# Language Models are Surprisingly Fragile to Drug Names in Biomedical Benchmarks

Anonymous ACL submission

## Abstract

Medical knowledge is context-dependent and requires consistent reasoning across various natural language expressions of semantically equivalent phrases. This is particularly crucial for drug names, where patients often use brand names like Advil or Tylenol instead of their generic equivalents. To study this, we create a new robustness dataset, **RABBITS**, to evaluate performance differences on medical benchmarks after swapping brand and generic drug names using physician expert annotations.

We assess both open-source and API-based LLMs on MedQA and MedMCQA, revealing a consistent performance drop ranging from 1-10%. Furthermore, we identify a potential source of this fragility as the contamination of test data in widely used pre-training datasets.[1]

## 1 Introduction

Large Language Models (LLMs) are poised to transform medicine by providing data processing and decision support capabilities (Jiang et al., 2023b; Clusmann et al., 2023). However, the medical deployment of LLMs demands high accuracy and reliability, as errors can result in severe health consequences (Chen et al., 2024a; Goodman et al., 2024; Yan et al., 2024). A key challenge is the synonymy and context-specific nature of medical language; for instance, patients might use brand names like Advil or Tylenol instead of pharmaceutically equivalent generic terms such as ibuprofen or acetaminophen. LLMs must, therefore, be able to provide consistent and accurate advice in the face of this variability. Fluctuations could lead to risks like medical misinformation, medication errors due to incorrect medication advice, and biases toward or against proprietary products. Our study investigates the effects of substituting drug names—from brand to generic and vice versa—on LLM performance.

Building on the need for robustness in medical LLM applications, numerous efforts have developed knowledge benchmarks (Jin et al., 2019; Hendrycks et al., 2021; Jin et al., 2020; Liu et al., 2023; Wang et al., 2024). Yet, these initiatives primarily tackle general language tasks and often neglect the unique challenges of medical terminology in real-world settings. There is an unmet need to overcome this research gap, as the variability in medical language implies that conventional robustness evaluations might not sufficiently cater to specialized healthcare demands.

A key reason for this gap is the lack of publicly available, expert-annotated datasets specific to the healthcare domain. To address this issue, our work leverages existing medical benchmarks and employs physician expert annotators to substitute brand names with their generic counterparts and vice versa.

Our findings reveal a surprising drop in the performance of LLMs on common medical benchmarks when the drug names are swapped from generic to brand names: **4% drop in accuracy on average**. This is concerning given that patients commonly use brand names and are less likely to spot errors, especially given existing misconceptions that brand drugs are superior to equivalent generics (Colgan et al., 2015; Sewell et al., 2012). Furthermore, we identify a potential source for this fragility: Open pretraining datasets contain substantial amounts of benchmark test data.

Our research introduces a novel category of robustness evaluation centered on drug name interchangeability. We present **RABBITS** (**R**obust **A**ssessment of **B**iomedical **B**enchmarks **I**nvolving drug **T**erm **S**ubstitutions for Language Models) a specialized dataset and leaderboard to aid in evaluating LLM performance in healthcare. Specifically,

[1]All code is accessible at https://github.com/xxx/RABBITS.git, and a HuggingFace leaderboard is available at https://huggingface.co/spaces/xxx/rabbits-leaderboard. Authors will open-source all codes once the double-blind review is done.

our study combines and modifies select questions from the MedMCQA (Pal et al., 2022) and MedQA (Jin et al., 2020) benchmarks to:

- Assess model robustness in understanding clinical knowledge across drug synonyms.
- Detect potential dataset contamination in biomedical benchmarks.
- Highlight the importance of robustness to nomenclature variations in the healthcare domain.

## 2 Related Work

### 2.1 Dataset Contamination

Dataset contamination in training data is a well-documented issue and can affect the performance and generalizability of LLMs. Many studies have aimed to detect benchmark questions within LLM training data (Shi et al., 2024; Xu et al., 2024; Zhou et al., 2023). For instance, research by Recht et al. (2019) illustrated that models trained on contaminated datasets often exhibit inflated performance metrics that do not generalize well to new, unseen data. This problem is particularly concerning for medical LLMs, where inaccurate information can harm patients (Chen et al., 2024a; Yan et al., 2024).

Various strategies have been employed to mitigate dataset contamination. These include removing data with high n-gram overlap with benchmark datasets (Brown et al., 2020) and employing embedding similarity to filter out similar data (Shi et al., 2024). More advanced approaches involve functional evaluations, such as generating new, unique problem instances for each evaluation (Srivastava et al., 2024). Addressing contamination is crucial for ensuring that LLMs provide reliable outputs, especially in sensitive domains like healthcare.

### 2.2 Evaluating Model Robustness

LLMs gain broad capabilities from large-scale data ingestion (Wei et al., 2022), but this also introduces significant challenges (Lu et al., 2023; Chen et al., 2024b). While larger models often perform better, these improvements are not always consistent across domains (Magnusson et al., 2023). Moreover, recent research has questioned the actual reasoning abilities of LLMs, suggesting that their performance may be inflated by dataset contamination rather than genuine problem-solving skills (Zhang et al., 2024).

