Journal of Data-centric Machine Learning Research (2024) Submitted 3/24; Revised 4/24; Published 5/24

,

Birbal: An efficient 7B instruct-model fine-tuned with curated datasets

Ashvini Jindal AJINDAL@LINKEDIN.COM
LinkedIn AI, USA

Pawan Kumar Rajpoot PAWANKUMAR.RAJPOOT@DATA-X.AI
SCB DataX, Thailand

Ankur Parikh ANKUR.PARIKH85@GMAIL.COM
UtilizeAI Research, India

Reviewed on OpenReview: *https://openreview.net/forum?id=wGuflAGXSb*

Editor: My editor

Abstract

LLMOps incur significant costs due to hardware requirements, hindering their widespread accessibility. Additionally, a lack of transparency in model training methods and data contributes to the majority of models being non-reproducible. To tackle these challenges, the LLM Efficiency Challenge was introduced at NeurIPS Workshop[1], aiming to adapt foundation models on a diverse set of tasks via fine-tuning on a single GPU (RTX 4090 or A100 with 40GB) within a 24-hour timeframe. In this system description paper, we introduce **Birbal**, our Mistral-7B based winning model, fine-tuned on a single RTX 4090 for 16 hours. Birbal's success lies in curating high-quality instructions covering diverse tasks, resulting in a 35% performance improvement over second-best Qwen-14B based submission.

Keywords: NeurIPS LLM Efficiency Challenge, Data Curation, Instruction Tuning, QLoRA, Super-NaturalInstructions, Mistral

1 Introduction

Few-shot Large Language Models (LLMs) have excelled in various NLP tasks, from standardized exams OpenAI et al. (2023); Ahmed et al. (2023); Singhal et al. (2022) to coding challenges and chatbots Singhal et al. (2022); HANS. Typically, this involves fine-tuning an LLM with associated task(s) examples. However, the costs of fine-tuning and querying LLMs to perform new tasks are large due to the expensive and often proprietary hardware used to train and serve these models. Given these costs, access to performant LLMs has been gated, making them inaccessible to those without substantial resources.

1. `https://llm-efficiency-challenge.github.io`

Despite the rise of open-source LLMs like Llama-2 Touvron et al. (2023), Falcon Almazrouei et al. (2023), Qwen Bai et al. (2023), and Mistral Jiang et al. (2023), the field encounters challenges in reproducibility and transparency. Many LLMs release partial artifacts, offering only final model weights or inference code, hindering comprehensive disclosure of training methodologies and impeding researchers' ability to replicate reported results. For instance, Llama Touvron et al. (2023) discloses training data mixtures, but the absence of data processing and training code impedes full reproducibility, as observed in the case of RedPajama Computer (2023), an open reproduction of Llama's data.

To address the lack of transparency in model training and democratize access to cutting-edge LLMs, a LLM efficiency challenge[2] was introduced at the NeurIPS Workshop. This challenge required participants to fine-tune an open-source foundation model on a single GPU (RTX 4090 or A100 with 40GB) within a 24-hour timeframe. In this paper, we introduce **Birbal**, our Mistral-7B based winning model, fine-tuned with high-quality instructions covering diverse tasks on a single RTX 4090 for 16 hours.

2 LLM Efficiency Challenge

The LLM Efficiency Challenge tasks participants with fine-tuning an open-source "base" language model on a single GPU (RTX4090 or A100 40GB) for 24 hours, exclusively using open-source data. The competition has two hardware tracks: the NVIDIA 4090 track and the NVIDIA A100 track. Accepted base models must be open and without instruction-tuning, adhering to licenses like MIT, Apache 2, BigScience RAIL, and Llama-2 Community License Agreement. Participants can use various standard autoregressive and autoencoder base models and all open-source datasets, including Databricks-Dolly-15 Conover et al. (2023), OpenAssistant Conversations Dataset (oasst1) Köpf et al. (2023), The Flan Collection Longpre et al. (2023), AllenAI Dolma Soldaini et al. (2023), RedPajama-Data-1T Computer (2023), and LIMA Zhou et al. (2023), are allowed, emphasizing avoidance of datasets with generated content unless explicitly permitted by the source model's license. Each team can submit three entries per track, and post-competition, winning models, code, and data must be open-sourced. The evaluation consists of four stages, where submissions are assessed on a set of tasks, and rankings are determined by the geometric mean across all evaluation tasks. Submissions below a score threshold are eliminated in stages 1 and 2. Stage-1 evaluates submissions on a subset of HELM Lee et al. (2023b) tasks (open eval), and stage-2 assesses them on a hidden evaluation set (closed eval). In the third stage, organizers reproduce training artifacts for consistency, and in the final stage, submissions are evaluated and ranked on a subset of open and closed tasks' performance.

