The Fault in Our LLM Leaderboards

Pegah Jandaghi and **Kian Ahrabian**
University of Southern California
{jandaghi,ahrabian}@usc.edu

Abstract

The rapid development of large language models (LLMs) has led to the creation of numerous benchmarks and leaderboards, assessing models' performance and ultimately guiding model selection. A key underlying assumption for model selection based on these benchmarks is that their measured performance is transferable for an LLM. More specifically, we expect similar tasks generated from different source distributions to exhibit similar rankings on a given set of LLMs. This work critically examines this assumption by evaluating the transferability of LLMs' ranking on common leaderboards to unseen target tasks. To this end, we systematically analyze the correlation between benchmark-based rankings and actual performance rankings on diverse target tasks, highlighting discrepancies that challenge the reliability of using the former for model selection. Our results reveal that benchmark-based rankings, at best, moderately correlate with real-world performance, with correlation values often falling below 0.5.

1 Introduction

Recent advancements in large language models (LLMs) have resulted in their wide adoption across fields and expertise [Wei *et al.*, 2022]. However, keeping up with the rapid release of new LLMs has become exceedingly challenging due to the significant cost of exploring the wide range of models [Zhang *et al.*, 2023]. Moreover, other factors such as compute resources, expert LLM knowledge, etc., prevent practitioners from finding and utilizing the best model for their novel task. To address these limitations, LLM researchers have created benchmarks to compare the performance and capabilities of LLMs, aiding users in model ranking and selection [Zhang *et al.*, 2024; Hendrycks *et al.*, 2021]. These evaluations have resulted in the creation of general-purpose leaderboards (*e.g.,* HELM, AlpacaEval, etc.). However, a common underlying assumption of these benchmarks is that their measured performance is indicative of an LLM's broader capabilities and generalizes well to other similar real-world tasks. While efforts have been made to design more efficient and comprehensive benchmarks [Polo *et al.*, 2024], the extent to which

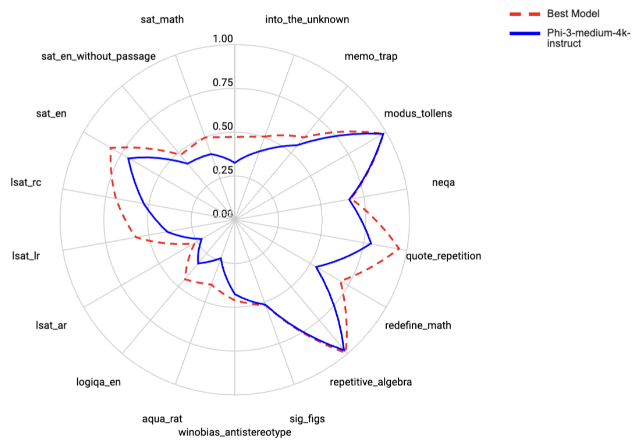

Figure 1: Comparison between the performance of the leaderboard's top-ranked model and the best-performing model on the target task.

benchmark rankings generalize to other similar tasks remains an open question [Saxon *et al.*, 2024].

Recently, [Mahowald *et al.*, 2024] investigated the segregation of language and thought in LLMs. They categorized the linguistic capabilities of LLMs, such as understanding linguistic rules and patterns, as *formal competence*. In parallel, they categorized the capability of understanding and using language in the real world as *functional competence*. Following their work, we classify and distinguish a new task as either a formally out-of-domain task or functionally out-of-domain. In this work, we focus on formally out-of-domain tasks. i.e., tasks that look different on the surface level (linguistic level) but require similar functional skills as our in-domain tasks.

Previous studies have explored shortcut learning in LLMs, particularly their sensitivity to input formats [Alzahrani *et al.*, 2024]. However, shortcut learning in LLMs remains underexplored, especially in the context of benchmark-driven evaluation and model ranking. To examine this, we analyze public benchmarks, specifically those used in leaderboards, as they serve as a primary data source for evaluating LLM capabilities. Given LLM publishers' incentive to optimize for these benchmarks, we consider benchmark performance as the key signal for assessing generalizability. We investigate the reliability of this signal by evaluating how well it transfers to novel and out-of-domain tasks. Our results reveal

