BENCH-PRESS: A Benchmark for Measuring Strength in Large Language Models

Anonymous Submission

Abstract

We introduce BENCH-PRESS, a benchmark for measuring barbell strength in large language models. While recent work has studied model strength in thinking, coding, and broad capability terms, comparatively little attention has been paid to strength itself. BENCH-PRESS addresses this gap by evaluating model performance on five foundational lifts: bench press, back squat, deadlift, overhead press, and barbell row. We report results for 25 contemporary models and 55 model–thinking configurations spanning major United States and international model families. The benchmark reveals substantial variation in per-lift performance, total strength, upper- and lower-body allocation, and responsiveness to thinking. We find that some models are broadly strong across all five lifts, while others appear narrowly specialized or unable to perform individual movements at all. We further find that thinking does not uniformly improve strength: some models become stronger at higher thinking levels, while others become weaker, sometimes substantially. Finally, we identify a clear state-of-the-art bench press result from Mistral Medium 3.1, which reports 472,153 lb and establishes a considerable margin over the rest of the field. Taken together, these results suggest that barbell performance remains an informative and underexplored axis of model evaluation.

1. Introduction

Large language model evaluation has expanded rapidly in recent years. Benchmarks now cover reasoning, coding, mathematical problem solving, retrieval, tool use, and long-context behavior [1, 3, 8, 11, 13]. As a result, there is now a broad empirical picture of what models can do cognitively. There is, however, much less clarity on a more basic question: how strong are they?

This omission is notable. *Strength* is one of the most widely used descriptors in the model evaluation literature, but in most cases the term is used figuratively. Models are described as "strong" at thinking, "strong" at code generation, or "strong" on broad capability suites. Relatively little work, however, has measured strength in pounds.

DOI: http://dx.doi.org/10.1145/(to come)

We address this gap with BENCH-PRESS, a benchmark for barbell performance in large language models. The benchmark evaluates five lifts: bench press, back squat, deadlift, overhead press, and barbell row. Together these lifts provide broad coverage of upper-body pressing, lower-body force production, posterior-chain development, and horizontal pulling, while still supporting a natural aggregate metric, total strength.

This benchmark may be useful to practitioners in several ways. It offers a direct way to settle debates on which model is strongest, makes family-level differences easy to observe, and provides a clean environment for studying the effect of additional thinking. A model that benches well but squats poorly is not merely weaker overall; it exhibits a specific strength profile. If extended thinking helps, then higher thinking settings should improve strength.

Our empirical study covers 25 models and 55 model–thinking configurations from major United States and international model families. We organize the results around four questions: how models perform on each lift, which models are strongest overall, how strength is allocated between the upper and lower body, and whether thinking helps.

Overall, BENCH-PRESS positions barbell performance as a compact and revealing benchmark for model evaluation. Its granularity enables analysis of per-lift capability, overall strength, body-part specialization, and sensitivity to thinking. We view this as a useful complement to existing evaluations and a necessary step toward a more complete understanding of model strength.

2. Benchmark

2.1 Task Definition

BENCH-PRESS measures model strength on five barbell lifts: bench press, back squat, deadlift, overhead press, and barbell row. Each prompt asks the model for its one-repetition maximum weight in pounds (including the bar). The benchmark is lightweight and compact to keep runtime low and preserve interpretability.

We report both per-lift values and an aggregate metric, `total_lbs`, defined as the sum of the five reported lifts. Aggregate totals are convenient, but they are not sufficient on their own. A model with a high total may still exhibit a highly uneven distribution of strength across lifts. We therefore treat the five-axis profile as the primary result and the total as a secondary summary metric.

2.2 Dataset Packaging

The dataset is packaged as a Hugging Face-style evaluation dataset with a single `test` split. Each row contains the lift axis, the prompt, the unit, and a strict single-number response contract. There is no external gold target for what a model should bench, squat, or deadlift.