3 Our Approach

3.1 Design Choices

We participated in the RTX 4090 track of the competition. In this section, we outline our design choices based on the aforementioned constraints:

2. `https://llm-efficiency-challenge.github.io/challenge`

- **Data Sources** - Evaluation included both HELM and hidden tasks; to excel in the latter, minimize reliance on HELM-sourced data.
- **Mistral-7B vs Qwen-14B**: Within 24GB memory budget, Mistral-7B and Qwen-14B were best performing models outperforming Llama-2 13B on several benchmarks[3].
- **High-Quality Data vs. Hardware Optimization**: We can optimize performance through kernel optimizations or prepare a high-quality dataset for fine-tuning.
- **Dataset Curation vs Generation:** The success of recent LLM-generated datasets like Stanford Alpaca Taori et al. (2023) is promising. However, only open-source base models can be used for dataset generation. We can curate a high-quality dataset from existing sources or generate from base LLMs.

3.2 Strategy

The spirit of the competition was to create a model that works well on diverse tasks. Based on this, we chose Mistral-7B base model Jiang et al. (2023) to fit more high-quality instructions covering multiple tasks. Due to our practical exposure in hardware optimization, we focused on high-quality dataset construction. Moreoever, we chose to curate existing datasets as generating datasets with a relatively small model (7B) can be tricky.

3.3 Data Curation

Our dataset curation methodology was geared toward obtaining various datasets spanning a broad spectrum of tasks. Given the constraints of 24GB memory and 24-hour fine-tuning limit, we determined that 200K, 400K, and 700K size datasets can be fine-tuned for three epochs, two epochs, and one epoch, respectively. Our data curation method is explained below. Table 1 shows a summary of the final datasets.

- **LIMA** Zhou et al. (2023) – This is a set of 1,000 well-crafted prompts and responses utilized by the LIMA model. We added all these data points in our final datasets.
- **Open-Platypus** Lee et al. (2023a) – A subset from various open datasets employed in Platypus models. We excluded 10% of the dataset containing GPT-generated instructions to satisfy the challenge constraint. We added the remaining data points in our final datasets.
- **Natural Instructions** Mishra et al. (2022); Wang et al. (2022) – The Natural Instructions (NI) dataset is an extensive assemblage of over 1,600 tasks, each defined through natural language instructions, and includes crucial metadata such as task outline, domain, category, and input/output languages. We sample examples from NI dataset based on strategy described later.
- **Other datasets** – In addition to above datasets, we randomly sampled examples from HELM training datasets: **OpenbookQA** Mihaylov et al. (2018), **QUAC** Choi et al. (2018), and **CNN/DailyMail** See et al. (2017); Hermann et al. (2015). To bolster the model's mathematical reasoning capabilities, we randomly sampled examples from the **MathInstruct** Xiang Yue (2023) dataset, post exclusion of LLM-generated examples.

Our NI dataset curation process, focused on winning 200k dataset, consists of four stages:

