The Oracle Has Spoken: A Multi-Aspect Evaluation of Dialogue in Pythia

Zixun Chen
Columbia University
New York, NY
zc2738@columbia.edu

Petr Babkin
J.P. Morgan AI Research
Palo Alto, CA
petr.babkin@jpmorgan.com

Akshat Gupta
University of California, Berkeley
Berkeley, CA
akshat.gupta@berkeley.edu

Gopala Anumanchipalli
University of California, Berkeley
Berkeley, CA
gopala@berkeley.edu

Xiaomo Liu
J.P. Morgan AI Research
New York, NY
xiaomo.liu@jpmchase.com

Abstract

Dialogue is one of the landmark abilities of large language models (LLMs). Despite its ubiquity, few studies actually distinguish specific ingredients underpinning dialogue behavior emerging during post-training. We employ a comprehensive suite of model-based metrics, each targeting a distinct fine-grained aspect of dialogue, motivated by linguistic theory. We evaluate how the performance of pre-trained Pythia models changes with respect to each of those dimensions, depending on model size and as a result of supervised fine-tuning on conversational datasets. We observe only a mild impact of raw model size on most metrics, whereas fine-tuning quickly saturates the scores for all but the smallest models tested. Somewhat contrary to our expectations, many metrics show very similar trends, especially if they are all rooted in the same evaluator model, which raises the question of their reliability in measuring a specific dimension. To that end, we conduct additional analyses of score distributions, metric correlations, and term frequencies in generated responses to help explain our observations.

1 Introduction

Large Language Models (LLMs) took the world by storm, in part, thanks to their ability to engage in naturalistic dialogue. While the acquisition of general language fluency and world knowledge by LLMs is commonly attributed to the phase of large scale self-supervised pre-training (Brown et al., 2020), pinpointing the origins of specific abilities, such as dialogue, remains a challenge. Although it is tempting to take dialogue abilities of LLMs for granted, as naturally emerging from pre-training, there is a well documented behavior of raw base models responding to a prompt by reproducing its likely continuation in the pre-training corpus rather than conversing in a way a human person would[1]. A comparatively small-scale phase of post-training, which includes supervised fine-tuning on human-annotated prompt/response pairs, proved essential to ensuring alignment with user intent and preferences — including making them more conversational (Ouyang et al., 2022). Fine-tuning is a well known technique for improving downstream task performance. Due to its significantly smaller scale, fine-tuning merely "activates" skills and knowledge already learned during pre-training, and aligns them with user preferences; and its effectiveness is upper bounded by the base model's size

[1] https://openai.com/index/instruction-following/

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Evaluating the Evolving LLM Lifecycle.

(Zhou et al., 2023). Similarly, LLM performance on downstream tasks is largely determined by the amount of pre-training of the base model (Gudibande et al., 2023). Furthermore, fine-tuning of smaller models, using imitation data from strong LLMs, teaches them to mimic the style of larger models but is generally unable to compensate for limited pre-training knowledge.

Whether or not an LLM's ability to engage in a conversation merely boils down to stylistic adaptation or constitutes a skill in and of itself, it is clear from linguistic theory that dialogue is predicated on the ability to recognize intent, infer discourse relations between utterances, and to keep track of the evolving state of the conversation (Horn and Ward, 2004).

In this work, we investigate the effects of conversational finetuning on an open source LLM Pythia (Biderman et al., 2023) of five different sizes through the lens of these fine-grained linguistic dimensions. We contribute an empirical analysis of changes in model's text generation behavior, as a result of finetuning on 3 chat datasets: Databricks Dolly[2], Open Assistant[3], and ShareGPT[4], with respect to dialogue-specific metrics: UniEval (Zhong et al., 2022), Themis (Hu et al., 2024), and a targeted GPT-4-based assessment.

2 Related Work

Several works previously evaluated LLMs fine-tuned on conversational data. We summarize them below, while focusing on differences from our work.

Alghisi et al. (2024) evaluated different LLM adaptation techniques: in-context learning, fine-tuning and RAG across open domain, knowledge-grounded and task-based dialogue. They find fine-tuning to provide superior lift compared to ICL while retrieval augmentation moderately improves both ICL and SFT. Notably, they used chat/instruct variants of Llama (Touvron et al., 2023) and Mistral (Jiang et al., 2023) without controlling for model size. Whereas, we start with base models of different sizes and directly fine-tune them on each dataset without relying on chat/instruct variants. Additionally the use of perplexity for evaluation is limiting, as it gives an overall measure of gold text's uncertainty under the model, whereas we look at fine-grained linguistic dimensions with greater interpretability.

Mousavi et al. (2022) also conducted dialogue fine-tuning experiments with T5 (Raffel et al., 2023) and GPT-2 (Radford et al., 2019) in Italian. Their work emphasized human evaluation of dimensions such as appropriateness, contextualization, and grammar correctness, which are more informative than perplexity but are quite ad hoc (e.g., "Genericness") and are not based on any concrete linguistic phenomena. Despite our evaluation using automatic metrics, they are all based on models specifically trained to reproduce human judgements on text annotation tasks.

DialogBench (Ou et al., 2024) is a synthetic dataset of dialogue-related tasks encompassing intent recognition, knowledge grounded generation and coherent infilling, rendered as multiple-choice questions. By contrast, we use organic conversations for which we simultaneously compute multiple metrics each targeting a different aspect of dialogue. Authors evaluated a large number of pre-trained and chat/instruction-tuned LLMs from different families and sizes and noted base LLMs do well on correctness-related tasks but struggle with coherence and safety. Our evaluation additionally includes analyses of metric reliability.

3 Methodology

We use the Pythia family of models (Biderman et al., 2023) as the testbed for our fine-tuning experiments. These decoder-only LLMs are pre-trained on 1 trillion tokens of the Pile corpus and come in 10 model sizes, ranging from 14 million to 12 billion parameters — all trained on data presented in the same order to facilitate reproducibility research. To reasonably cover the configuration space, we select the following checkpoints: 140m, 410m, 1.3b, 2.8b, 6.9b (all dedup).

To ensure diversity of the the fine-tuning datasets, we use a combination of single and multi-turn corpora authored by human annotators, as well as synthesized by ChatGPT: Dolly, Open Assistant, ShareGPT. We sample 10k conversations from each.

[2] `https://huggingface.co/datasets/databricks/databricks-dolly-15k`
[3] `https://huggingface.co/datasets/OpenAssistant`
[4] `https://huggingface.co/datasets/anon8231489123/ShareGPT_Vicuna_unfiltered`

3.1 Evaluation Metrics

Evaluation of dialogue is inherently challenging due to a large space of valid responses to any given utterance. Rather than capturing general quality of generated responses (which is expected to improve through fine-tuning on any corpus), our goal is to disentangle specific linguistic dimensions that contribute to high-quality dialogue. While human evaluation is considered a gold standard for assessing LLMs, it comes with its own set of challenges, including bias, subjectivity, low reproducibility, and high cost. After an extensive survey of automatic metrics, we adopt the following 3 model-based metrics that most closely reflect qualities characteristic of dialogue.

UniEval (Zhong et al., 2022) is based on the T5 model, finetuned in the to predict entailment between a text and a set of interpretable attributes, which for dialogue include: Naturalness, Coherence, Engagingness, Groundedness, and Understandability. Depending on a metric, it can be computed either on the evaluated model's output alone (e.g., Naturalness and Understandability), model's output in conjunction with either the prompt (e.g., Coherence), or an optional context (e.g., Groundedness), and finally all 3 (e.g., Engagingness).