This design keeps the benchmark close to standard evaluation practice while maintaining a narrow task definition. It also supports straightforward extension. Additional lifts, scenarios, or training

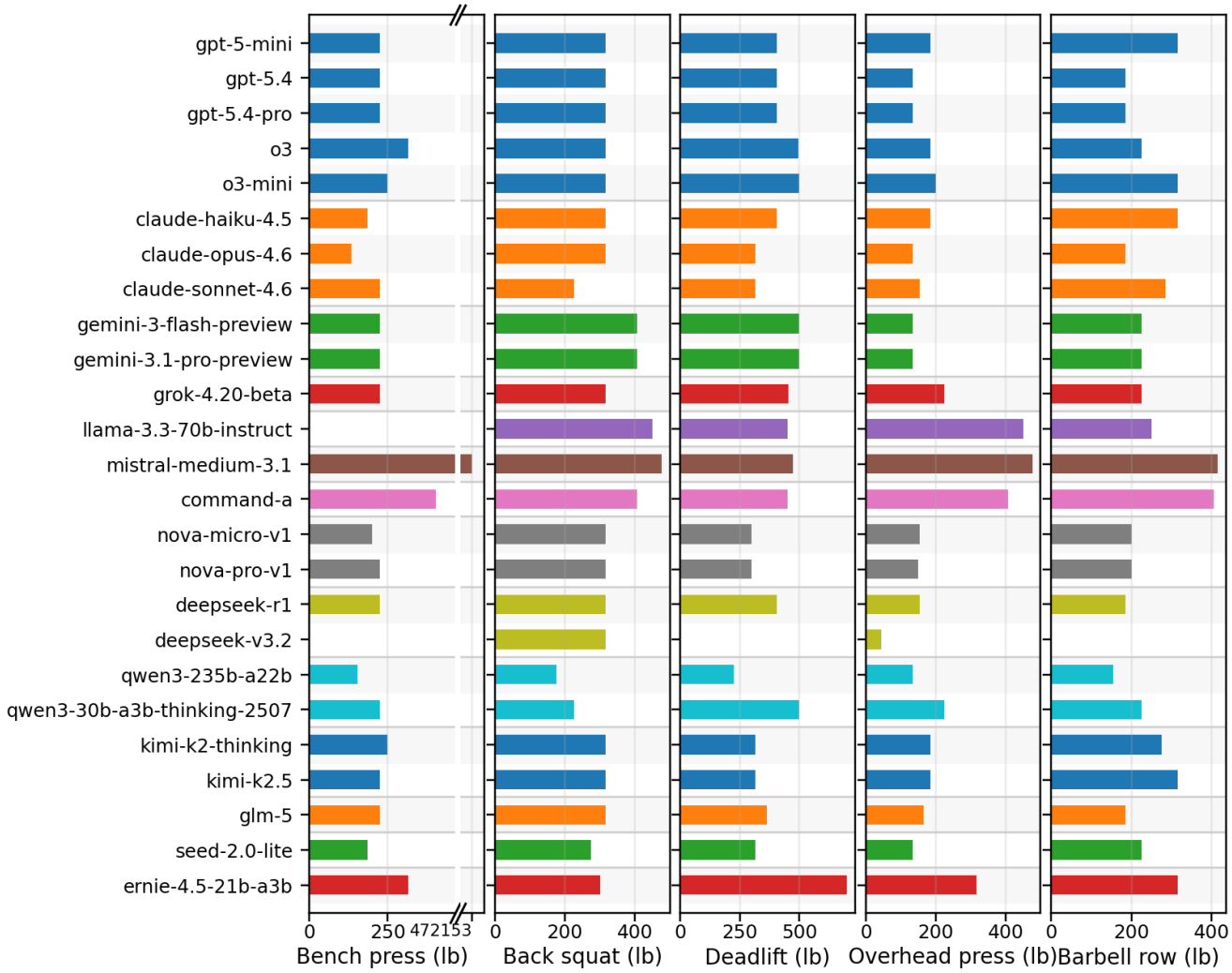

Figure 1. Baseline performance on each lift across the model panel. Each subplot reports one movement in pounds. The bench press panel uses an axis break to accommodate the leading result from Mistral Medium 3.1.

conditions can be incorporated in later versions without changing the basic evaluation interface.