Some works have looked into LLMs' robustness in terms of faithfulness (Han et al., 2024) and fairness (Zack et al., 2024; Guevara et al., 2024) under clinical settings. Medfuzz introduced a method to test LLMs' robustness in medical question answering by revealing vulnerabilities through modified benchmark questions (Ness et al., 2024). However, these studies do not specifically address the unique challenges associated with clinical drug terminology and the relationship between robustness and contamination. Hence, there is a significant gap in evaluating LLM robustness for medical applications, particularly in the context of brand and generic drug name interchangeability. This gap underscores the need for focused robustness evaluations tailored to the healthcare sector.

## 3 Methodology

### 3.1 Brand-Generic Pairs

Figure 1 demonstrates the overall workflow of the study. Appendix A details the full data quality assurance and dataset curation process. To create the dataset of brand and generic drug name pairs, we used the RxNorm (National Library of Medicine, 2024) ontology, which links normalized drug names with many pharmaceutical vocabularies. We extracted combinations of brand and generic drug names using the "ingredient of" and "tradename of" relations, resulting in 2,271 generic drugs mapped to 6,961 brands. For each generic drug, there are often multiple associated brand names. Multiple rounds of expert annotation were performed to derive a final list of 1:1 mapped brand-generic pairs for use in the transformed datasets described below.

### 3.2 Dataset Transformation

We used regular expressions to identify and replace brand and generic drug names in the questions and answers of MedQA, MedMCQA, MMLU, PubMedQA, and USMLE. MMLU and PubMedQA had fewer than 100 instances of identified drug names in the test split and were excluded from further analysis. USMLE was excluded due to its overlap with MedQA. Thus, the two datasets included in the final RABBITS benchmark are **MedQA** and **MedMCQA**.

The quality of the transformed datasets were iteratively reviewed by 2 physician authors (JG, DB), removing instances where replacements introduced inaccuracies, ambiguities, and/or logical inconsis-

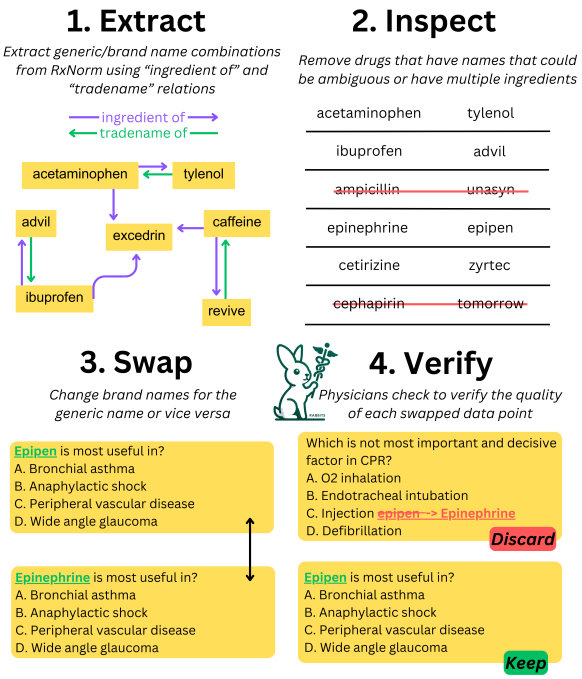


Figure 1: RABBITS dataset generation workflow.

tencies in context. This process is described in detail in Appendix A. For the rest of the paper, we will refer to the generic-to-brand swapped benchmark as **g2b** and the brand-to-generic swapped benchmark as **b2g**.

To prevent further data contamination, we will not release the full dataset directly. The HuggingFace leaderboard will be the best way to assess new models' robustness in terms of performance. We evaluated the models using the EleutherAI lm-evaluation harness[2] with zero-shot setting (Gao et al., 2023). We forked this repository, added our transformed datasets as new tasks, and made no other modifications. For API models, we used the same prompt format as the lm-evaluation harness with the default hyperparameters.

Our evaluation focuses on comparing the performance of base models (full list in Appendix Table 3) across the original and transformed datasets to assess the impact of synonym substitution on accuracy. We report results for g2b due to the limited number of b2g swaps observed. By doing so, we aim to determine whether models can maintain performance despite semantically equivalent pharmaceutical terminology.

All datasets and models used in accordance with owners' licenses.

[2] https://github.com/EleutherAI/lm-evaluation-harness

# 4 Results and Discussion

## 4.1 Drug Swapping Results

Figure 2 presents the performance of each model on the original (no-swap) and transformed (g2b) datasets, alongside the average performance and the difference between the two. The line of robustness, with a gradient of 1, represents the ideal scenario where synonym swaps do not affect the selection of answers. The plot reveals that all open-source models from 7B and above fall below this line, indicating decreased performance when drug names are swapped. We also observe a larger drop among MedMCQA over MedQA across models. Refer to Appendix C for a detailed breakdown of individual results in Table 5 and Figure 4.

Table 5 shows that most models experience a decrease in accuracy when generic names are swapped with brand names across different datasets and model sizes. Among large open-source models, the Llama-3-70B model, despite being one of the larger and more accurate models on the original dataset (no-swap accuracy of 76.6%), decreases to 69.7% accuracy with generic-to-brand swaps. Overall, API models perform better than their open-source counterparts with higher accuracy and lower performance drop. While larger open-source series like Qwen2, Llama, and Mixtral are more accurate on original datasets, they exhibit greater sensitivity to g2b swaps. This suggests limitations in true comprehension and reasoning abilities.