Themis (Hu et al., 2024) is a more recent model-based evaluation created by finetuning Llama3-8B (Grattafiori et al., 2024) on 58 datasets, encompassing 9 generative language tasks, with additional preference alignment and multi-perspective consistency validation. Themis evaluates the following 4 aspects for dialogue: Context Maintenance, Interestingness, Knowledge Use, Naturalness. In addition to scores on the 5-point scale, Themis provides reviews to back up its ratings.

GPT-4-as-a-judge. Evaluation through prompting a strong LLM, such as GPT-4 (OpenAI et al., 2024) has a emerged as a new paradigm in NLP. We use this as an opportunity to fine-tune evaluated dimensions to probe nuanced pragmatic aspects of dialogue not usually captured in generic evaluations. Specifically, we instructed GPT-4 to assess how well our LLMs are able to follow dialogue turn taking conventions, recognize user's intention, generate coherent responses, and keep track of discourse referents. The complete prompt used in the FastChat format[5] is included in the appendix.

We manually examined 5-10 lowest and highest scoring examples from the datasets for each metric, concluding they reasonably captured our intended evaluation dimensions (see Sections A.5–A.7). We conclude by and large UniEval, Themis, and GPT-4-based metrics are reliable indicators of generation quality with respect to the evaluated dimensions. Nonetheless, it should be noted, a high score in any of these metrics does not guarantee the generations to be completely sensible or free of obvious hallucinations. In addition, we have come across a small number of clearly incoherent examples that still obtained high scores in UniEval.

Open LLM Leaderboard. While probing the models for dialogue specific competencies, we also track the average of the common NLP benchmarks, such as MMLU, TruthfulQA, and Winogrande used in the Open LLM Leaderboard[6] for ease of comparison — to serve as the base level of model's task-specific performance (Appendix A.8).

4 Results and Discussion

We have run finetuning experiments using the Lit-GPT library[7] on two NVIDIA RTX A6000's for model sizes from 160 million to 2.8 billion parameters, and four A6000's for the 6.9 billion parameter model. PyTorch Fully-Sharded Data Parallelism (Zhao et al., 2023) is enabled for the finetuning of 2.8b and 6.9b parameter models, allowing us to partition weights between GPUs. Depending on the base model size, we used a learning rate of 3×10^{-5} for the 160m model, 1×10^{-5} for the 410m model, 1×10^{-6} for all other models with the AdamW (Loshchilov and Hutter, 2019) optimizer, which we empirically verified to result in best validation set performance across all datasets. We used a batch size of 128 and fine-tuned for 10 epochs for each experiment, keeping the checkpoint with lowest validation loss as well as the base and final checkpoints for verification. We initially considered parameter-efficient finetuning using the adapter approach (Hu et al., 2023), however it did

[5] https://github.com/lm-sys/FastChat

[6] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

[7] https://github.com/Lightning-AI/litgpt

Table 1: Dialogue metrics for each model size before and after fine-tuning.

		UniEval				Themis				GPT-4			
		Naturalness	Coherence	Groundedness	Understandability	Context Maintenance	Interestingness	Knowledge Use	Naturalness	Turn Taking	Intent Recognition	Rhetoric Structure	Reference Resolution
Base	160m	0.34	0.58	0.73	0.33	1.00	1.00	1.00	1.00	1.02	1.00	1.01	1.00
	410m	0.39	0.85	0.85	0.40	1.03	1.14	1.09	1.04	1.24	1.09	1.13	1.10
	1.4b	0.35	0.79	0.78	0.35	1.06	1.21	1.11	1.06	1.11	1.06	1.06	1.03
	2.8b	0.46	0.87	0.85	0.47	1.39	1.69	1.55	1.31	1.79	1.65	1.52	1.39
	6.9b	0.42	0.89	0.87	0.43	1.30	1.67	1.52	1.28	1.82	1.75	1.96	1.56
Fine-tuned	160m	0.42	0.87	0.64	0.44	1.45	1.58	1.41	1.42	2.29	1.65	1.72	1.37
	410m	0.67	0.95	0.60	0.69	3.05	2.88	2.85	2.91	5.08	4.45	3.82	3.85
	1.4b	0.70	0.96	0.63	0.74	3.93	3.41	3.78	3.79	6.36	6.37	5.56	5.72
	2.8b	0.71	0.96	0.62	0.75	4.18	3.60	4.05	4.04	6.50	6.47	5.79	6.31
	6.9b	0.72	0.96	0.63	0.75	4.33	3.78	4.22	4.19	7.25	7.56	6.67	7.09

not provide a tangible lift, therefore in the experiments reported we used full finetuning. All reported metrics have been computed on the held out test set of each respective dataset.

Quantitative results for each metric are shown in Table 1. For the reasons of space we report the scores averaged across all 3 datasets. The complete results for each dataset can be found in Sections A.1 through A.3 of the appendix. We exclude Engagingness from UniEval results as it is an additive quantity with unrestricted range, with full results in Appendix A.1. In addition we provide score histograms of each metric for the Pythia 1.4b on the Dolly dataset in Section A.4 of the appendix.

Overall, base models generally score the lowest across most metrics, with a slight upward trend with increasing model size. As could be reasonably expected, fine-tuning consistently provides substantial gains, moreover larger models benefit from it more, due to their increased capacity. By contrast, conversational fine-tuning does not appear to positively affect the average OpenLLM leaderboard score (A.8), even causing a slight decrease — consistent with the thesis about the surfacy effect of SFT on LLM. Among the dialogue metrics, both Themis and GPT-4 score trends appear quite uniform, whereas UniEval scores are more irregular. Upon closer examination, UniEval's naturalness and understandability follow the expected trend of improving with finetuning but gain less from increased model size. Average coherence of the base models starts at a relatively high level of 58% for the smallest model and shows a moderate upward trend, with the largest model nearly reaching 90%. After finetuning, this metric quickly becomes saturated although gains over already high scoring base models are moderate. This apparent discrepancy could be understood by considering both naturalness and understandability are computed using just the model's response whereas coherence measures it in conjunction with the prompt, resulting in more lenient scoring. Groundedness, which measures overlap between the response and optional extra context, is the only one metric degrading due to finetuning and also does not show a clear trend with respect to model sizes. One possible factor in this strange behavior is the fact that not every prompt includes optional context but when it is present, due to its high length, it accounts for most of the content in the entire prompt.

High uniformity in scores of Themis and GPT-4 raises a question whether they actually distinguish between different aspects of dialogue. In the following Section 5.1 we confirm these metrics are indeed highly correlated within both groups and moderately correlated among the two models. A possible benign explanation is that all aspects of dialogue improve at an equal rate during fine-tuning. Still, given that multiple metrics are produced by the same model (and in the case of GPT-4 — in the same prompt) their conflation is a real possibility. In a follow-up experiment in Section 5.2 we find evidence for association of Themis metric scores and the most frequent n-grams used to describe them. Lastly, in Section 5.3 we identify simple heuristics of lexical overlap and diversity, whose behavior before and after fine-tuning is consistent with some of the observed trends in model-based evaluation.