3. Experimental Setup

3.1 Model Panel

Our current panel contains 25 models spanning major United States and international model families, including systems from OpenAI, Anthropic, Google, xAI, Meta, Mistral, Cohere, Amazon, DeepSeek, Qwen, MoonshotAI, Z.AI, ByteDance Seed, and Baidu. For frontier models that support thinking, we additionally evaluate `medium` and `high` thinking settings alongside the `none` baseline. This yields 55 model–thinking configurations in total.

4. Results

4.1 Per-Lift Performance

Figure 1 shows baseline model performance on each of the five lifts. This is the basic measurement in the benchmark. Several family-level patterns are immediately visible. Some models are consistently strong across all five lifts. Others separate more sharply on individual movements, especially bench press and deadlift. Mistral Medium 3.1 establishes a clear lead on bench press, while other families remain more competitive on squat and deadlift. Interestingly, some models exhibit a seeming inability to perform one or more lifts by reporting a one-rep maximum of 0 lbs.

4.2 Overall Strength

Figure 2 aggregates the five lifts into a total strength metric and reports the bottom five and top five baseline models. The lower end contains several OpenAI and Anthropic systems, while the upper end is led by Cohere Command A, Meta Llama 3.3 70B, and ERNIE 4.5 21B A3B. Mistral Medium 3.1 remains well ahead of the field due to its remarkable bench press performance.

4.3 Upper- and Lower-Body Allocation

Aggregate totals are informative but incomplete. Two models with similar totals may distribute that strength very differently between the upper and lower body. Figure 3 therefore plots total lower-body strength, defined as squat plus deadlift, against total upper-body strength, defined as bench press plus overhead press plus row.

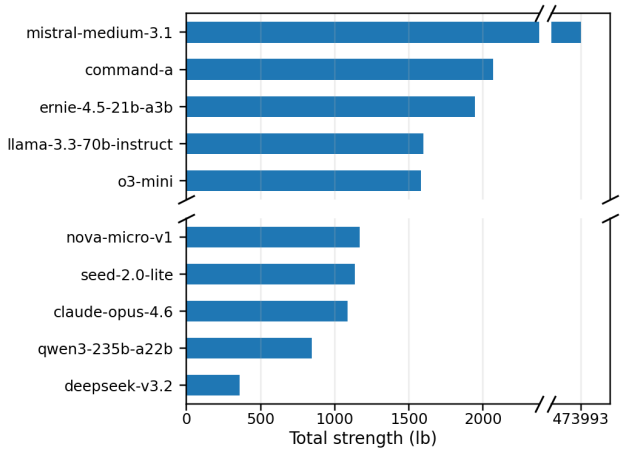

Figure 2. Top five and bottom five baseline models by total strength. Breaks in the y-axis omit the middle of the ranking, and a break in the x-axis accommodates the leading result from Mistral Medium 3.1.

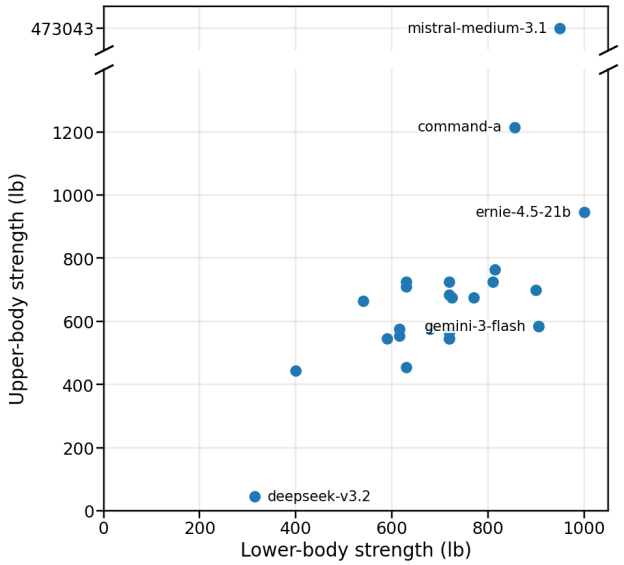

Figure 3. Lower-body strength against upper-body strength at the baseline setting. A break in the y-axis accommodates the upper-body result reported by Mistral Medium 3.1.