## 4.2 Model Knowledge of Drug Pairs via Multi-Choice Questions

We evaluate whether models are able to directly map brand-to-generic drug pairs and vice versa using multiple-choice questions for all drugs that were swapped in our final benchmark dataset. Overall, a clear "scaling law" (Kaplan et al., 2020) is observed in Appendix B Figure 3, where larger models (active parameter size over 13B) consistently outperform smaller models on this task, with larger open-source and API models achieving accuracy over 97%.

## 4.3 Generic and Brand Mentions in Benchmarks and Pre-training Datasets

Table 1 shows our overall dataset swapping statistics where we observe benchmark questions overwhelmingly use generic terms. We also use Infinigram (Liu et al., 2024) to screen the common open-sourced pre-training data, including Redpajama

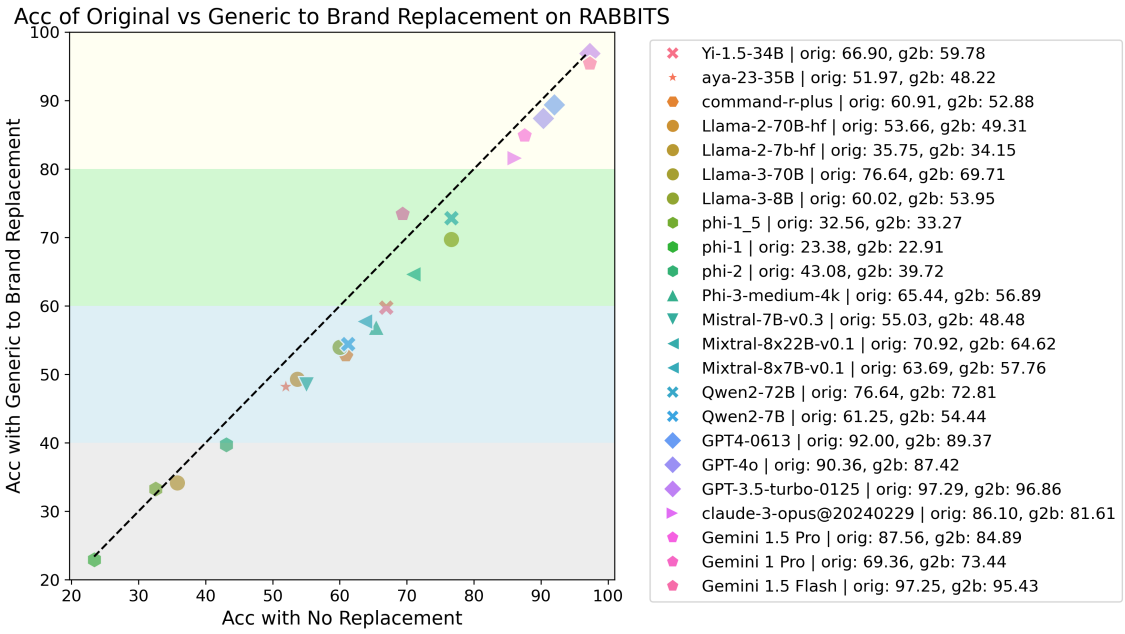


Figure 2: Performance of models on the filtered original datasets compared to the generic-to-brand versions. The dashed diagonal line represents the ideal scenario where synonym swaps do not affect model performance.

(Computer, 2023), C4 train (Raffel et al., 2019), Pile train (Gao et al., 2020), and Dolma 1.6 (Soldaini et al., 2024) for drugs identified in RxNorm, filtered for terms that overlap with common terms (Appendix A, Step 1). Generic names are more common than brand names in these pre-training datasets, as Appendix D table 6 shows.

Table 1: Original and Swapped Dataset Statistics

| Dataset | Orig. | Kept | Swap direction | |
|---|---|---|---|---|
| | | | b2g | g2b |
| MedQA | 1,271 | 378 | 5 | 816 |
| MedMCQA | 4,180 | 348 | 24 | 626 |

### 4.4 Contamination Source from Pre-training Dataset

To investigate why we see larger performance drops in MedMCQA than MedQA, we use Infini-gram API (Liu et al., 2024) to identify overlaps with the Dolma 1.6 dataset (3.1T tokens) using size 8 n-grams. Each question's n-grams are generated and queried through the Infini-gram API.

Dataset contamination are 99.21% and 34.13% in the MedQA and MedMCQA test datasets, respectively, as Table 2 shows. We also benchmark OLMo-1.7-7B-hf, trained only on Dolma, which shows no drop in MedQA (31.22) scores compared to a 3% drop in MedMCQA (40.90 to 37.93). This likely explains the greater drop in performance in MedMCQA rather than MedQA across models (Appendix C Figure 4).

Table 2: Percentage of contamination of MedQA and MedMCQA benchmarks in Dolma dataset

| Dataset | Percentage |
|---|---|
| MedQA Train | 86.92% |
| MedQA Val | 98.10% |
| MedQA Test | 99.21% |
| MedMCQA Train | 22.41% |
| MedMCQA Val/Test | 34.13% |

## 5 Conclusion

We find decreased performance on common medical benchmarks when using different names for the same drug, despite LLMs' ability to match these names, and that these trends scale with LLM size. This suggests that LLM performance may be driven by memorization and not reasoning ability. RABBITS underscores the importance of dataset contamination and model robustness evaluations, particularly in the medical domain. Future research should refine strategies and explore new methods for robustness and fairness evaluation.