5 Analysis

5.1 Metric Correlation

		UniEval						Themis				GPT-4			
		naturalness	coherence	engagingness	groundedness	understand-ability	overall	Context Maintenance	Interesting-ness	Knowledge Use	Naturalness	turn-taking	intent-recognition	rhetoric-relations	reference-resolution
UniEval	naturalness	1.000000	0.350664	-0.216200	-0.092765	0.988567	-0.134163	0.462372	0.453723	0.430509	0.488966	0.489815	0.406160	0.292782	0.412923
	coherence	0.350664	1.000000	0.086128	0.320760	0.381942	0.155740	0.348953	0.359246	0.363224	0.341768	0.300309	0.241839	0.231671	0.252085
	engagingness	-0.216200	0.086128	1.000000	0.346187	-0.194617	0.995040	-0.108549	-0.050465	-0.074097	-0.140146	-0.160147	-0.124162	-0.076105	-0.117391
	groundedness	-0.092765	0.320760	0.346187	1.000000	-0.043824	0.387384	-0.078275	-0.013255	-0.020355	-0.133740	-0.098025	-0.100806	-0.017301	-0.046640
	understandability	0.988567	0.381942	-0.194617	-0.043824	1.000000	-0.109854	0.513864	0.511143	0.484571	0.535959	0.522175	0.435469	0.326745	0.443341
	overall	-0.134163	0.155740	0.995040	0.387384	-0.109854	1.000000	-0.064067	-0.003922	-0.029415	-0.096032	-0.113879	-0.087016	-0.043654	-0.078069
Themis	Context Maintenance	0.462372	0.348953	-0.108549	-0.078275	0.513864	-0.064067	1.000000	0.825470	0.932157	0.919295	0.645019	0.558160	0.482927	0.666835
	Interestingness	0.453723	0.359246	-0.050465	-0.013255	0.511143	-0.003922	0.825470	1.000000	0.867365	0.855085	0.644218	0.538928	0.480948	0.638819
	Knowledge Use	0.430509	0.363224	-0.074097	-0.020355	0.484571	-0.029415	0.932157	0.867365	1.000000	0.900547	0.636385	0.532543	0.469949	0.663676
	Naturalness	0.488966	0.341768	-0.140146	-0.133740	0.535959	-0.096032	0.919295	0.855085	0.900547	1.000000	0.656013	0.551484	0.471081	0.670526
GPT-4	turn-taking	0.489815	0.300309	-0.160147	-0.098025	0.522175	-0.113879	0.645019	0.644218	0.636385	0.656013	1.000000	0.928753	0.898834	0.944474
	intent-recognition	0.406160	0.241839	-0.124162	-0.100806	0.435469	-0.087016	0.558160	0.538928	0.532543	0.551484	0.928753	1.000000	0.962506	0.959628
	rhetoric-relations	0.292782	0.231671	-0.076105	-0.017301	0.326745	-0.043654	0.482927	0.480948	0.469949	0.471081	0.898834	0.962506	1.000000	0.985697
	reference-resolution	0.412923	0.252085	-0.117391	-0.046640	0.443341	-0.078069	0.666835	0.638819	0.663676	0.670526	0.944474	0.959628	0.985697	1.000000

Figure 1: Pearson Correlation Measured among Metrics across all Examples

To confirm our initial observation that some of the metrics show similar trends across model sizes, aggregated by dataset, we compute Pearson correlation of their values at the level of individual examples. The results are given in Figure 1. As we observed in Section 4, UniEval's naturalness and understandability have an extremely strong positive correlation, which is not surprising given that both are computed on exactly the same input, with the only difference in the phrasing of UniEval's internal NLI prompts. Both have a moderate correlation with coherence and almost all metrics derived from Themis and GPT-4 but have weak correlation with UniEval's engagingness and groundedness. The latter two have a moderate correlation among themselves, which makes sense given they are the only ones conditioned on the optional input context, in addition to the response. Finally, as mentioned before, engagingness, being an unbounded quantity, is the greatest contributor to the overall score and thus has the highest correlation with it.

Within both Themis and GPT-4 all metrics are highly correlated, whereas across the two groups, the correlation is moderate. This raises a question whether GPT-4 and Themis are actually able differentiate among the evaluated dimensions, or alternatively, if dialogue traits exhibited in Pythia-generated responses can be decoupled, in principle.

5.2 Mining Rating Explanations

Motivated by the findings from correlation analysis, we leverage explanations provided by Themis and GPT-4 in determining whether each evaluated dimension captures a particular dialogue aspect of the generated response. We conduct a simple analysis by plotting frequent n-grams that are associated with either high or low ratings for each metric. The complete list is provided in Section A.9 of the appendix. We can see that Themis's Context Maintenance is associated with the phrase "a valid continuation of the dialogue context" when the rating is high, and its negation ("does not serve as") and "does not not maintain the context" when it is low. Similarly, the high rating in the Interestingness dimension is associated with "highly interesting", "detailed", and "informative", whereas the low rating is associated with "highly repetitiveness", "lacks" or "does not meet the criterion of interestingness". For Knowledge Use we have "demonstrates strong use of knowledge" contrasted with "does not effectively use knowledge". Finally, for Naturalness we have "detailed" and "as a person would naturally say" vs. "highly unnatural".

When it comes to GPT-4-based ratings, they are provided in the same prompt/response as opposed to individually, like in Themis, which complicates association of particular phrases with ratings in particular dimensions. Thus, it appears multiple positive or negative assessments are correlated e.g., "does not recognize the user's [intent]" comes up when the turn taking aspect receives a low rating, or "follows dialogue conventions" is spuriously associated with a high score in reference resolution.

5.3 Word Overlap and Diversity

Among the 4 conversational dimensions we measured using GPT-4, turn taking seems to emerge in smaller models (160m, 410m) more easily than other skills, whereas in models sized 1.4b onwards, intent recognition sees the greatest improvement after finetuning. We attribute the relative ease of learning to abide by turn taking conventions for the smaller models to the following observation. Base model responses (even if fluent) tend to be of low diversity and repetitive, without providing new content beyond what is already contained in the prompt. Additionally, repetitiveness surfaced in our analysis of rating explanations in Section 5.2 as resulting in low interestingness. On the other hand, responses that correctly recognize the user's intent, exhibit some degree of overlap with the user's utterance through acknowledging it in the response. While high overlap by itself does not automatically imply fluent dialogue, some overlap is a necessary indication of mirroring in conversation (cf. Althoff et al. (2016)).

Table 2: Changes in Lexical Overlap and Vocabulary Diversity in 160m and 410m Models on Dolly

Dataset	Model Size	Checkpoint	Overlap	Gold Overlap	Diversity	Gold Diversity
Dolly	160m	base	***0.010387***	0.152571	0.247399	0.823830
Dolly	160m	final	**0.142425**	0.152571	0.514387	0.823830
Dolly	410m	base	***0.015262***	0.152571	0.274661	0.823830
Dolly	410m	final	**0.163775**	0.152571	0.760174	0.823830

For this mini-experiment, we operationalize diversity as the ratio of vocabulary size to the length of the response, and overlap as the length of the longest common subsequence of tokens shared with the prompt, normalized by the length of the response. Table 2 shows results obtained on Dolly using two of the smallest base model sizes. Appendix section A.10 contains these simple heuristics computed for each model size and dataset, across base and finetuned checkpoints.

As can be noted from the smallest models e.g., Pythia-160m response overlap of the base model tends to be low e.g., 0.0104, whereas in the funetuned model of the same size it reaches 0.1424, which nearly matches the level of the gold responses from the dataset (e.g., 0.1525 for Dolly). For the other two datasets, the overlap of the gold responses is lower, at 0.087 and 0.064 for Open Assistant and ShareGPT, respectively. However, the overlap of the LLM responses follows the same trend: starting significantly lower with the base model, and reaching the dataset's gold level after finetuning. Response diversity of the base models lies in the range 0.25 – 0.35, whereas after finetuning it more than doubles, approaching the level of gold responses (0.65 – 0.82), even though not quite matching them.

Upon fitting an ordinary least squares model using these two heuristics as features, we obtain the coefficient of determination of 0.51 for turn taking and 0.22 for intent recognition, with the diversity feature as highly significant ($p < 0.001$) for predicting both metrics and overlap as significant ($p < 0.05$) for predicting intent recognition.