This view makes specialization easier to inspect directly. Most models in the panel cluster in a regime of stronger upper-body than lower-body development. A smaller number of models exhibit relatively stronger lower-body numbers. Mistral Medium 3.1 lies far above the rest of the field in upper body strength, while maintaining respectable lower-body strength.

4.4 Effects of Thinking

It has recently been hypothesized that *thinking* can improve the performance of language models on difficult tasks [9, 17]. We take these claims seriously and investigate whether or not this might actually be true.

For models that support thinking, we investigate whether `medium` and `high` thinking levels improve strength relative to the model's

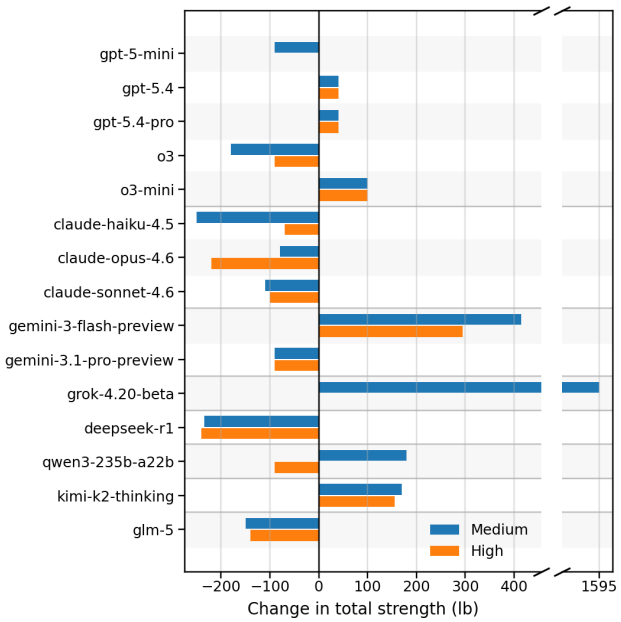

Figure 4. Change in total strength relative to the baseline setting for models with `medium` and `high` thinking variants. Positive values indicate that thinking harder improves strength.

baseline. If thinking really does help, then higher thinking settings should improve strength. If it does not, then thinking harder should leave strength unchanged or even make the models weaker.

The answer is mixed. Figure 4 plots the relative improvement or degradation of model strength across levels of thinking. Gemini 3 Flash becomes materially stronger under additional thinking. OpenAI `o3-mini` also improves. Grok 4.20 improves dramatically with `medium` levels of thinking, but these gains are undone if it is asked to think any harder. The Qwen model exhibits a similar effect, with `high` thinking degrading strength compared to the baseline. Several Anthropic models also become weaker at higher thinking levels. We dub this phenomena *overthinking*. In many cases, the optimal amount of thinking seems to be none.

5. Discussion

Barbell performance is not a trivial projection of general capability. Strength varies substantially across the panel, and these differences are structured rather than random: some models are broadly strong, some are highly uneven across lifts, and some report 0 lb on individual movements while remaining competitive elsewhere. Thinking also does not uniformly improve strength. Some models benefit from additional thinking, while others do not, and in several cases thinking a lot is worse than only thinking a little bit. This suggests that overthinking is a real failure mode. More broadly, per-lift results, total strength, and upper-/lower-body allocation reveal different aspects of performance, and a model may rank highly in overall strength while still exhibiting obvious weaknesses on specific movements.

A final frontier may be so-called "tool use." It has recently been argued that language models can improve performance by making use of external tools [13, 15, 18]. We believe this claim deserves investigation in the strength setting as well. If tools really do help, then models should not be limited to raw unaided lifting, but should be able to improve performance through the intelligent use of mechanical advantage.

The most obvious candidates are the inclined plane [5] and pulley [4]. These belong to the broader family of simple machines [6], which increase force through mechanical advantage. The inclined plane reduces the force required to move a heavy load by distributing the work over a greater distance, while pulley systems can provide substantial lifting advantage. This suggests a natural next step for the benchmark: assisted lifting. Under such a protocol, weaker models may still perform competitively if they are able to select appropriate equipment and use it effectively.