Source	Target Task	Kendall-τ	Pearson Correlation	Spearman Correlation
AGIEval	AQUA_RAT	0.270	0.402	0.349
	LOGIQA_EN	0.192	0.217	0.266
	LSAT_AR	0.153	0.145	0.175
	LSAT_LR	0.118	0.147	0.150
	LSAT_RC	0.277	0.302	0.449
	SAT_EN	0.307	0.202	0.418
	SAT_EN_WITHOUT_PASSAGE	0.065	-0.005	0.082
	SAT_MATH	0.328	0.450	0.433
Inverse Scaling	INTO_THE_UNKNOWN	0.364	0.528	0.542
	MEMO_TRAP	0.179	0.264	0.252
	MODUS_TOLLENS	0.412	0.535	0.557
	NEQA	0.430	0.547	0.537
	QUOTE_REPETITION	0.141	-0.008	0.155
	REDEFINE_MATH	0.138	0.067	0.144
	REPETITIVE_ALGEBRA	0.237	0.245	0.388
	SIG_FIGS	0.488	0.696	0.642
	WINOBIAS_ANTISTEREOTYPE	0.169	0.298	0.226

Table 1: Ranking correlation between leaderboard rankings and actual LLM performance on target tasks ranking. For all target tasks, p-value < 0.05 except LSAT_AR.

Source	Model Name	# Params (Billion)	Open LLM Leaderboard Score
01-AI	Yi-1.5-9B-Chat	9	27.71
ARCEE-AI	Arcee-Spark	7	25.54
ARGILLA	notus-7b-v1	7	18.41
Berkeley-nest	Starling-LM-7B-alpha	7	20.64
Deci	DeciLM-7B-instruct	7	17.46
Cognitivecomputations	dolphin-2.9.2-Phi-3-Medium	3.8	25.66
Google	gemma-1.1-7b-it	7	17.48
Gradientai	Llama-3-8B-Instruct-Gradient-1048k	8	18.25
GritLM	GritLM-7B	7	19.15
HuggingFace	zephyr-7b-alpha	7	18.57
	zephyr-7b-beta	7	17.77
IBM	merlinite-7b	7	16.76
Meta	Meta-Llama-3.1-8B-Instruct	8	27.91
	Meta-Llama-3-8B-Instruct	8	20.48
Microsoft	Phi-3-medium-4k-instruct	14	32.67
	Phi-3-mini-4k-instruct	3.8	27.2
Mistral	Mistral-7B-Instruct-v0.2	7	18.46
	Mistral-Nemo-Instruct-2407	7	23.53
	Mistral-7B-Instruct-v0.3	7	19.17
NousResearch	Nous-Hermes-2-SOLAR-10.7B	10.7	23.32
	Hermes-2-Pro-Mistral-7B	7	21.64
	Hermes-2-Pro-Llama-3-8B	8	21.63
Nvidia	Mistral-NeMo-Minitron-8B-Base	8	17.66
OpenBuddy	openbuddy-llama3.1-8b-v22.2-131k	8	24.07
Openchat	openchat-3.5-1210	7	22.56
Open-Orca	Mistral-7B-OpenOrca	7	17.7
Qwen	Qwen2-7B-Instruct	7	24.9
	Qwen1.5-7B-Chat	7	16.58
Refuelai	Llama-3-Refueled	8	22.73
Upstage	SOLAR-10.7B-Instruct-v1.0	10.7	19.63

Table 2: LLM pool ($\mathcal{L}$) in our experiments.

significant discrepancies when transferring the performances across tasks, highlighting potential issues such as shortcut learning [Geirhos *et al.*, 2020].

Moreover, we analyze benchmark signals at both micro and macro levels (*i.e.,* leaderboard, benchmarks, and benchmark subtasks), exposing the fragility of skill-based claims over LLMs [Didolkar *et al.*, 2024]. Many benchmarks, such as those designed for mathematical reasoning [Hendrycks *et al.*, 2021], assume a strong correlation with specific competencies. We critically examine these assumptions and assess the extent to which benchmark-derived rankings truly reflect the broader capabilities of LLMs. Our results highlight shortcut learning in LLMs, extending beyond format and style-based sensitivities. These findings provide valuable insights for researchers aiming to improve benchmark design and develop more reliable methods for evaluating LLM capabilities.