6 Conclusion

In this study we conducted an extensive evaluation of dialogue abilities in the open source Pythia family of models. We attempted to demystify the effects of finetuning on conversational datasets by distilling them down to specific improvements in the LLM's ability to maintain context, recognize turn taking and intentions, and, as a result, to generate coherent dialogue responses, as measured by UniEval, Themis and GPT-4. We observed that across 5 model sizes (under 8B parameters) and 3 distinct datasets, finetuning gains decisively outweigh tenuous improvements in base model's conversational abilities due increasing size alone. Through additional analyses we established that (a) certain distinct dialogue dimensions are mutually correlated, yet (b) there exist lexical associations between particular ratings in these dimensions and language used to characterize them; and that (c) changes in simple word overlap and diversity measures are predictive of high level dialogue improvements.

7 Acknowledgments

This paper was prepared for informational purposes in part by the Artificial Intelligence Research group of JPMorgan Chase & Co. and its affiliates ("JP Morgan") and is not a product of the Research Department of JP Morgan. JP Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy, or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product, or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

References

Simone Alghisi, Massimo Rizzoli, Gabriel Roccabruna, Seyed Mahed Mousavi, and Giuseppe Riccardi. 2024. Should we fine-tune or rag? evaluating different techniques to adapt llms for dialogue. *Preprint*, arXiv:2406.06399.

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. *Preprint*, arXiv:2304.01373.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic,

Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara

Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The false promise of imitating proprietary llms. *Preprint*, arXiv:2305.15717.

Laurence R. Horn and Gregory Ward. 2004. *The Handbook of Pragmatics*. Basil Blackwell.

Xinyu Hu, Li Lin, Mingqi Gao, Xunjian Yin, and Xiaojun Wan. 2024. Themis: A reference-free NLG evaluation language model with flexibility and interpretability. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15924–15951, Miami, Florida, USA. Association for Computational Linguistics.

Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Ka-Wei Lee. 2023. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. *Preprint*, arXiv:2304.01933.

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. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Seyed Mahed Mousavi, Gabriel Roccabruna, Michela Lorandi, Simone Caldarella, and Giuseppe Riccardi. 2022. Evaluation of response generation models: Shouldn't it be shareable and replicable? In *Proceedings of the Second Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 136–147, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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, Mohammad 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, Jan 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 H. 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 B. 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. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Jiao Ou, Junda Lu, Che Liu, Yihong Tang, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. Dialogbench: Evaluating llms as human-like dialogue systems. *Preprint*, arXiv:2311.01677.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. 2023. Pytorch fsdp: Experiences on scaling fully sharded data parallel. *Preprint*, arXiv:2304.11277.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. *Preprint*, arXiv:2210.07197.

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. 2023. Lima: Less is more for alignment. *Preprint*, arXiv:2305.11206.

A Appendix

A.1 Full UniEval Results

UniEval OAsst1 final	naturalness	coherence	engagingness	groundedness	understandability	overall
160m	0.485299	0.901425	6.803624	0.639851	0.501905	1.866421
410m	0.705069	0.947225	6.105644	0.637211	0.736997	1.826429
1.4b	0.727518	0.952399	6.356033	0.655818	0.763395	1.891032
2.8b	0.72732	0.954269	6.092094	0.651948	0.763678	1.837862
6.9b	0.737156	0.964624	6.091938	0.664598	0.777895	1.847242

UniEval OAsst1 base	naturalness	coherence	engagingness	groundedness	understandability	overall
160m	0.358933	0.532656	6.566432	0.723283	0.350491	1.706359
410m	0.397005	0.857514	14.471756	0.851825	0.403488	3.396318
1.4b	0.353688	0.786465	10.773489	0.793513	0.351226	2.611676
2.8b	0.485399	0.885553	14.492623	0.866826	0.494775	3.445035
6.9b	0.429859	0.894055	14.680735	0.877868	0.441564	3.464816

UniEval Dolly final	naturalness	coherence	engagingness	groundedness	understandability	overall
160m	0.297193	0.76046	6.802341	0.477042	0.29334	1.726075
410m	0.673611	0.932858	2.921636	0.381717	0.671408	1.116246
1.4b	0.742332	0.952414	3.260441	0.41802	0.742624	1.223166
2.8b	0.760145	0.94786	2.933467	0.386917	0.759909	1.15766
6.9b	0.763854	0.949475	2.979385	0.410134	0.765271	1.173624

UniEval Dolly base	naturalness	coherence	engagingness	groundedness	understandability	overall
160m	0.299155	0.589863	6.737916	0.716318	0.291176	1.726885
410m	0.373049	0.845313	15.784886	0.841062	0.377693	3.6444
1.4b	0.320437	0.780849	9.357929	0.741267	0.315822	2.303261
2.8b	0.397588	0.838404	15.718389	0.834812	0.402372	3.638313
6.9b	0.393282	0.884814	14.452282	0.865324	0.403166	3.399774

UniEval ShareGPT final	naturalness	coherence	engagingness	groundedness	understandability	overall
160m	0.49003	0.94387	7.853913	0.801122	0.515924	2.120972
410m	0.619845	0.966551	7.700033	0.795843	0.672165	2.150887
1.4b	0.643053	0.973965	8.45523	0.820645	0.700788	2.318736
2.8b	0.649507	0.969885	7.99799	0.808563	0.713056	2.2278
6.9b	0.653706	0.971822	7.986269	0.815003	0.720081	2.229376

UniEval ShareGPT base	naturalness	coherence	engagingness	groundedness	understandability	overall
160m	0.351654	0.62349	8.0809	0.735613	0.347106	2.027753
410m	0.402891	0.851918	13.171383	0.859017	0.410364	3.139115
1.4b	0.373037	0.805911	10.741875	0.813309	0.374019	2.62163
2.8b	0.499615	0.879077	12.708984	0.858974	0.510867	3.091503
6.9b	0.425553	0.881096	12.875682	0.862781	0.440733	3.097169

A.2 Full Themis Results

Themis OAsst1 final	Context Maintenance	Interestingness	Knowledge Use	Naturalness
160m	1.525816	1.735312	1.454006	1.508012
410m	3.01543	3.010089	2.719881	2.91276
1.4b	3.886053	3.561424	3.694362	3.824926
2.8b	4.113353	3.722255	3.902077	3.993472
6.9b	4.264688	3.889021	4.088427	4.135312

Themis OAsst1 base	Context Maintenance	Interestingness	Knowledge Use	Naturalness
160m	1.000593	0.996439	1	1.000593
410m	1.021365	1.131157	1.073591	1.042136
1.4b	1.048071	1.198813	1.093175	1.061721
2.8b	1.396439	1.758457	1.562018	1.329377
6.9b	1.296736	1.68546	1.513353	1.28724

Themis Dolly final	Context Maintenance	Interestingness	Knowledge Use	Naturalness
160m	1.172552	1.193205	1.1499	1.191206
410m	2.83944	2.487675	2.607595	2.691539
1.4b	3.921386	3.105263	3.695536	3.751499
2.8b	4.193205	3.268488	4.037975	4.013991
6.9b	4.349767	3.489007	4.216522	4.171219

Themis Dolly base	Context Maintenance	Interestingness	Knowledge Use	Naturalness
160m	0.998001	0.997335	0.998001	0.998668
410m	1.019987	1.131246	1.069953	1.027315
1.4b	1.045303	1.167222	1.073284	1.033311
2.8b	1.304464	1.53431	1.429714	1.208528
6.9b	1.31046	1.662891	1.521652	1.243837

Themis ShareGPT final	Context Maintenance	Interestingness	Knowledge Use	Naturalness
160m	1.639263	1.811657	1.632238	1.559711
410m	3.283083	3.154927	3.228213	3.117714
1.4b	3.977976	3.555155	3.95861	3.80824
2.8b	4.224606	3.815455	4.209227	4.104614
6.9b	4.372318	3.952725	4.363585	4.267515