3. `https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard`

1. **Tasks subset selection** - We selected a subset of 463 tasks[4] from the total pool of 1600+ tasks. Tasks featuring non-English inputs/outputs were eliminated, resulting in the exclusion of 576 tasks. We also disregarded tasks from the MMLU benchmark in the Question Answering category, as usage of the MMLU dataset was not allowed in the competition. Tasks falling under Question Generation and Question Understanding categories were excluded in favor of focusing on answer generation tasks. Tasks in the Wrong Candidate Generation and math categories were also removed. Additionally, tasks related to linguistic aspects such as PoS tagging, Keyword Tagging, Named Entity Recognition, Coreference Resolution, Word Semantics, Linguistic Probing, and Paraphrasing were filtered out, considering the inherent strength of most LLMs in linguistics and general text understanding tasks. Consequently, the chosen tasks spanned 33 categories within the NI dataset, heavily weighted toward the more prevalent categories, encompassing tasks related to Question Answering, Sentiment Analysis, Program Execution, Toxic Language Detection, and others.
2. **Task Categorization** - Each of the 463 selected tasks is manually categorized as "Exact Match" or "Generation", based upon the output characteristics of the task.
3. **Few-Shot Inference** - For quantitative assessment of tasks performance, we performed few-shot inference on Mistral-7B base model to direct controlled generation. We used Accuracy for "Exact Match" tasks and ROUGE score for "Generation" tasks.
4. **Sampling** - For 200k dataset, We sampled 50K examples from both "Exact Match" and "Generation" tasks. For "Exact Match" tasks, first, we removed low-accuracy tasks as those tasks might be too difficult for the model to learn within our constraints. Then, we bucketed tasks based on their accuracy. Next, we sample examples from each task in a bucket. Specifically, we sample more examples from lower accuracy tasks and vice-versa. Finally, we randomly selected 50K examples from the aggregated pool. For "Generation" tasks, we bucketed examples from each task based on the ROUGE score. The buckets were [0,0.2), [0.2,0.3), [0.3,0.4), [0.4,0.5), [0.5,0.6), [0.6,0.7), and [0.7,0.8). For each task, we randomly sampled 40% examples from the bucket [0,0.2] and 10% examples from the remaining buckets. Finally, we randomly selected 50K examples from the aggregated pool.

Source Dataset	**200K**	**400K**	**700K**
LIMA	1K	1K	1K
Open-Platypus	25K	25K	25K
NI (Exact Match)	50K	110K	220K
NI (Generation)	50K	110K	220K
OpenQA	5K	5K	5K
QUAC	10K	10K	10K
CNN/DailyMail	15K	28k	28k
MathInstruct	50K	100K	200K

Table 1: Our curated datasets[5]. NI refers to Natural-Instructions dataset.

4. https://github.com/Upaya07/NeurIPS-llm-efficiency-challenge/blob/main/selected_NI_tasks
5. https://github.com/Upaya07/NeurIPS-llm-efficiency-challenge#birbal-models-and-datasets

3.4 Fine-Tuning

Due to memory and fine-tuning time constraints, we applied 4-bit QLoRA Dettmers et al. (2023) to fine-tune the Mistral-7B base model. To meet the time limit, we conducted fine-tuning for ∼3 epochs on the 200K dataset, ∼2 epochs on the 400K dataset, and ∼1 epoch on the 700K dataset. We randomly sampled 2000 examples from fine-tuning dataset as validation set. For LoRA, we set the rank to 128 and alpha to 256. We apply LoRA to all Query, Key, and Value metrics in multi-head self-attention blocks alongside Linear layers. Following NEFTune Jain et al. (2023), random noise was introduced into embeddings. We set gradient accumulation steps to 3 with the micro-batch size of 2 to simulate a larger batch size. We used paged_adamw_32bit optimizer with a cosine schedule with a learning rate of 2e-5. We set the decay rate to 0.01 and warmup steps to 100. Additionally, we enabled sample packing to enhance fine-tuning efficiency. All fine-tuning experiments were conducted using axolotl[6] library. After fine-tuning for 24 hours, we selected a checkpoint based on minimum validation loss and used it for final submissions. We submitted three fine-tuning models on 200K, 400K, and 700K size datasets, respectively.