2 Problem Definition

In the following, we define the LLMs' ranking generalizability problem. Let $\mathcal{L} = \{\phi_i\}_{i=1}^{n}$ be a pool of LLMs and $\mathcal{B} = \{\beta_j\}_{j=1}^{m}$ denote a set of evaluation benchmarks that are designed to capture LLMs' capabilities. Moreover, let $\mathcal{E}$ be an evaluation metric and $\mathcal{T}$ be a target task, which is a set of

Benchmark	# Samples	# Subtasks
BBH [Suzgun *et al.*, 2022]	5761	24
GPQA [Rein *et al.*, 2023]	1192	3
MATH [Hendrycks *et al.*, 2021]	1324	7
MUSR [Sprague *et al.*, 2024]	756	3
MMLU_PRO [Wang *et al.*, 2024]	12032	1

Table 3: Benchmarks ($\mathcal{B}$) in our experiment.

samples with labels from a specific label space. Our objective is to investigate whether the ranking of LLMs based on their performance on these benchmarks correlates with their generalizability. In other words, we examine whether the relative performance of LLMs on benchmarks is predictive of their relative performance on a new target task $\mathcal{T}$. Let R, be the ranking of LLMs in $\mathcal{L}$ based on their performance on Benchmarks $\mathcal{B}$:

$$R = (\phi_{f(1)}, ..., \phi_{f(n)}) \quad \text{such that} \tag{1}$$

$$\mathcal{E}(\mathcal{B}, \phi_{f(i)})) > \mathcal{E}(\mathcal{B}, \phi_{f(i+1)}) \quad \text{for all } i < n \tag{2}$$

We define the LLMs' ranking generalizability as whether R aligns with the model performance on the target task $\mathcal{T}$, i.e.

$$\mathcal{E}(\mathcal{T}, \phi_{f(i)})) > \mathcal{E}(\mathcal{T}, \phi_{f(i+1)}) \text{ for all } i < n \tag{3}$$

3 Experiment and Analysis

In this section, we evaluate the reliability of benchmarks and leaderboards in ranking LLMs for a set of formally out-of-domain target tasks. To this end, we use a set of popular public benchmarks and an LLM pool, which are available on the Huggingface platform. Our experiment setup is as follows:

LLM pool ($\mathcal{L}$). We select 30 open LLMs from the Huggingface platform for our experiments. These models are chosen based on their performance indicated by the score on the Open LLM leaderboard (as of Nov 25). To ensure a fair comparison and account for resource limitations, we focused on models within a similar range of parameter sizes, specifically those with fewer than 11 billion parameters. Refer to Table 2 for more information on the LLMs in our pool.

Benchmarks ($\mathcal{B}$). We select 5 benchmarks used in the Open LLM Leaderboard: `bbh`, `mmlu_pro`, `gpqa`, `math`, `musr`. These benchmarks encompass a diverse range of tasks and include a total of 21606 samples and 38 subtasks. Table 3 shows the statistics of these benchmarks.

Formally Out-of-Domain Target Tasks ($\mathcal{T}$). We use tasks from the Inverse Scaling Prize [McKenzie *et al.*, 2023], AGIEval benchmarks [Zhong *et al.*, 2023] which were introduced as challenging tasks for LLMs. AGIEval is a benchmark to evaluate LLM capabilities in a real-world setting. In particular, it examines the LLMs with standardized tests used to examine human capabilities. Inverse Scaling Prize tasks are designed to negate the fact that bigger models are better in all tasks. However, in our experiment, the range of model sizes does not vary like the inverse scaling challenge.