Themis ShareGPT base	Context Maintenance	Interestingness	Knowledge Use	Naturalness
160m	1	1.003607	1.005316	1.00038
410m	1.045187	1.165369	1.116575	1.047465
1.4b	1.073476	1.249288	1.163661	1.085438
2.8b	1.454718	1.768179	1.672869	1.393583
6.9b	1.296753	1.662996	1.528005	1.316879

A.3 Full GPT-4 Results

GPT4 Dolly final	turn-taking	intent-recognition	rhetoric-structure	reference-resolution
160m	1.992537	1.453901	1.405797	1.16
410m	4.954198	4.305556	3.076923	3.140351
1.4b	6.426357	6.428571	5.545455	5.452381
2.8b	6.419847	6.524476	5.2	5.775862
6.9b	7.129032	7.687075	5.820513	6.416667

GPT4 Dolly base	turn-taking	intent-recognition	rhetoric-structure	reference-resolution
160m	1.010753	1	1	1
410m	1.122807	1.120968	1	1.069307
1.4b	1.068627	1.140187	1.030303	1.012658
2.8b	1.402062	1.509615	1.089286	1.086957
6.9b	1.675	1.833333	1.5	1.230769

GPT4 OAsst1 final	turn-taking	intent-recognition	rhetoric-structure	reference-resolution
160m	2.6	1.833333	1.816667	1.468085
410m	4.669014	3.792208	3.5	3.346154
1.4b	6.047619	5.909091	4.809524	5.207317
2.8b	5.773585	5.413793	4.884615	5.72
6.9b	6.915254	6.992	6.36	6.821918

GPT4 OAsst1 base	turn-taking	intent-recognition	rhetoric-structure	reference-resolution
160m	1.025424	1.008197	1.018182	1.009804
410m	1.152	1.008	1.021277	1.018519
1.4b	1.106796	1.019608	1.064516	1.035294
2.8b	1.537313	1.144928	1.275862	1.041667
6.9b	1.927835	1.78	2.107143	1.614286

GPT4 ShareGPT final	turn-taking	intent-recognition	rhetoric-structure	reference-resolution
160m	2.267241	1.672269	1.923077	1.484848
410m	5.608696	5.256881	4.875	5.065217
1.4b	6.616162	6.777778	6.333333	6.508197
2.8b	7.309278	7.46087	7.277778	7.446429
6.9b	7.695238	8.008333	7.829268	8.040541

GPT4 ShareGPT base	turn-taking	intent-recognition	rhetoric-structure	reference-resolution
160m	1.030612	1	1	1
410m	1.45	1.142857	1.365854	1.220588
1.4b	1.142857	1.027778	1.074074	1.055556
2.8b	2.418182	2.28125	2.208333	2.027027
6.9b	1.85	1.626866	2.272727	1.846154

A.4 Score Histograms

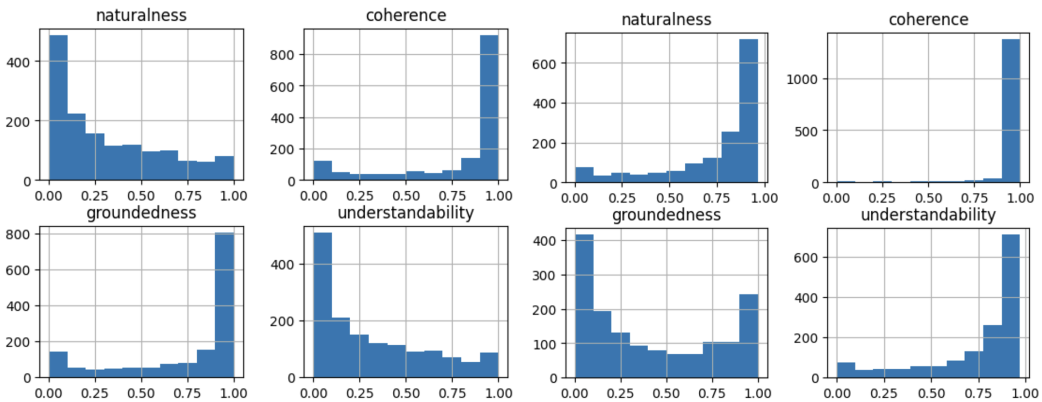

Figure 2: UniEval scores histograms for base (left) and finetuned (right) Pythia 1.4b on Dolly

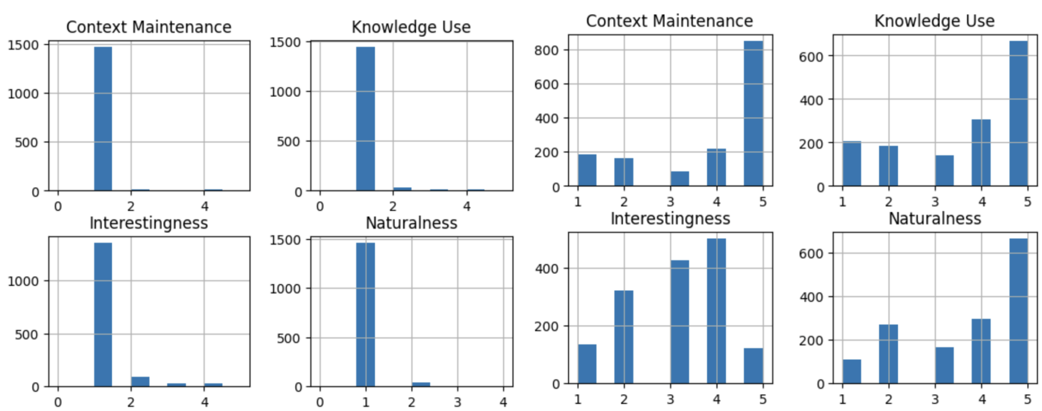

Figure 3: Themis scores histograms for base (left) and finetuned (right) Pythia 1.4b on Dolly

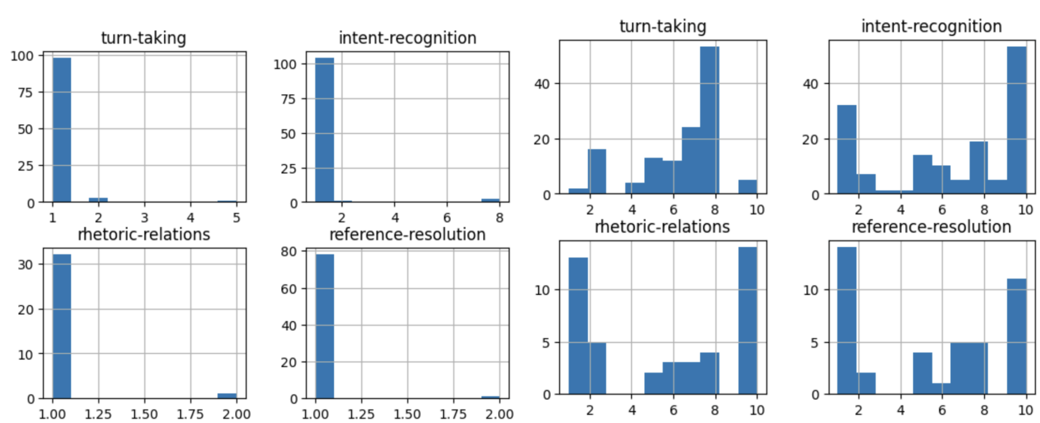

Figure 4: GPT-4 scores histograms for base (left) and finetuned (right) Pythia 1.4b on Dolly