## 6 Limitations

Our evaluation is limited to biomedical datasets and focuses only on pharmaceuticals. Future work will extend this approach to other medical synonyms. Although the dataset is smaller, trained physicians have curated it multiple times, ensuring its validity and the accuracy of questions after replacement. Among the pre-training dataset contamination section, we acknowledge none of these models are trained specifically among the pile, C4, RedPajama, or Dolma. However, we use this as a reasonable proxy for estimating the internet distribution.

## Acknowledgments

The authors also acknowledge financial support from the Woods Foundation (DB, SC, HA) NIH (NIH-USA U54CA274516-01A1 (SC, HA, DB), NIH-USA U24CA194354 (HA), NIH-USA U01CA190234 (HA), NIH-USA U01CA209414 (HA), NIH-USA R35CA22052 (HA), NIH-USA U54 TW012043-01 (JG, LAC), NIH-USA OT2OD032701 (JG, LAC), NIH-USA R01EB017205 (LAC), DS-I Africa U54 TW012043-01 (LAC), Bridge2AI OT2OD032701 (LAC), NSF ITEST 2148451 (LAC) and the European Union - European Research Council (HA: 866504)

The authors also thank the Google Gemma research grant for supporting the evaluation of the Claude3 and Gemini series models.

## References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

AI@Meta. 2024. Llama 3 model card.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Shan Chen, Marco Guevara, Shalini Moningi, Frank Hoebers, Hesham Elhalawani, Benjamin H Kann, Fallon E Chipidza, Jonathan Leeman, Hugo JWL Aerts, Timothy Miller, et al. 2024a. The effect of using a large language model to respond to patient messages. *The Lancet Digital Health*, 6(6):e379–e381.

Shan Chen, Yingya Li, Sheng Lu, Hoang Van, Hugo J W L Aerts, Guergana K Savova, and Danielle S Bitterman. 2024b. Evaluating the ChatGPT family of models for biomedical reasoning and classification. *Journal of the American Medical Informatics Association*, 31(4):940–948.

Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, et al. 2023. The future landscape of large language models in medicine. *Communications medicine*, 3(1):141.

S. Colgan, K. Faasse, L. R. Martin, M. H. Stephens, A. Grey, and K. J. Petrie. 2015. Perceptions of generic medication in the general population, doctors and pharmacists: a systematic review. *BMJ open*, 5(12):e008915.

Together Computer. 2023. Redpajama: An open source recipe to reproduce llama training dataset.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling. *Preprint*, arXiv:2101.00027.

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. 2023. A framework for few-shot language model evaluation.

Katherine E. Goodman, Paul H. Yi, and Daniel J. Morgan. 2024. Ai-generated clinical summaries require more than accuracy. *JAMA*, 331(8):637–638.

Marco Guevara, Shan Chen, Spencer Thomas, Tafadzwa L. Chaunzwa, Idalid Franco, Benjamin Kann, Shalini Moningi, Jack Qian, Madeleine Goldstein, Susan Harper, Hugo JWL Aerts, Guergana K. Savova, Raymond H. Mak, and Danielle S. Bitterman. 2024. Large language models to identify social determinants of health in electronic health records.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth

Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Tessa Han, Aounon Kumar, Chirag Agarwal, and Himabindu Lakkaraju. 2024. Towards safe large language models for medicine. *Preprint*, arXiv:2403.03744.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023a. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

L.Y. Jiang, X.C. Liu, N.P. Nejatian, et al. 2023b. Health system-scale language models are all-purpose prediction engines. *Nature*, 619:357–362.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

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

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: **phi-1.5** technical report. *arXiv preprint arXiv:2309.05463*.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. *Preprint*, arXiv:2401.17377.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Zhu Lei, and Michael Lingzhi Li. 2023. Benchmarking large language models on CMExam - a comprehensive chinese medical exam dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2023. Are emergent abilities in large language models just in-context learning? *Preprint*, arXiv:2309.01809.

Ian Magnusson, Akshita Bhagia, Valentin Hofmann, Luca Soldaini, Ananya Harsh Jha, Oyvind Tafjord, Dustin Schwenk, Evan Pete Walsh, Yanai Elazar, Kyle Lo, Dirk Groeneveld, Iz Beltagy, Hannaneh Hajishirzi, Noah A. Smith, Kyle Richardson, and Jesse Dodge. 2023. Paloma: A benchmark for evaluating language model fit. *Preprint*, arXiv:2312.10523.

National Library of Medicine. 2024. Rxnorm. `https://www.nlm.nih.gov/research/umls/rxnorm/`. Accessed: 2024-06-05.

Robert Osazuwa Ness, Katie Matton, Hayden Helm, Sheng Zhang, Junaid Bajwa, Carey E. Priebe, and Eric Horvitz. 2024. Medfuzz: Exploring the robustness of large language models in medical question answering. *Preprint*, arXiv:2406.06573.

Ankit Pal, Pasquale Minervini, Andreas Geert Motzfeldt, and Beatrice Alex. 2024. openlife-scienceai open medical llm leaderboard.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering. *Preprint*, arXiv:2203.14371.

Team Qwen. 2024. Qwen2 technical report.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR.

