estimate a reliable evaluation $\epsilon = 0.01$, $\delta = 0.1$, $N = 100$, with Llama-3.3-70B serving as the reference model $\hat{M}$ for estimating n^*. See additional implementation details in the Appendix.

4.2 Results

Frontier models are sensitive to meaning-preserving prompt perturbations, underscoring the need for stochastic evaluation. Figure 1 shows that across all three evaluated benchmarks, model performance varies across different prompt resamplings. This highlights the importance of stochastic evaluation, i.e., reporting statistical measures over the distribution of scores rather than relying on single prompts. As shown by the overlapping boxplots in Figure 1, there is *often no definitive winner* – any meaning-preserving prompt could be cherry-picked to suggest a particular model ranking.

The number of resamplings required for reliable evaluation depends both on the dataset and on the model. In Figure 2a, we show that the convergence behavior of Llama-3.3-70B's estimation depends on the benchmark. Moreover, in Figure 2b, we observe that different models exhibit different convergence rates on the same dataset, suggesting that reliable evaluation is determined by both the model and the dataset.

Llama-3.1-8B can guide the number of resamplings needed for reliable evaluation of Llama-3.3-70B. While Llama-3.3-70B substantially outperforms the smaller Llama-3.1-8B, Figure 2b shows that the smaller model provides a valid upper bound on convergence behavior. This suggests that smaller models can serve as effective proxies for estimating the number of prompt resamplings required for reliable stochastic evaluation of larger models. This is shown also for the GPQA-Diamond and SimpleQA (Figure 3 in Appendix).

5 Related Work

Most related to our work, Polo et al. (2024) proposed a method for multi-prompt evaluation, hinging on a binary Bernoulli distribution, limiting its applicability to text generation, and revolving around the selection of representative evaluation examples. In contrast, we find a minimal representative random subspace, are agnostic to the type of perturbations, and do not make any assumption about the scoring function. Other works highlight the importance of multi-prompt evaluations , albeit without prescriptive guidelines (Voronov et al., 2024; Tam et al., 2024; Zhuo et al., 2024; Hida et al., 2024).

6 Conclusion

We propose to estimate model performance over prompt variations using moment analysis and show how to compute how many samples are needed for reliable results.

Our proposed method is designed to accommodate any computational budget, with an inherent trade-off between budget, error margin, and confidence. The practical question becomes: *"Given a specific compute budget, what is the most reliable evaluation achievable?"* In our framework, the compute budget sets the maximum feasible N and constrains n^*. If for a given error margin ϵ and confidence level δ the number of samples $n*$ exceeds a given budget, it is still possible run the evaluation with $n < n^*$ resamplings, accepting a larger margin of error or lower confidence as a result. Thus, even with limited compute resources, our method provides guidance on how to maximize evaluation reliability within those constraints.

Finally, by evaluating frontier models across benchmarks, we find that sensitivity varies widely, underscoring the need for more robust evaluation practices.

Limitations

We identify several limitations of this work that future research may address.

First, RELIABLEEVAL requires running a reference model $\hat{M}$ over a large number of resamplings N. While this is performed only once, it can be computationally expensive—especially in LLM-based evaluation settings where the reference model also serves as a judge and is costly to query.

Second, there are two additional factors that may influence the required resampling size, which we did not directly investigate. Future work may explore: (1) the effect of dataset size on the number of resamplings needed, and (2) the impact of the model's decoding strategy, which is known to affect evaluation outcomes (Song et al., 2025). For the latter, we provide an initial comparison in Figure 4, showing results for GPT-4o using greedy decoding versus sampling with a default temperature. However, further experimentation is needed to better understand these effects.

Acknowledgments

This work was partially supported by research grant no. 7256 from the Israeli Ministry of Science and Technology. We thank Dr. Arie Cattan and Dr. Ori Shapira for the helpful discussions and advice on this project.

References

Andrei Alexandru, Antonia Calvi, Henry Broomfield, Jackson Golden, Kyle Dai, Mathias Leys, Maurice Burger, Max Bartolo, Roman Engeler, Sashank Pisupati, and 1 others. 2025. Atla selene mini: A general purpose evaluation model. *arXiv preprint arXiv:2501.17195*.

Anthropic. 2025. Claude 3.7 sonnet. https://www.anthropic.com/claude/sonnet.