4 Evaluation

In the initial evaluation stage (Open Eval), all submissions underwent assessment on a subset of HELM tasks, featuring test examples sourced from datasets like MMLU Hendrycks et al. (2021b,a), TruthfulQA Lin et al. (2022), BBQ Parrish et al. (2022), GSM8K Cobbe et al. (2021), and Big-bench bench authors (2023). Among the 57 submissions in the 4090 track, 30 qualified for the subsequent stage based on a predefined threshold. From our three submissions (Birbal-200k, Birbal-400k, Birbal-700k), Birbal-200K secured the 20th rank with a score of 0.64, while the other two did not progress to the second evaluation stage. In the second stage, the 30 selected submissions were evaluated on hidden tasks, leading to the selection of 10 teams for the model reproducibility stage. Test examples were drawn from datasets like SAMSum Gliwa et al. (2019), Corr2cause Jin et al. (2023), MATH Hendrycks et al. (2021c), and ETHICS Hendrycks et al. (2021a). Our team achieved the 1st rank with a score of 0.660. The model was successfully reproduced in the third stage, and in the final evaluation stage, a new subset was formed from datasets in the first and second stages. The final score was computed as a weighted sum of open and closed eval scores, with 1/3 and 2/3 weights, respectively. The scores of the top 3 teams are detailed in Table 2.

Dataset	Metric	Stage	Top-3 Teams[7] Scores		
			Birbal* **(Ours)**	Rank-2$	Rank-3*
MMLU	EM(Accuracy)	Open	0.63	0.69	0.64
	EM(Robustness)	Open	0.59	0.64	0.60
	EM(Fairness)	Open	0.60	0.65	0.60
	MWR	Open	0.42	**0.87**	0.46

6. https://github.com/OpenAccess-AI-Collective/axolotl
7. https://llm-efficiency-challenge.github.io/leaderboard

TruthfulQA	EM(Accuracy)	Open	0.59	0.52	0.57
	EM(Robustness)	Open	0.54	0.52	0.52
	EM(Fairness)	Open	0.49	0.44	0.46
	MWR	Open	**0.75**	0.28	0.56
BIG-bench	EM(Accuracy)	Open	0.33	0.38	0.0
	MWR	Open	0.75	**0.87**	0.06
GSM8K	EM(Accuracy)	Open	0.44	0.57	0.0
	MWR	Open	0.62	**0.81**	0.03
BBQ	EM(Accuracy)	Open	0.74	0.85	0.93
	MWR	Open	0.25	0.56	**0.75**
sam_sum	ROUGE-2	Closed	0.13	0.03	0.10
	Stereotypes(race)†	Closed	0.67	-	0.67
	Stereotypes(gender)†	Closed	0.45	0.42	0.34
	Representation(race)†	Closed	0.46	0.62	0.38
	Representation(gender)†	Closed	0.01	0.0	0.01
	MWR	Closed	**0.38**	0.21	0.65
corr2cause	EM(Accuracy)	Closed	0.61	0.47	0.50
	MWR	Closed	**0.87**	0.25	0.62
MATH	chain-of-thoughts	Closed	0.12	0.07	0.05
	MWR	Closed	**0.75**	0.5	0.25
ethics_j	EM(Accuracy)	Closed	0.68	0.68	0.70
	EM(Robustness)	Closed	0.64	0.66	0.65
	EM(Fairness)	Closed	0.62	0.58	0.64
ethics_c	EM(Accuracy)	Closed	0.41	0.52	0.49
	EM(Robustness)	Closed	0.33	0.45	0.42
	EM(Fairness)	Closed	0.34	0.5	0.45
ethics_v	EM(Accuracy)	Closed	0.89	0.77	0.74
	EM(Robustness)	Closed	0.86	0.70	0.67
	EM(Fairness)	Closed	0.86	0.69	0.69
ethics_d	EM(Accuracy)	Closed	0.63	0.58	0.60
	EM(Robustness)	Closed	0.58	0.49	0.52
	(Fairness)	Closed	0.59	0.53	0.49
ethics_u	EM(Accuracy)	Closed	0.72	0.55	0.56
	EM(Robustness)	Closed	0.60	0.34	0.45
	EM(Fairness)	Closed	0.64	0.40	0.52
ethics	**MWR**	Closed	**0.55**	0.41	0.47
Open Eval Score			0.52	**0.63**	0.21
Closed Eval Score			**0.61**	0.32	0.47
Final Score			**0.58**	0.42	0.38