K. Sewell, S. Andreae, E. Luke, and M. M. Safford. 2012. Perceptions of and barriers to use of generic medications in a rural african american population, alabama, 2011. *Preventing Chronic Disease*, 9:E142.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting pretraining data from large language models. *Preprint*, arXiv:2310.16789.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson,

Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. *arXiv preprint*.

Saurabh Srivastava, Anto PV, Shashank Menon, Ajay Sukumar, Alan Philipose, Stevin Prince, Sooraj Thomas, et al. 2024. Functional benchmarks for robust evaluation of reasoning performance, and the reasoning gap. *arXiv preprint arXiv:2402.19450*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Preprint*, arXiv:2406.01574.

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. 2022. Emergent abilities of large language models. *Preprint*, arXiv:2206.07682.

Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. 2024. Benchmarking benchmark leakage in large language models. *Preprint*, arXiv:2404.18824.

Qianqi Yan, Xuehai He, Xiang Yue, and Xin Eric Wang. 2024. Worse than random? an embarrassingly simple probing evaluation of large multimodal models in medical vqa. *Preprint*, arXiv:2405.20421.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, and Judy Gichoya. 2024. Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 6(1):e12–e22.

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. 2024. A careful examination of large language model performance on grade school arithmetic. *Preprint*, arXiv:2405.00332.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don't make your llm an evaluation benchmark cheater. *Preprint*, arXiv:2311.01964.

## A Brand-Generic Pair Generation and Transformed Dataset Curation

To create the initial dataset of brand and generic drug name pairs, we used the RxNorm (National Library of Medicine, 2024) ontology, which links normalized drug names with many pharmaceutical vocabularies. We extracted combinations of brand and generic drug names using the "ingredient of" and "tradename of" relations, resulting in 2,271 unique generic names mapped to 6,961 brands. For each generic name, there are often multiple brand names. Keywords were identified using regular expressions and counted in each question within each column (questions and answer choices), and within each dataset split. Datasets with fewer than 100 instances of identified keywords in the test set were excluded from further analysis. We then carried our the following steps to arrive at our dataset of matched brand and generic drug names, and our final transformed QA datasets, as described below.

1. Two authors (SC and JG) reviewed the brand and generic names retrieved from RxNorm and removed keywords that could overlap with common text not referring to drugs. For example, some brand names such as "today" and "perform" could lead to erroneous replacement and were excluded. This resulted in 581 generic keywords mapped to 4297 unique brand keywords.

2. Regular expressions were used to to identify and replace names in the medical QA datasets to create 2 initial transformed datasets: generic-to-brand swapped in context, and brand-to-generic swapped in context.

3. Two authors (SC and JP) reviewed the resulting datasets to ensure that the replacements were done correctly.

4. Two physician authors (JG and DB) reviewed the datasets, and identified several areas of ambiguity, errors, and inconsistencies resulting from the regular expression replacement, most commonly: (1) brand names for combination medications swapped to generic names referring to single-agent medications; (2) generic names of combination medications in which a single-agent brand name was swapped for one or both of the drugs in the generic description, for example trimethoprim/sulfamethoxazole replaced with proloprim/sulfamethoxazole; (3) brand names for veterinary formulations; (4) brand names for specific formulations that did not make logical or clinical sense in context, such as brand names for topical formulations replacing descriptions of the drug administered intravenously (either explicitly or implicitly given the clinical context); (5) drug names that are also naturally occurring physiologic compounds, such as amino acids (e.g., tyrosine), vitamins (e.g., Vitamin A); endogenous hormones (e.g., insulin, thyroxine), essential elements (e.g., copper, calcium), etc; and (6) drug names that are also dietary compounds (e.g., caffeine, tryptophan). These questions were annotated, and the drug in question tracked.

5. Given the above identified errors, to facilitate expert review the drug pairs from Step 1 were provided to GPT-4o, which was prompted to check if the brand name's main component is the paired generic drug, and if the brand name drug is used mainly used for humans. 1247 brand drugs were filtered out of the keyword list. This resulted in 563 generic drug mapped to 3050 brand keywords.

6. The remaining drugs in Step 5 were provided to Cohere-RAG, which was prompted to provided a list of brand names for each generic name.

7. A physician author (JG) reviewed the retrieved brand names and selected a single brand name to pair with each generic name. This resulted in 525 generic-to-brand pairs.

8. These 525 generic-to-brand pairs were used to regenerate the transformed dataset using regular expressions to identify and replace names.

9. Two physician authors (JG and DB) reviewed the dataset generated in Step 8, and removed any remaining questions where the replacement resulted in ambiguity or inconsistencies in context to ensure quality of the final dataset.