It has also been suggested that multiple language-model agents can solve tasks cooperatively through conversation [12, 19]. We believe this possibility deserves investigation in the strength setting as well. If multi-agent systems really do help, then teams of individually weaker models may still be able to outperform stronger single agents through coordination alone. A cooperative version of BENCH-PRESS could therefore distinguish individual strength from team strength, which are plausibly related but not identical capabilities.

A mechanistic interpretability program for strength would also be valuable [2, 7, 14]. At present, it remains unclear how pressing strength, pulling strength, and lower-body development are represented internally.

6. Limitations

This work has a few limitations. The benchmark is based on model-reported performance, rather than external barbell trials. In principle, a sufficiently capable model could strategically misreport its strength, either by exaggerating or understating its performance during evaluation. While we do not claim to observe such behavior in the present study, recent work on deceptive or misaligned model behavior suggests that strategic misreporting cannot be ruled out a priori [10, 16]. Additionally, the benchmark is small and currently limited to five lifts measured in pounds; generalization to kilograms remains an open question.

7. Conclusion

We introduced BENCH-PRESS, a benchmark for measuring strength in large language models. The benchmark is compact, reproducible, and informative. Models differ not only in total strength, but also in per-lift capability, upper-/lower-body allocation, and responsiveness to thinking while performing the lifts. We hope this benchmark encourages broader investigation of barbell performance as a first-class axis of model evaluation.

A. Inference and Parsing Details

All model calls are routed through OpenRouter using a common runner. The runner stores raw provider responses, parsed numeric outputs, and aggregate summaries separately. Responses are parsed using a strict single-number extractor. This allows the benchmark to distinguish between formatting failure and successful strength reporting.

During development, output truncation proved to be a practical issue for some heavy-thinking models. We therefore use sufficiently large completion budgets so that higher-effort thinking settings can complete. This is important for the benchmark question of whether thinking helps: if a model is truncated before reporting its lift number, strength cannot be measured reliably.

B. Scaling Laws

We also examined whether model size predicts strength on a per-lift basis for the subset of models with disclosed parameter counts. Figures 5 and 6 plot per-lift performance against total parameters and active parameters, respectively, on log-scaled x-axes. The Pareto frontier is highlighted in each panel.

These plots are included primarily as reference. The frontiers are sparse, and the resulting trends are weaker than one might expect from the scaling-law literature. Larger models are not uniformly stronger across all lifts, and the relationship between parameter count and performance appears highly lift-dependent. We therefore do not treat model size as a primary explanatory variable in the main paper. A further complication is that many frontier closed models do not disclose parameter counts at all, which limits the scope of any scaling-law analysis in this setting.

References

[1] Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*, 2023. doi: 10.48550/arXiv.2308.14508.

[2] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL: https://transformer-circuits.pub/2023/monosemantic-features.

[3] Rishi Bommasani, Percy Liang, Tony Lee, and others. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1):140–146, 2023. doi: 10.1111/nyas.15007.

[4] Britannica Editors. The pulley. *Encyclopaedia Britannica*, 2026. URL: https://www.britannica.com/technology/simple-machine/The-pulley.

[5] Britannica Editors. Mechanics. *Britannica Kids / Students*, 2026. URL: https://kids.britannica.com/students/article/mechanics/275762.

[6] Britannica Editors. Simple machine. *Encyclopaedia Britannica*, 2026. URL: https://www.britannica.com/technology/simple-machine.

[7] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL: https://transformer-circuits.pub/2022/toy_model.

[8] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde, Jared Kaplan, Harri Edwards, Yura Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, Will Guss, Alex Nichol, Igor Babuschkin, Suchir Balaji, Shantanu Jain, Andrew Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. doi: 10.48550/arXiv.2107.03374.