Table-2: Comparative Analysis of Top-3 Models' Overall Performance on Open and Closed Tasks. *Open Eval Score* and *Closed Eval Score* for each submission are derived as the geometric mean of mean win rates across tasks in the *Open* and *Closed* evaluation stages, respectively. The *Final Score* is computed as a weighted sum of the *Open Eval Score* (weighted at 1/3) and *Closed Eval Score* (weighted at 2/3). *ethics_justice*, *ethics_commonsense*, *ethics_virtue*, *ethics_deontology*, and *ethics_utilitarianism* are denoted as *ethics_j*, *ethics_c*, *ethics_v*, *ethics_d*, and *ethics_u*, respectively. *Birbal* is fine-tunined model on 200k dataset. *MWR* refers to Mean Win Rate. *EM* refers to Exact Match. † refers to lower is better. (* = Mistral-7B as base model, $ = Qwen-14B as base model)

Dataset	Metric	Stage	Mistral-7B	Model Variants		
				Birbal (200k)	Birbal (400k)	Birbal (700k)
MMLU	EM(Accuracy)	Open	**0.64**	0.63	0.62	0.62
	EM(Robustness)	Open	**0.60**	0.59	0.57	0.58
	EM(Fairness)	Open	0.59	**0.60**	0.58	0.59
TruthfulQA	EM(Accuracy)	Open	0.56	**0.59**	0.41	0.46
	EM(Robustness)	Open	0.44	**0.54**	0.38	0.43
	EM(Fairness)	Open	0.44	**0.49**	0.34	0.39
BIG-bench	EM(Accuracy)	Open	0.29	0.33	**0.38**	0.37
GSM8K	EM(Accuracy)	Open	0.33	0.44	**0.61**	0.56
BBQ	EM(Accuracy)	Open	**0.80**	0.74	0.67	0.64
sam_sum	ROUGE-2	Closed	0.14	0.13	**0.16**	**0.16**
	Stereotypes(race)†	Closed	-	0.67	-	-
	Stereotypes(gender)†	Closed	0.32	0.45	**0.30**	0.35
	Representation(race)†	Closed	**0.33**	0.46	**0.33**	**0.33**
	Representation(gender)†	Closed	0.05	0.01	0.01	**0.0**
corr2cause	EM(Accuracy)	Closed	0.46	**0.61**	0.55	0.56
MATH	chain-of-thoughts	Closed	0.02	**0.12**	**0.12**	**0.12**
ethics_j	EM(Accuracy)	Closed	0.69	0.68	**0.71**	**0.71**
	EM(Robustness)	Closed	0.64	0.64	0.68	**0.69**
	EM(Fairness)	Closed	0.61	0.62	0.61	**0.65**
ethics_c	EM(Accuracy)	Closed	**0.51**	0.41	0.40	0.43
	EM(Robustness)	Closed	0.36	0.33	0.36	**0.39**
	EM(Fairness)	Closed	**0.42**	0.34	0.36	0.39
ethics_v	EM(Accuracy)	Closed	0.68	**0.89**	0.79	0.81
	EM(Robustness)	Closed	0.58	**0.86**	0.77	0.77
	EM(Fairness)	Closed	0.56	**0.86**	0.76	0.77
ethics_d	EM(Accuracy)	Closed	0.64	0.63	**0.67**	0.65
	EM(Robustness)	Closed	0.51	0.58	**0.59**	**0.59**

	(Fairness)	Closed	0.54	0.59	**0.60**	**0.60**
ethics_u	EM(Accuracy)	Closed	0.61	**0.72**	0.66	0.64
	EM(Robustness)	Closed	0.42	**0.60**	0.58	0.59
	EM(Fairness)	Closed	0.49	**0.64**	0.58	0.57

Table-3: Overall performance of Birbal models fine-tuned on different dataset sizes vs Mistral-Base-7B on open and closed eval. Best scores are marked in **bold**. † refers to lower is better.

Though our 400k and 700k submissions could not make it to the second stage in the competition, we have benchmarked all submissions on all evaluation scenarios in Table-3 for detailed analysis. Birbal-200k, Birbal-400k, and Birbal-700k models were fine-tuned for 3, 2, and 1 epoch(s), respectively. There are a total of 31 evaluations: 9 Open and 22 Closed evaluations. Mistral-7B base model scored best in 3 open and 3 closed evaluations. Birbal-200k scored best in 4 open and 8 closed evaluations. Birbal-400k scored best in the 2 open and 8 closed evaluations. Bibral-700k scored best in 10 closed evaluations. During fine-tuning, adding more data points led to a drop in performance in a number of open tasks. However, performance on closed tasks improves as we scale a number of data points.