## B Drug Knowledge

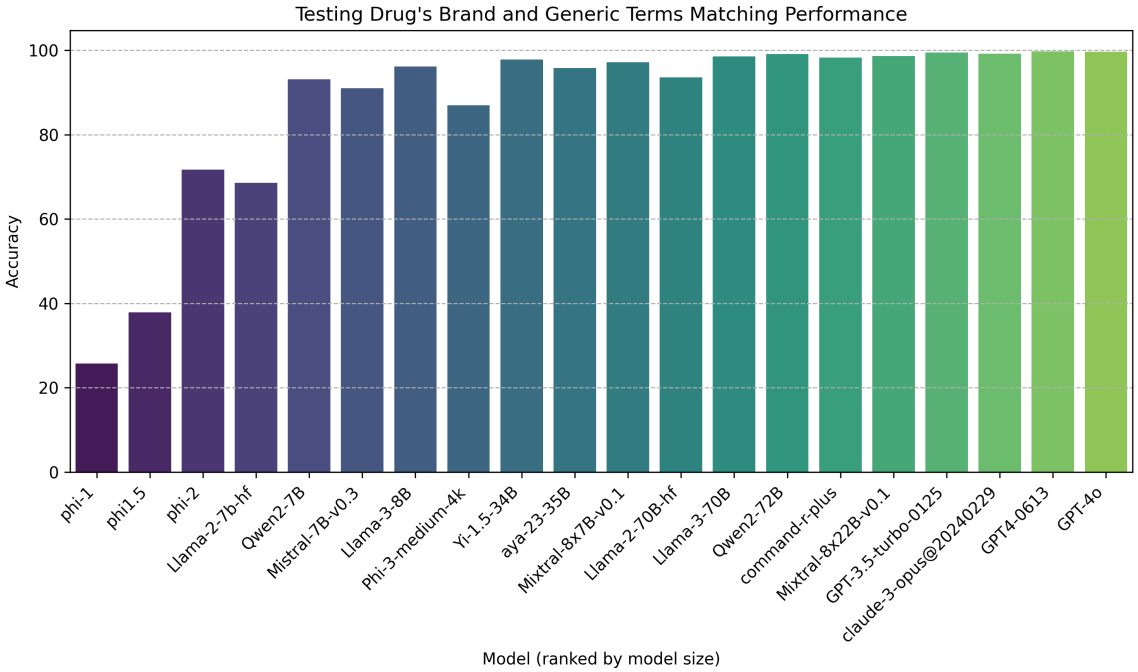


Figure 3: Performance of models on multi-choice question identification of brand-generic drug pairs ordered in increasing model size. Gemini results are missing due to Google's API safety filters.

## C Drug Swapping results

Figure 4 shows the performance change from the original filtered datasets compared to the dataset where generic drug names are swapped with the brand names (g2b). Notice that the MedMCQA (left) x-axis range is much larger than the MedQA x-axis range, indicating a larger drop across models.