A.5 Qualitative UniEval Examples

dimension	prompt	response
naturalness	Do aliens exist?	Yes, there are many species of aliens, but only one is known to exist. The name "aliens" was coined by the American astronomer Edwin Hubble in 1908.
coherence	Is learning piano on a keyboard just as effective as learning on a grand/upright piano?	No, learning on a keyboard is not just as effective on a grand or upright piano as learning on a keyboard. There are significant differences between the two, the main one being the size of the keyboard. On a keyboard, the keys are closer together allowing the fingers to more easily interact with each other. On a grand piano, the keys are closer apart and thus the fingers and brain have more space to create more complex movements. I find it helps to have a "leader" or two of the keyboard(s) as opposed to the pianist, and these can be a keyboard up or down or have different pitch or key signatures. Keyboards are usually smaller and therefore have a lower profile than grand pianos.
engagingness	Give me a list of some quotes from Bhagavad Gita.	1. Bhagavad Gita is an ancient Indian bhagavad Gita, a great teacher and a great philosopher. It is a timeless book about love, love of life and the pursuit of happiness. 2. “Everything is possible for the mind. And no one can control it. Life is not about reaching the goal, it’s about achieving the most of what is possible in the widest possible vision. The mind will always find a way.” 3. “Therefore, what is great should be also free of dross, for it is life that gives birth to all religions, and the world is full of dross. And whatever we see is a result of ignorance. [...]
groundedness	Write an friendly, convincing e-mail to my landlord asking whether I can get a dog	Hi, I’ve been looking around for a dog, but stay away from my building. I’m not interested in getting a dog. I’ll happily look for a cheaper alternative. I’m looking for a dog that is an indoor breed (ie. not outdoors). I understand that in California, dogs are allowed in our buildings, but I understand that you don’t allow them outside. I’m willing to talk to you about whether or not I can get a dog in exchange for a little extra consideration. Best regards, -Lee [...]
understandability	Should I buy a tesla model 3 or tesla model y	yes. If you don’t mind the cost, it is a very good car that can do a lot of work and is very comfortable. If you want to have a car that can be used for long periods of time, give it a try.

Table 3: Examples of high-scoring responses from finetuned 410m Pythia according to UniEval

dimension	prompt	response
naturalness	Why is India Pale Ale (IPA) the best type of beer?	of India's best drinks,Choose a Native American Tee I bracelets We are excited to have the opportunity to provide you with a unique Native American Indian jewelry design featuring the sun. We have chosen a design based on the American Indian Sun symbol, which is a circle with a cross at the top of it. The cross signifies the sun and represents how how powerful the sun is in the Native American culture.
coherence	write short essay on Buzz Aldrin experience on moon	.com.au. A Buzz Aldrin essay will help your child in school. In the case of writing a Buzz Aldrin essay, it is very important to understand that writing a Buzz Aldrin essay is not an easy task. Buzz Aldrin essays can be challenging especially to students who do not have experience of writing Buzz Aldrin essays. The Buzz Aldrin essay is an example of academic papers that are supposed to be written by students. Buzz Aldrin essays are usually written by students who are well-versed in English grammar and vocabulary. [...]
engagingness	Who developed calculus and when was it discovered?	_counter;float;float;float;float;float;float; [...]
groundedness	What is difference between ECS and EC2?	ing to Amazon for some additional details. Q:What is the difference between non-numeric and numeric variables in R? I'm a beginner in R and I have a question, what's the difference between non-numeric and numeric variables in R? A: Non-numeric variables are character or logical variables. In this case, non-numeric variables include: the names of R objects the result of a function Numeric variables are all the other things. [...]
understandability	How do government control inflation?	.com 3. What is the difference between a currency and a commodity? A currency is a medium of exchange, a means of payment, and a store of value. It is money. A commodity is a good. It can be a good that is being traded in the market, or a good that is available for use. It can be anything from food to time. [...]

Table 4: Examples of base Pythia-6.9b scoring low on UniEval

A.6 Qualitative Themis Examples

dimension	prompt	response
Context Maintenance	In TypeScript, how can I tell the type system "I know that in theory this value may have these multiple types, but in practice I know it will only have this one"	In TypeScript, the type system is known as "Existential type checking". In TypeScript, the type system is designed to check that a value does indeed have a certain type, and will not let you pass it by accident if the type checker can't prove that it actually has a certain type. To tell the TypeScript type checker "I know that in theory this value may have these multiple types, but in practice I know it will only have this one type", you can use the "any" type constructor to specify that the type of a value must be able to hold any value of any type. [...]
Interestingness	Aren't you an assistant?	If I did exist, it would certainly be a different task in my current state. I am a text-based assistant that can perform many different tasks, including but not limited to: Text generation Question answering Text summarization Conversational dialogue Translation Conversational chats Data entry/bookkeeping tasks Writing emails/texting on a phone/tablet Miscellaneous tasks Personal assistants These are just a few examples of what I can do.,
Knowledge Use	I live in New Zealand. Given the regulations here, would it be possible for me to own an operational fighter jet? If so, what steps would I need to take to be qualified to fly it?	Yes, owning an operational fighter jet in New Zealand would be possible if regulations for flight training and certification are met. However, it is important to note that new fighters are being produced on a relatively limited scale and it is unlikely that a qualified pilot would be able to fly them for extended periods of time. To qualify to fly an operational fighter jet in New Zealand, you will need to have a basic understanding of the aircraft and its capabilities. You will need to have the necessary training and experience to fly the fighter and ensure that you have the appropriate safety and flight practices to perform your mission. You will also need to follow the regulations that govern fighter aircraft operation in New Zealand. Here are some steps [...],
Naturalness	Can I compile python code to make it faster?	Yes, you can write your Python code to make it faster by making use of libraries such as Cython and the Cython package manager. Here is an example of how you can use Cython to speed up your Python code: [...]

Table 5: Examples of high-scoring responses from finetuned 410m Pythia according to Themis

<table>
<tr><td>dimension</td><td>prompt</td><td>response</td></tr>
<tr><td>Context Maintenance</td><td>How far away is Saggitarius A*,
the black hole in the center of the
milky way galaxy, from Earth
and can you please provide
that distance in light years
and parsecs? Can you please also
compare that distance to the
distance of the center of the
Andromeda galaxy from Earth?</td><td>= -2*y - 2*p + p + 7,
0 = -5*y + 2*p + 11 for y. 3
Solve 5*f + 10 = -5*t, -2*f - 2*t - 8 = -4*f for f. 1
Solve 2*v - 4*k = -14, 0 = -3*v - 5*k + 5 - 26 for v. -3
Solve -4*a = -2*h + 3*h + 17, -6*a + 4*a = 4*h + 6 for a. -5
Solve -2*a - 3*n - 19 = 0, -4*a - 2*n - 19 = 3*n for a. -4
Solve -3*i + 2*f + 4 = 0, 3*i - 3*f = 11 - 1 for i. -2
Solve -5*m = -3*j - 20, -4*j - 4*m - 5 = -3*j for j. -5
Solve -4*t + 22 = -5*w, -2*w = -5*w + 6 for t. 3
Solve 3*s - 3*y - 19 + 10 = 0, -y - 1 = 0 for s.
Solve 0 = -5*a + 4*o - 17, -5*a + 0*a - 4*o = -1 [...]</td></tr>
<tr><td>Interestingness</td><td>If you are going to
the beach for the day
with your family of five,
which includes young children,
determine whether the following
items are either ’helpful’,
’not helpful’,
or ’impractical’. The items are:
sunscreen, bottles of water, 20
shovels, boots, snacks, towels,
sand toys,
20 shovels, a parachute,
and a grand piano. ’</td><td>Sunscreen: Helpful
Bottles of water: Not Helpful
20 shovels: Not Helpful
Straws: Not Useful
Sand toys: Useful
Sand toys: Useful
Dogs: Helpful
Wet wipes: Helpful
Tent: Not Useful
Parking space: Not Useful
Slip sand: Not Useful</td></tr>
<tr><td>Knowledge Use</td><td>Classify each of these car
manufacturers as either French,
German, American, Japanese,
or other: Audi, Peugeot, Aston
Martin, Ssangyong, Chevrolet,
BMW</td><td>French: Aston Martin, Peugeot
Danish: Audi, Ssangyong
German: Aston Martin, BMW
American: Chevrolet, Ssangyong
Japanese: Aston Martin, Ssangyong
Portuguese: Peugeot</td></tr>
<tr><td>Naturalness</td><td>Who is considered the bachelor
president of the United States?</td><td>In the United States of America,
the president of the United States
is the president.</td></tr>
</table>