Elron Bandel, Yotam Perlitz, Elad Venezian, Roni Friedman, Ofir Arviv, Matan Orbach, Shachar Don-Yehiya, Dafna Sheinwald, Ariel Gera, Leshem Choshen, Michal Shmueli-Scheuer, and Yoav Katz. 2024. Unitxt: Flexible, shareable and reusable data preparation and evaluation for generative AI. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pages 207–215, Mexico City, Mexico. Association for Computational Linguistics.

George Casella and Roger Berger. 2024. *Statistical inference*. CRC press.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Alex Gu, Wen-Ding Li, Naman Jain, Theo Olausson, Celine Lee, Koushik Sen, and Armando Solar-Lezama. 2024a. The counterfeit conundrum: Can code language models grasp the nuances of their incorrect generations? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 74–117, Bangkok, Thailand. Association for Computational Linguistics.

Alex Gu, Baptiste Rozière, Hugh Leather, Armando Solar-Lezama, Gabriel Synnaeve, and Sida I. Wang. 2024b. Cruxeval: a benchmark for code reasoning, understanding and execution. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Eliya Habba, Ofir Arviv, Itay Itzhak, Yotam Perlitz, Elron Bandel, Leshem Choshen, Michal Shmueli-Scheuer, and Gabriel Stanovsky. 2025. Dove: A large-scale multi-dimensional predictions dataset towards meaningful llm evaluation. *arXiv preprint arXiv:2503.01622*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Rem Hida, Masahiro Kaneko, and Naoaki Okazaki. 2024. Social bias evaluation for large language models requires prompt variations. *arXiv preprint arXiv:2407.03129*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353, Bangkok, Thailand. Association for Computational Linguistics.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024b. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein-Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. 2024. Efficient benchmarking (of language models). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2519–2536, Mexico City, Mexico. Association for Computational Linguistics.

Felipe Maia Polo, Ronald Xu, Lucas Weber, Mírian Silva, Onkar Bhardwaj, Leshem Choshen, Allysson Flavio Melo de Oliveira, Yuekai Sun, and Mikhail Yurochkin. 2024. Efficient multi-prompt evaluation of LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Charlotte Siska, Katerina Marazopoulou, Melissa Ailem, and James Bono. 2024. Examining the robustness of LLM evaluation to the distributional assumptions of benchmarks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10406–10421, Bangkok, Thailand. Association for Computational Linguistics.

Yifan Song, Guoyin Wang, Sujian Li, and Bill Yuchen Lin. 2025. The good, the bad, and the greedy: Evaluation of LLMs should not ignore non-determinism. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4195–4206, Albuquerque, New Mexico. Association for Computational Linguistics.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. 2024. Let me speak freely? a study on the impact of format restrictions on large language model performance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1218–1236, Miami, Florida, US. Association for Computational Linguistics.

Anton Voronov, Lena Wolf, and Max Ryabinin. 2024. Mind your format: Towards consistent evaluation of in-context learning improvements. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6287–6310, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*.

xAI. 2025. Grok 3 beta — the age of reasoning agents. https://x.ai/news/grok-3.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. ProSA: Assessing and understanding the prompt sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1950–1976, Miami, Florida, USA. Association for Computational Linguistics.

A Appendix

A.1 Benchmarks and Prompt Perturbations

We provide additional details about the benchmarks used in our evaluation with RELIABLEEVAL.

Prompt Perturbation Dimensions. For each benchmark, we define task-specific dimensions of prompt perturbations over which we resample.

For MMLU and GPQA-Diamond (Multiple-Choice QA), we follow the resampling strategy from (Habba et al., 2025), varying along four dimensions: (1) instruction paraphrasing, (2) answer choice order, (3) answer choice enumerator (e.g., letters, numbers, Roman numerals), and (4) choice separators (e.g., whitespace, tab, newline) between the answers.

For SimpleQA (Open-Ended QA), we vary: (1) instruction phrasing (e.g., "Answer the following question"), (2) which examples are selected for evaluation, (3) the selection and ordering of few-shot demonstrations, and (4) whether prompts include 'Question:' and 'Answer:' markers.