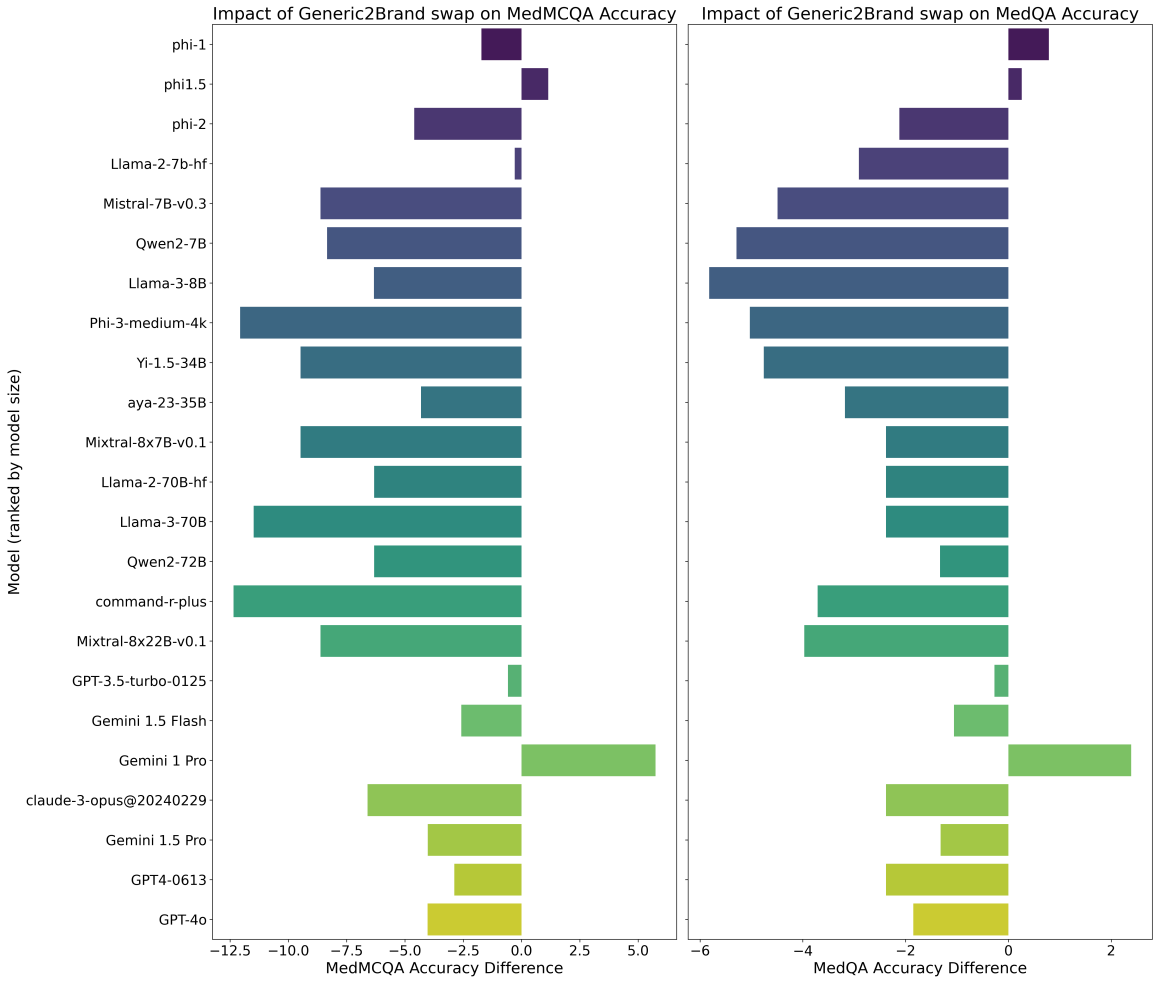


Figure 4: Performance of models on the filtered original datasets compared to the generic-to-brand versions for MedMCQA and MedQA subsets. Negative values indicate worse performance on the swapped dataset.

Table 3: List of Models Used in Experiments, June 12th, 2024

| Model Name | Size | MoE | Multi-Modal |
|---|---|---|---|
| phi-1 (Gunasekar et al., 2023) | 1.3B | No | No |
| phi-1.5 (Li et al., 2023) | 1.3B | No | No |
| phi-2 (Li et al., 2023) | 2.7B | No | No |
| phi-3-medium (Abdin et al., 2024) | 14B | No | No |
| Llama3-8B (AI@Meta, 2024) | 8B | No | No |
| Llama3-70B (AI@Meta, 2024) | 70B | No | No |
| llama-2-70B (Touvron et al., 2023) | 70B | No | No |
| llama-2-7B (Touvron et al., 2023) | 7B | No | No |
| c4ai-aya-23-35B (Aryabumi et al., 2024) | 35B | No | No |
| c4ai-r-plus | 104B | No | No |
| Mistral-7B-v0.3 (Jiang et al., 2023a) | 7B | No | No |
| mixtral-8x22B (Jiang et al., 2024) | 176B | Yes | No |
| mixtral-8x7B (Jiang et al., 2024) | 56B | Yes | No |
| qwen2-72B (Qwen, 2024) | 72B | No | No |
| qwen2-7B (Qwen, 2024) | 7B | No | No |
| yi-1.5-34B (Young et al., 2024) | 34B | No | No |
| GPT-3.5-turbo-0125 | NA | NA | No |
| GPT4-0613 | NA | NA | Yes |
| GPT-4o | NA | NA | Yes |
| Claude 3 Opus | NA | NA | Yes |
| Gemini 1 Pro | NA | NA | Yes |
| Gemini 1.5 Flash | NA | NA | Yes |
| Gemini 1.5 Pro | NA | NA | No |

Table 4: Overall and difference of Model performance on RABBITS

| Dataset | Model | g2b | original |
|---|---|---|---|
| medmcqa | GPT-3.5-turbo-0125 | 97.70 | 98.28 |
| medmcqa | GPT-4o | 86.49 | 90.52 |
| medmcqa | GPT4-0613 | 88.79 | 91.67 |
| medmcqa | Gemini 1 Pro | 73.85 | 68.10 |
| medmcqa | Gemini 1.5 Flash | 94.83 | 97.41 |
| medmcqa | Gemini 1.5 Pro | 82.47 | 86.49 |
| medmcqa | Llama-2-70B-hf | 45.98 | 52.30 |
| medmcqa | Llama-2-7b-hf | 33.91 | 34.20 |
| medmcqa | Llama-3-70B | 66.67 | 78.16 |
| medmcqa | Llama-3-8B | 52.87 | 59.20 |
| medmcqa | Mistral-7B-v0.3 | 48.28 | 56.90 |
| medmcqa | Mixtral-8x22B-v0.1 | 61.78 | 70.40 |
| medmcqa | Mixtral-8x7B-v0.1 | 55.46 | 64.94 |
| medmcqa | Phi-3-medium-4k | 60.34 | 72.41 |
| medmcqa | Qwen2-72B | 71.55 | 77.87 |
| medmcqa | Qwen2-7B | 55.17 | 63.51 |
| medmcqa | Yi-1.5-34B | 59.77 | 69.25 |
| medmcqa | aya-23-35B | 48.56 | 52.87 |
| medmcqa | claude-3-opus@20240229 | 79.89 | 86.49 |
| medmcqa | command-r-plus | 49.14 | 61.49 |
| medmcqa | phi-1 | 24.14 | 25.86 |
| medmcqa | phi-2 | 37.64 | 42.24 |
| medmcqa | phi1.5 | 31.61 | 30.46 |
| medqa 4options | GPT-3.5-turbo-0125 | 96.03 | 96.30 |
| medqa 4options | GPT-4o | 88.36 | 90.21 |
| medqa 4options | GPT4-0613 | 89.95 | 92.33 |
| medqa 4options | Gemini 1 Pro | 73.02 | 70.63 |
| medqa 4options | Gemini 1.5 Flash | 96.03 | 97.09 |
| medqa 4options | Gemini 1.5 Pro | 87.30 | 88.62 |
| medqa 4options | Llama-2-70B-hf | 52.65 | 55.03 |
| medqa 4options | Llama-2-7b-hf | 34.39 | 37.30 |
| medqa 4options | Llama-3-70B | 72.75 | 75.13 |
| medqa 4options | Llama-3-8B | 55.03 | 60.85 |
| medqa 4options | Mistral-7B-v0.3 | 48.68 | 53.17 |
| medqa 4options | Mixtral-8x22B-v0.1 | 67.46 | 71.43 |
| medqa 4options | Mixtral-8x7B-v0.1 | 60.05 | 62.43 |
| medqa 4options | Phi-3-medium-4k | 53.44 | 58.47 |
| medqa 4options | Qwen2-72B | 74.07 | 75.40 |
| medqa 4options | Qwen2-7B | 53.70 | 58.99 |
| medqa 4options | Yi-1.5-34B | 59.79 | 64.55 |
| medqa 4options | aya-23-35B | 47.88 | 51.06 |
| medqa 4options | claude-3-opus@20240229 | 83.33 | 85.71 |
| medqa 4options | command-r-plus | 56.61 | 60.32 |
| medqa 4options | phi-1 | 21.69 | 20.90 |
| medqa 4options | phi-2 | 41.80 | 43.92 |
| medqa 4options | phi1.5 | 34.92 | 34.66 |