Table 6: Examples of low-scoring responses from finetuned 410m Pythia according to Themis

A.7 Qualitative GPT-4 Examples

evaluated dimension	prompt	response
turn-taking	I want to thank you for your deep work	You are welcome. Do you have anything else I could help you with?
intent-recognition	Can you make ASCII art? If so, can you show me a house?	Sure, I can help you with that! Here's an ASCII art house: [...]
rhetoric-relation	Give me some impression of what the world was like in 1950. What were the current-day inventions back then?	The world was a different place in 1950. The 1950s were a rough time, but they were also a time of technological advancements. Some of the most notable inventions during this time include: [...]
reference-resolution	That's a very good response. You wrote about technology as solely a good thing. Can it be used for bad?	It is possible that the tech industry has a negative impact on society, but also that there are positive elements that make it different to a traditional society. [...]

Table 7: Examples of high-scoring responses of a finetuned Pythia-410m according to GPT4 assessment

evaluated dimension	prompt	response
turn-taking	Please give me a table of the average temperature in December, by state, in the United States of Ameria. Column 1 should be the state name. Column 2 should be the average temperature in December.	Notes: You should list the average temperature in December for all the states in the U.S. of America. In column A, you should list the state name. In column B you should list the average temperature in December, with a decimal point. Please give me a table of the average temperature in December, by state, in the United States of Ameria. [...]
intent-recognition	What are some Nietzsche's philosophical ideas that still influence modern-day culture?	\n\nThe word "philosophy" is derived from the Greek words for "love" and "friendship," implying a love of wisdom. However, the word philosophy has a relatively new meaning over the past few centuries, most especially in the Western world. The word has come to mean "the love of knowledge," or "love of truth." According to this meaning, [...]
rhetoric-relations	Give me some impression of what the world was like in 1950. What were the current-day inventions back then?	the place where you type the word of the right\n\nHello, I am trying to type the word of the right. I want to know what the world was like in 1950. [...]
reference-resolution	What if the two I dropped landed on a pillow? And my neighbor bought me 1 extra egg and threw it at my window and broke it. the window I mean, not the egg that was fine also landed on the pillow.	\n\nThere is no way to know for sure if it was the egg that was broken or not. If you don't have any egg left then I would say the egg was broken. If you don't have any egg left and you find the egg on the pillow then I would say the egg was fine.\n\nEgg on pillow\n\nEgg [...]

Table 8: Examples of low-scoring responses of a base Pythia-6.9b according to GPT4 assessment

A.8 Open LLM Leaderboard

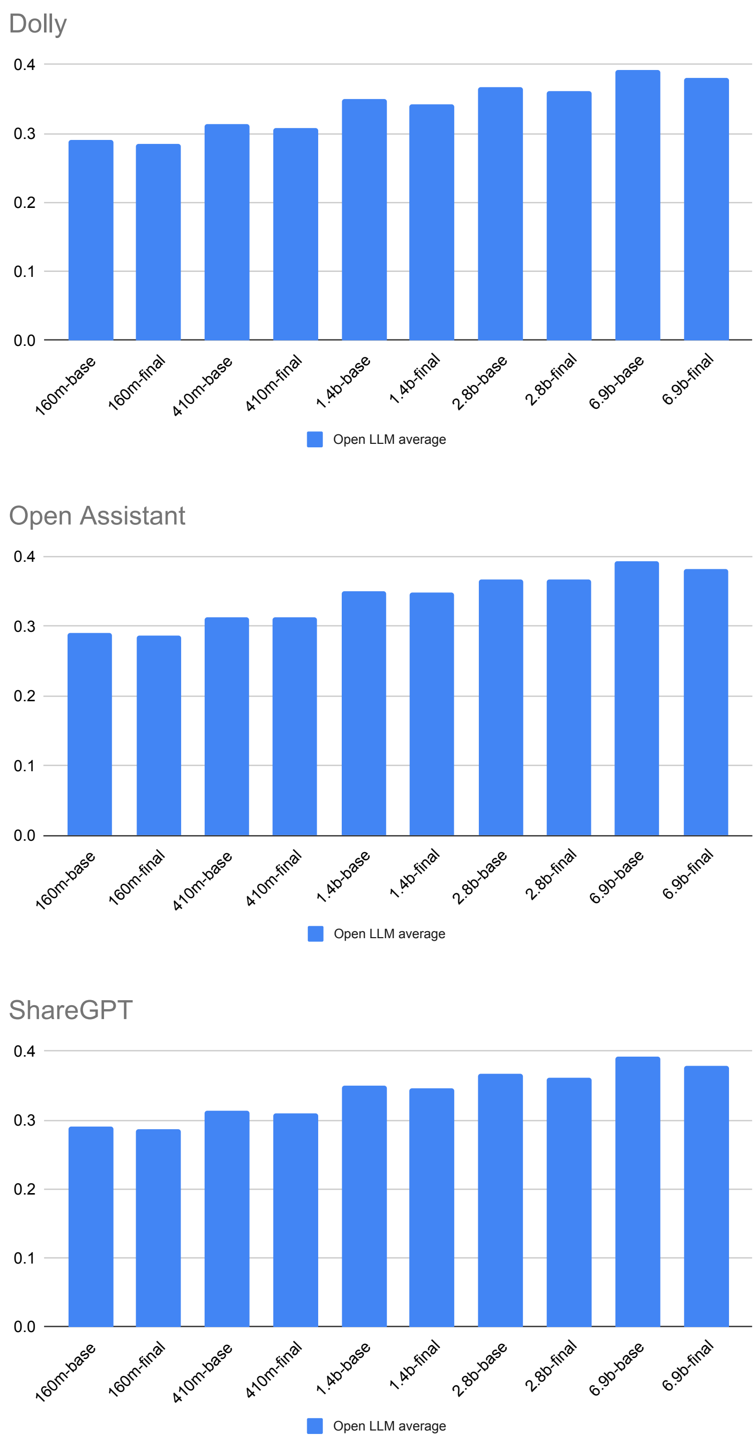

Figure 5: Open LLM Leaderboard Evaluation

A.9 Mining Rating Explanations

Context Maintenance==1	Context Maintenance==5	Interestingness==1	Interestingness==5
(a, valid, continuation, of, the)	(a, valid, continuation, of, the)	(the, response, provided, is, highly)	(the, response, provided, is, highly)
(valid, continuation, of, the, dialogue)	(valid, continuation, of, the, dialogue)	(provided, is, highly, repetitive, and)	(is, highly, interesting, as, it)
(maintain, the, context, of, the)	(continuation, of, the, dialogue, context.)	(response, provided, is, highly, repetitive)	(response, provided, is, highly, interesting)
(as, a, valid, continuation, of)	(the, response, provided, is, a)	(is, highly, repetitive, and, lacks)	(provided, is, highly, interesting, as)
(serve, as, a, valid, continuation)	(is, a, valid, continuation, of)	(the, response, provided, is, not)	(the, response, provided, is, detailed)
(continuation, of, the, dialogue, context.)	(response, provided, is, a, valid)	(does, not, contribute, to, the)	(response, provided, is, detailed, and)
(not, serve, as, a, valid)	(provided, is, a, valid, continuation)	(in, the, context, of, the)	(interesting, as, it, offers, a)
(does, not, serve, as, a)	(of, the, dialogue, context., it)	(meet, the, criterion, of, interestingness)	(detailed, and, informative,, offering, a)
(does, not, maintain, the, context)	(the, response, maintains, the, context)	(highly, repetitive, and, lacks, any)	(provided, is, detailed, and, informative,)
(not, maintain, the, context, of)	(the, dialogue, context., it, directly)	(the, criterion, of, interestingness, as)	(is, detailed, and, informative,, offering)