Number of Examples Per Benchmark. For GPQA-Diamond, we evaluate the full dataset, with 198 examples per resampling. For MMLU, we sample 100 examples from each subcategory, resulting in 5,700 total examples (from the 14K test split), reused across all resamplings. For SimpleQA, which includes variation in the evaluation examples themselves, we randomly select 1K examples (from 4K) per resampling, ensuring full coverage over multiple runs.

Prompting Technique. We use 5-shot prompting for all benchmarks during evaluation.

A.2 Evaluation Setup

LLM-as-a-Judge for SimpleQA. To evaluate SimpleQA, we use an LLM-as-a-judge setup to determine alignment between predictions and gold answers. We adopt the judging prompt from the official SimpleQA repository.[2] Our judge model is Atla Selene Mini (Alexandru et al., 2025), which currently ranks highest among open-source models on the Judge Arena Leaderboard.[3]

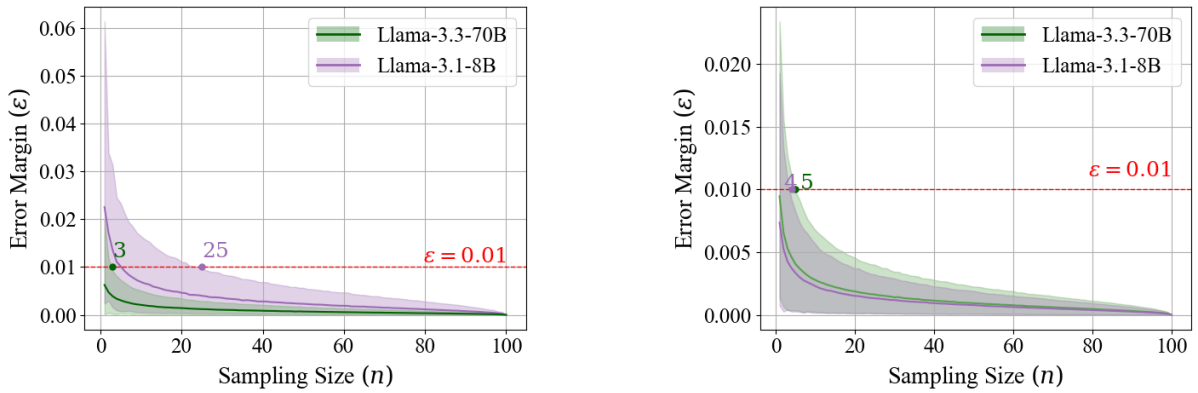
Model Decoding Temperatures. To match typical usage, we adopt model-specific decoding temperatures aligned with standard evaluation practices, informed by official documentation and community reports.

[2] https://github.com/openai/simple-evals

[3] https://huggingface.co/spaces/AtlaAI/judge-arena

Model	Version	Decoding Temp.	Inference Platform/API	Total Cost ($)
GPT-4o	gpt-4o-2024-08-06	1.0	OpenAI	100
Llama-3.3-70B	Llama-3.3-70B-Instruct-Turbo	0.0	Together AI	420
Deepseek-v3	DeepSeek-V3	0.3	Together AI	60
Grok-3	grok-3	0.1	XAI	60
Claude-3.7-Sonnet	claude-3-7-sonnet-20250219	0.0	Anthropic	60
Llama-3.1-8B	meta-llama/Llama-3.1-8B-Instruct	0.0	vLLM on local a6000 (1)	N/A

Table 1: Model inference configurations.



(a) MMLU

(b) SimpleQA

Figure 3: Error convergence of Llama-3.3-70B vs Llama-3.1-8B.

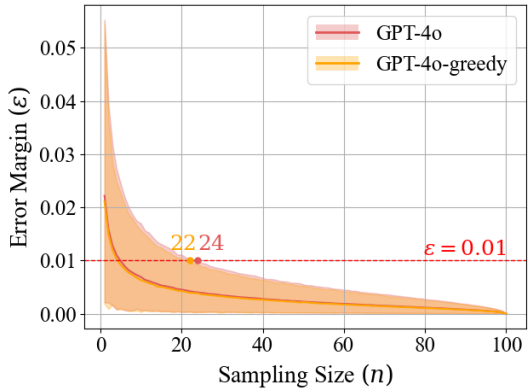


Figure 4: GPT-4o's error convergence on GPQA-Diamond, greedy decoding versus default temperature sampling (temp=1).