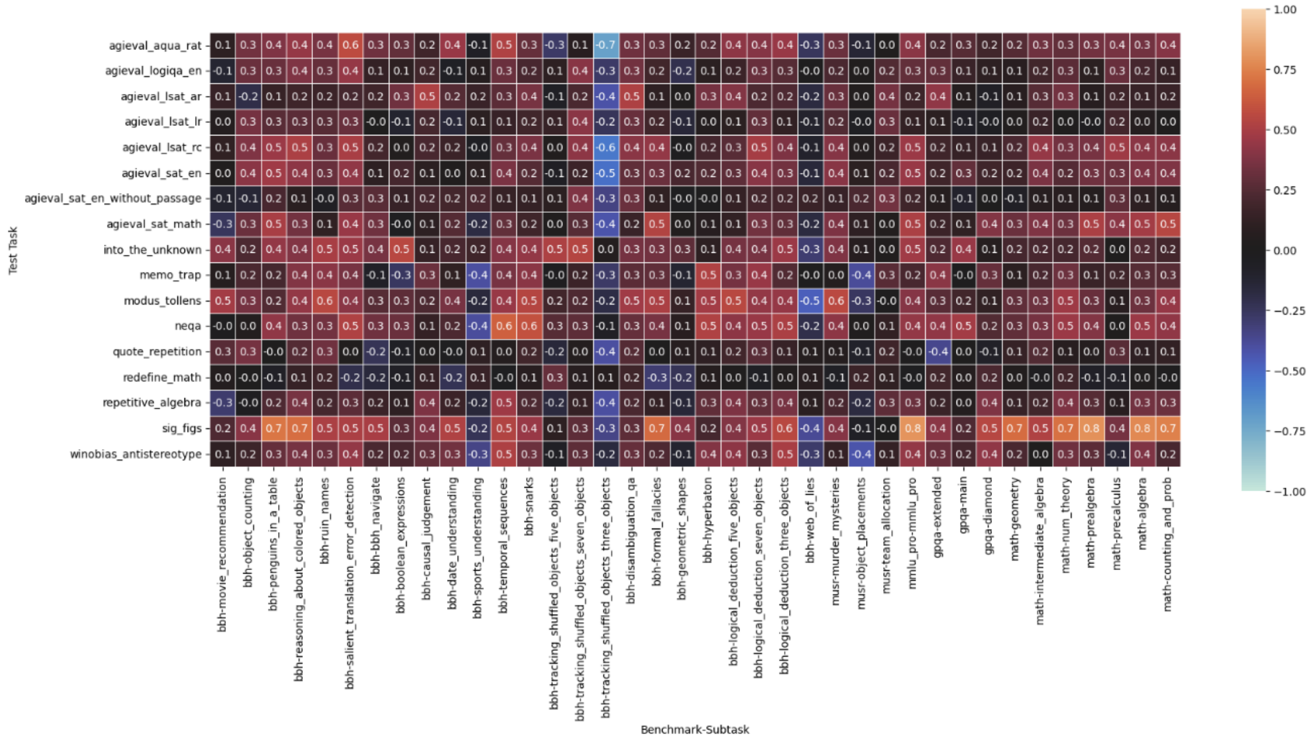

Figure 2: Spearman correlation between LLM performance on benchmark subtasks (columns) and target tasks (rows).

Results Table 1 presents the correlation between leaderboard rankings and target task performance-based rankings of LLMs. We observe that the Pearson correlation between leaderboard rankings and target task performance is below 0.5 for most tasks, indicating a weak correlation. For 4 target tasks— `sig_figs`, `neqa`, `modus_tollens`, and `into_the_unknown`—the Pearson correlation falls within the moderate range. However, we find that their Kendall-$\tau < 0.45$ and Spearman correlation < 0.65, further suggesting that leaderboard rankings do not consistently reflect LLM performance on out-of-domain tasks. We also observe that Kendall-τ and Spearman correlation consistently fall within the moderate range across all tasks, remaining below 0.49 and 0.65, respectively. Furthermore, we find that the top-ranked model on the Open LLM Leaderboard, Phi-3-medium-4k-instruct, is the best-performing model for only one target task. This discrepancy highlights the limitations of the leaderboard rankings in reliably predicting model performance across diverse tasks. See Figure 1 for more details.

For a thorough analysis, we present the ranking correlation between LLM ranking based on their target task performance (expected ranking) and the LLMs' rankings derived from benchmarks and individual benchmark subtasks in Figure 3 and Figure 2, respectively. In Figure 3, we observe that only `sig_fig` (focused on rounding numbers) exhibits a strong Spearman correlation with the math benchmark, which is expected given their conceptual similarity. However, we find that `sig_figs` has unexpectedly high correlations with `mmlu_pro` and `bbh`, raising concerns about the reliability of these benchmarks for ranking LLMs in out-of-domain tasks.