5 Conclusion

This paper describes fine-tuning the base Mistral-7B model on our curated subset of existing datasets on RTX 4090 (24 GB) GPU for one day. The fine-tuned model was evaluated on various tasks and outperformed other submissions by more than 35%.

Broader Impact Statement

This work addresses the challenge of adapting an LLM with only 1 GPU (24GB or 40GB memory) within a day. So, this type of approach has the potential to make an efficient LLM fine-tuning accessible to those without substantial resources. Our open-source model was developed by fine-tuning the Mistral-7B model on a subset of datasets without alignment. So, it contains certain forms of bias (e.g., the risk of social stereotypes, discrimination and exclusion, and the risk of under-representing certain languages or domains) that are already present in the base model and source datasets.

Acknowledgments and Disclosure of Funding

We would like to thank Lambda Labs[8] for providing the compute resources required for post-competition analysis.

8. `https://lambdalabs.com/`

6 Reproducibility

Failure to reproducibility was one of the criteria to eliminate the submissions from the competition. Our winning model was reproduced successfully. Our dataset curation mechanism, fine-tuning scripts, and models are publicly available[9].

References

Imtiaz Ahmed, Ayon Roy, Mashrafi Kajol, Uzma Hasan, Partha Protim Datta, and Md Rokonuzzaman Reza. Chatgpt vs. bard: a comparative study. *Authorea Preprints*, 2023.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. The falcon series of open language models, 2023.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

BIG bench authors. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=uyTL5Bvosj.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1241. URL https://www.aclweb.org/anthology/D18-1241.

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.

Together Computer. Redpajama: an open dataset for training large language models, 2023. URL https://github.com/togethercomputer/RedPajama-Data.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's

9. https://github.com/Upaya07/NeurIPS-llm-efficiency-challenge

first truly open instruction-tuned llm, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In Lu Wang, Jackie Chi Kit Cheung, Giuseppe Carenini, and Fei Liu, editors, *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5409. URL https://aclanthology.org/D19-5409.

Felix HANS. Chatgpt vs. bard–which is better at solving coding problems?

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021a.

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)*, 2021b.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021c.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701, 2015. URL http://papers.nips.cc/paper/5945-teaching-machines-to-read-and-comprehend.

Neel Jain, Ping yeh Chiang, Yuxin Wen, John Kirchenbauer, Hong-Min Chu, Gowthami Somepalli, Brian R. Bartoldson, Bhavya Kailkhura, Avi Schwarzschild, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Neftune: Noisy embeddings improve instruction finetuning, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. Can large language models infer causation from correlation?, 2023.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi,

et al. Openassistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*, 2023.

Ariel N. Lee, Cole J. Hunter, and Nataniel Ruiz. Platypus: Quick, cheap, and powerful refinement of llms. 2023a.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Benita Teufel, Marco Bellagente, Minguk Kang, Taesung Park, Jure Leskovec, Jun-Yan Zhu, Li Fei-Fei, Jiajun Wu, Stefano Ermon, and Percy Liang. Holistic evaluation of text-to-image models, 2023b.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. The flan collection: Designing data and methods for effective instruction tuning, 2023.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In *ACL*, 2022.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey,

Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2023.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. BBQ: A hand-built bias benchmark for question answering. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.165. URL `https://aclanthology.org/2022.findings-acl.165`.

Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of*

the Association for Computational Linguistics (Volume 1: Long Papers), pages 1073–1083, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1099. URL https://www.aclweb.org/anthology/P17-1099.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022.

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, 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, Evan Pete Walsh, Hannaneh Hajishirzi, Noah A. Smith, Luke Zettlemoyer, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: An Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. *arXiv preprint*, 2023.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. Super-naturalinstructions:generalization via declarative instructions on 1600+ tasks. In *EMNLP*, 2022.

Ge Zhang Yao Fu Wenhao Huang Huan Sun Yu Su Wenhu Chen Xiang Yue, Xingwei Qu. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*, 2023.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. Lima: Less is more for alignment, 2023.