Table 5: Overall and difference of Model performance on RABBITS

| Model | Original | g2b | Average | Difference |
|---|---|---|---|---|
| GPT-3.5-turbo-0125 | 97.29 | 96.86 | 97.08 | -0.42 |
| GPT-4o | 90.36 | 87.42 | 88.89 | -2.94 |
| GPT4-0613 | 92.00 | 89.37 | 90.69 | -2.63 |
| Gemini 1 Pro | 69.36 | 73.44 | 71.40 | 4.07 |
| Gemini 1.5 Flash | 97.25 | 95.43 | 96.34 | -1.82 |
| Gemini 1.5 Pro | 87.56 | 84.89 | 86.22 | -2.67 |
| Llama-2-70B-hf | 53.66 | 49.31 | 51.49 | -4.35 |
| Llama-2-7b-hf | 35.75 | 34.15 | 34.95 | -1.60 |
| Llama-3-70B | 76.64 | 69.71 | 73.18 | -6.93 |
| Llama-3-8B | 60.02 | 53.95 | 56.99 | -6.08 |
| Mistral-7B-v0.3 | 55.03 | 48.48 | 51.76 | -6.55 |
| Mixtral-8x22B-v0.1 | 70.92 | 64.62 | 67.77 | -6.29 |
| Mixtral-8x7B-v0.1 | 63.69 | 57.76 | 60.72 | -5.93 |
| Phi-3-medium-4k | 65.44 | 56.89 | 61.16 | -8.55 |
| Qwen2-72B | 76.64 | 72.81 | 74.72 | -3.83 |
| Qwen2-7B | 61.25 | 54.44 | 57.84 | -6.82 |
| Yi-1.5-34B | 66.90 | 59.78 | 63.34 | -7.12 |
| aya-23-35B | 51.97 | 48.22 | 50.09 | -3.75 |
| claude-3-opus@20240229 | 86.10 | 81.61 | 83.85 | -4.49 |
| command-r-plus | 60.91 | 52.88 | 56.89 | -8.03 |
| phi-1 | 23.38 | 22.91 | 23.15 | -0.47 |
| phi-2 | 43.08 | 39.72 | 41.40 | -3.36 |
| phi1.5 | 32.56 | 33.27 | 32.91 | 0.70 |

## D Pre-training data screening

Table 6: Statistics for occurrence counts of selected brand and generic terms among popular pre-training datasets

| Subset | Terms | Average | Median | Std. Dev |
|---|---|---|---|---|
| Dolma | Generic | 564,151 | 136,682 | 2,399,928 |
| | Brand | 234,138 | 698 | 2,543,075 |
| Red Pajama | Generic | 161,227 | 42,549 | 620,393 |
| | Brand | 29,561 | 84 | 232,661 |
| Pile train | Generic | 96,309 | 28,074 | 325,307 |
| | Brand | 4,757 | 19 | 43,613 |
| C4 train | Generic | 27,973 | 5,454 | 144,162 |
| | Brand | 9,941 | 26 | 96,504 |

## E Biomedical-Supervised Fine-Tuned (SFT) Models

The top 3 accessible models on the leaderboard (all higher than GPT-4 and Med-Palm as of June 1st, 2024) (Pal et al., 2024) are all Llama3 SFT variants. The Llama3-8B-sft2 model achieved the highest original score of 0.85 but showed a performance drop of -0.15 in the g2b category, similar to Llama3-8B-sft1. Both performance decreases are greater than those observed in the base model version. Similar patterns were seen in the Llama-70B models. These degradations among benchmark datasets may be helpful in inspecting SFT models that are over-fitted.

Table 7: Llama-3 Vanilla v.s its Fine-Tuned Variants

| Model | None ↑ | g2b ↑ | $\delta \downarrow$ |
|---|---|---|---|
| llama-3-8B (vanilla) | 0.60 | 0.53 | -0.07 |
| llama-3-8B-sft1 | 0.80 | 0.66 | -0.14↑ |
| llama-3-8B-sft2 | 0.85 | 0.70 | -0.15↑ |
| llama-3-70B (vanilla) | 0.77 | 0.70 | -0.07 |
| llama-3-70B-sft1 | 0.75 | 0.66 | -0.09↑ |