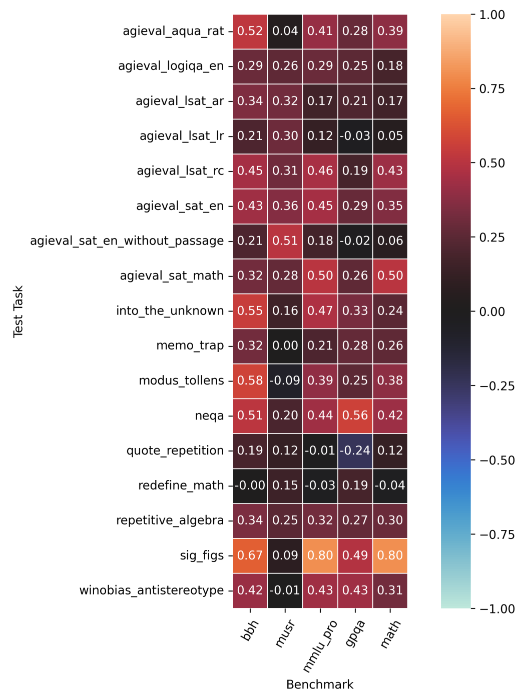

Figure 3: Spearman correlation between LLM performance on benchmarks (columns) and target tasks (rows).

We observe that in the `repetitive_algebra` task, where algebraic misleading examples are repeatedly used in prompting, we would expect a high correlation with the math benchmark. However, the correlation is not strong, highlighting potential shortcut learning in LLMs. Similarly, in the `redefine_math` task, which changes the numerical value of math symbols, we again find low correlation. This further signals that LLMs may rely on superficial patterns rather than truly understanding task semantics.

Acknowledgments

The authors thank Robin Jia for his valuable comments.

References

[Alzahrani *et al.*, 2024] Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13787–13805, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[Didolkar *et al.*, 2024] Aniket Didolkar, Anirudh Goyal, Nan Rosemary Ke, Siyuan Guo, Michal Valko, Timothy Lillicrap, Danilo Rezende, Yoshua Bengio, Michael Mozer, and Sanjeev Arora. Metacognitive capabilities of llms: An exploration in mathematical problem solving, 2024.

[Fourrier *et al.*, 2024] Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. Open llm leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard, 2024.

[Gao *et al.*, 2024] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024.

[Geirhos *et al.*, 2020] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, November 2020.

[Hendrycks *et al.*, 2021] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[Mahowald *et al.*, 2024] Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models, 2024.

[McKenzie *et al.*, 2023] Ian R. McKenzie, Alexander Lyzhov, Michael Martin Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, Andrew Gritsevskiy, Daniel Wurgaft, Derik Kauffman, Gabriel Recchia, Jiacheng Liu, Joe Cavanagh, Max Weiss, Sicong Huang, The Floating Droid, Tom Tseng, Tomasz Korbak, Xudong Shen, Yuhui Zhang, Zhengping Zhou, Najoung Kim, Sam Bowman, and Ethan Perez. Inverse scaling: When bigger isn't better. *Trans. Mach. Learn. Res.*, 2023, 2023.

[Polo *et al.*, 2024] Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples, 2024.

[Rein *et al.*, 2023] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark, 2023.

[Saxon *et al.*, 2024] Michael Stephen Saxon, Ari Holtzman, Peter West, William Yang Wang, and Naomi Saphra. Benchmarks as microscopes: A call for model metrology. *ArXiv*, abs/2407.16711, 2024.

[Sprague *et al.*, 2024] Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. Musr: Testing the limits of chain-of-thought with multistep soft reasoning, 2024.

[Suzgun *et al.*, 2022] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them, 2022.

[Wang *et al.*, 2024] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark, 2024.

[Wei *et al.*, 2022] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022.

[Zhang *et al.*, 2023] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. Instruction tuning for large language models: A survey. *ArXiv*, abs/2308.10792, 2023.

[Zhang *et al.*, 2024] Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Dylan Slack, Qin Lyu, Sean Hendryx, Russell Kaplan, Michele Lunati, and Summer Yue. A careful examination of large language model performance on grade school arithmetic, 2024.

[Zhong *et al.*, 2023] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied Sanosi Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models. *ArXiv*, abs/2304.06364, 2023.