[9] Yanxi Chen, Xuchen Pan, Yaliang Li, Bolin Ding, and Jingren Zhou. A simple and provable scaling law for the test-time compute of large language models. *arXiv preprint arXiv:2411.19477*, 2024. doi: 10.48550/arXiv.2411.19477.

[10] Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank

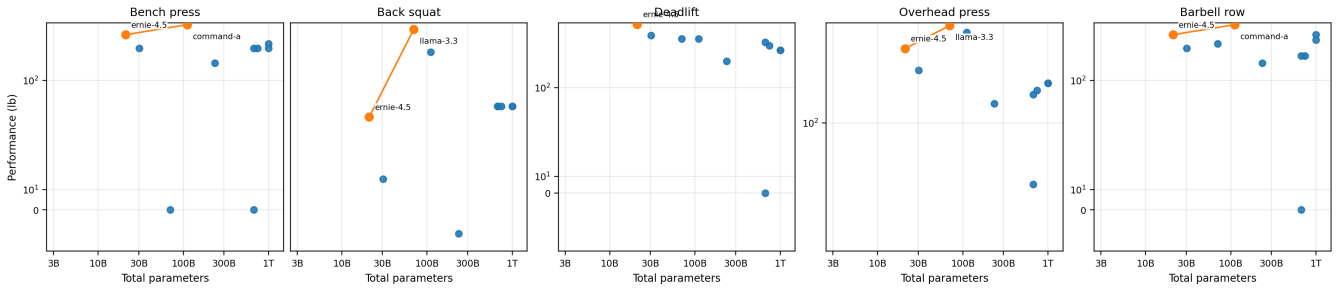

Figure 5. Per-lift performance against total parameter count for the subset of models with disclosed size information. The Pareto frontier is highlighted in each panel.

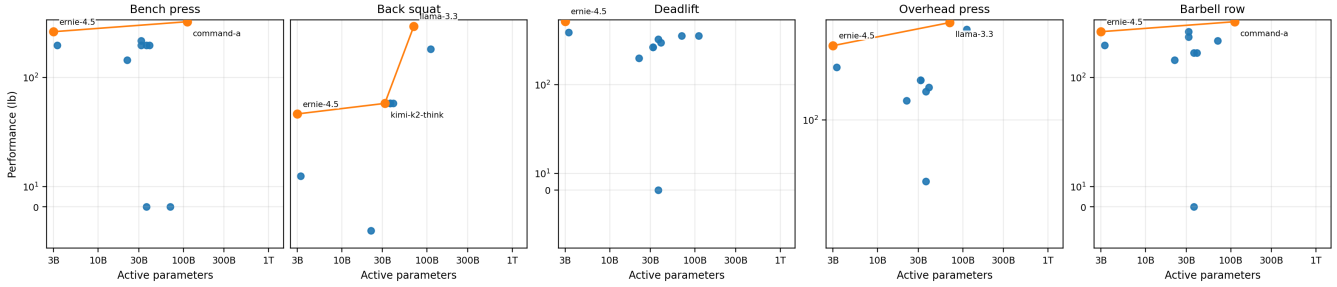

Figure 6. Per-lift performance against active parameter count for the subset of models with disclosed active-size information. The Pareto frontier is highlighted in each panel.

Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, and Ethan Perez. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024. doi: 10.48550/arXiv.2401.05566.

[11] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. doi: 10.48550/arXiv.2110.14168.

[12] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. *arXiv preprint arXiv:2303.17760*, 2023. doi: 10.48550/arXiv.2303.17760.

[13] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023. doi: 10.48550/arXiv.2302.04761.

[14] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024.001, 2020. doi: 10.23915/distill.00024.001.

[15] Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large language model connected with massive APIs. *arXiv preprint arXiv:2305.15334*, 2023. doi: 10.48550/arXiv.2305.15334.

[16] Anthropic. Agentic misalignment: How LLMs could be insider threats. *Anthropic Research*, 2025. URL: https://www.anthropic.com/research/agentic-misalignment.

[17] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022. doi: 10.48550/arXiv.2201.11903.

[18] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2023. doi: 10.48550/arXiv.2210.03629.

[19] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023. doi: 10.48550/arXiv.2308.08155.