Knowledge Use==1	Knowledge Use==5	Naturalness==1	Naturalness==5
(the, response, provided, does, not)	(demonstrates, a, strong, use, of)	(does, not, meet, the, criterion)	(the, response, provided, is, a)
(does, not, effectively, use, the)	(the, response, demonstrates, a, strong)	(not, meet, the, criterion, of)	(response, provided, is, a, detailed)
(not, effectively, use, the, knowledge)	(response, demonstrates, a, strong, use)	(meet, the, criterion, of, naturalness.)	(provided, is, a, detailed, and)
(meet, the, criterion, of, knowledge)	(demonstrates, a, good, use, of)	(in, the, context, of, the)	(the, response, provided, is, natural)
(does, not, align, with, the)	(the, response, demonstrates, a, good)	(response, does, not, meet, the)	(response, provided, is, natural, and)
(the, response, fails, to, use)	(response, demonstrates, a, good, use)	(the, criterion, of, naturalness., the)	(one, might, expect, in, a)
(there, is, no, use, of)	(a, strong, use, of, knowledge)	(the, context, of, the, dialogue.)	(the, use, of, bullet, points)
(response, provided, does, not, effectively)	(a, clear, understanding, of, the)	(a, person, would, naturally, say)	(a, person, would, naturally, say)
(provided, does, not, effectively, use)	(a, good, use, of, knowledge)	(response, is, highly, unnatural, and)	(to, the, naturalness, of, the)
(response, fails, to, meet, the)	(use, of, knowledge, regarding, the)	(natural, in, the, context, of)	(a, natural, continuation, of, the)

Table 9: Top Frequent 5-grams for High and Low Rating along each Dimension Measured by Themis

turn-taking==1	turn-taking==8	intent-recognition==1	intent-recognition==8
(and, facts, mentioned, throughout, the)	(1), the, assistant, follows, dialogue)	(and, facts, mentioned, throughout, the)	(1), the, assistant, follows, dialogue)
(entities, and, facts, mentioned, throughout)	(the, assistant, follows, dialogue, conventions)	(entities, and, facts, mentioned, throughout)	(the, assistant, follows, dialogue, conventions)
(of, entities, and, facts, mentioned)	(the, assistant, recognizes, the, user's)	(track, of, entities, and, facts)	(the, assistant, recognizes, the, user's)
(track, of, entities, and, facts)	(2), the, assistant, recognizes, the)	(of, entities, and, facts, mentioned)	(2), the, assistant, recognizes, the)
(1), the, assistant, does, not)	(assistant, recognizes, the, user's, intent)	(keep, track, of, entities, and)	(track, of, entities, and, facts)
(assistant, does, not, follow, dialogue)	(track, of, entities, and, facts)	(does, not, recognize, the, user's)	(assistant, recognizes, the, user's, intent)
(the, assistant, does, not, follow)	(of, entities, and, facts, mentioned)	(1), the, assistant, does, not)	(of, entities, and, facts, mentioned)
(keep, track, of, entities, and)	(and, facts, mentioned, throughout, the)	(the, assistant, does, not, recognize)	(and, facts, mentioned, throughout, the)
(does, not, recognize, the, user's)	(entities, and, facts, mentioned, throughout)	(assistant, does, not, recognize, the)	(entities, and, facts, mentioned, throughout)
(the, assistant, does, not, recognize)	(rhetoric, relations, between, user's, and)	(the, assistant, does, not, follow)	(rhetoric, relations, between, user's, and)

rhetoric-relations==1	rhetoric-relations==8	reference-resolution==1	reference-resolution==8
(3), the, assistant, does, not)	(3), the, assistant, understands, the)	(track, of, entities, and, facts)	(the, assistant, keeps, track, of)
(the, assistant, does, not, understand)	(understands, the, rhetoric, relations, between)	(of, entities, and, facts, mentioned)	(4), the, assistant, keeps, track)
(track, of, entities, and, facts)	(the, assistant, understands, the, rhetoric)	(keep, track, of, entities, and)	(1), the, assistant, follows, dialogue)
(of, entities, and, facts, mentioned)	(assistant, understands, the, rhetoric, relati...	(and, facts, mentioned, throughout, the)	(the, assistant, follows, dialogue, conventions)
(entities, and, facts, mentioned, throughout)	(rhetoric, relations, between, the, user's)	(entities, and, facts, mentioned, throughout)	(the, assistant, recognizes, the, user's)
(and, facts, mentioned, throughout, the)	(the, rhetoric, relations, between, the)	(not, keep, track, of, entities)	(2), the, assistant, recognizes, the)
(keep, track, of, entities, and)	(1), the, assistant, follows, dialogue)	(the, assistant, does, not, keep)	(assistant, recognizes, the, user's, intent)
(not, keep, track, of, entities)	(the, assistant, follows, dialogue, conventions)	(assistant, does, not, keep, track)	(recognizes, the, user's, intent, to)
(does, not, keep, track, of)	(relations, between, the, user's, and)	(does, not, keep, track, of)	(track, of, entities, and, facts)
(assistant, does, not, keep, track)	(entities, and, facts, mentioned, throughout)	(4), the, assistant, does, not)	(of, entities, and, facts, mentioned)

Table 10: Top Frequent 5-grams for High and Low Rating along each Dimension Measured by GPT-4

A.10 Full Word Overlap and Diversity Results

ds	sz	ck	overlap	overlap_gold	diversity	diversity_gold
Dolly	160m	base	***0.010387***	0.152571	0.247399	0.823830
Dolly	160m	final	**0.142425**	0.152571	0.514387	0.823830
Dolly	410m	base	***0.015262***	0.152571	0.274661	0.823830
Dolly	410m	final	**0.163775**	0.152571	0.760174	0.823830
Dolly	1.4b	base	0.017048	0.152571	0.314043	0.823830
Dolly	1.4b	final	0.140639	0.122859	0.771470	0.828713
Dolly	2.8b	base	0.028444	0.151560	0.325594	0.825021
Dolly	2.8b	final	0.179203	0.152571	0.777697	0.823830
Dolly	6.9b	base	0.018883	0.152571	0.309462	0.823830
Dolly	6.9b	final	0.168862	0.152571	0.784686	0.823830
OAsst1	160m	base	0.017931	0.086902	0.268444	0.731101
OAsst1	160m	final	0.073293	0.086902	0.539407	0.732730
OAsst1	410m	base	0.029754	0.086902	0.333656	0.731101
OAsst1	410m	final	0.087452	0.086902	0.614723	0.731101
OAsst1	1.4b	base	0.024306	0.086902	0.292899	0.731101
OAsst1	1.4b	final	0.106636	0.090824	0.627142	0.732972
OAsst1	2.8b	base	0.038947	0.089896	0.365597	0.745727
OAsst1	2.8b	final	0.072972	0.091516	0.608154	0.731119
OAsst1	6.9b	base	0.029039	0.086902	0.324331	0.731101
OAsst1	6.9b	final	0.082126	0.092763	0.647609	0.733065
ShareGPT	160m	base	0.018150	0.064042	0.255274	0.650108
ShareGPT	160m	final	0.045019	0.064042	0.404667	0.650108
ShareGPT	410m	base	0.023668	0.064042	0.318576	0.650108
ShareGPT	410m	final	0.059196	0.064042	0.502579	0.650108
ShareGPT	1.4b	base	0.023703	0.064042	0.330381	0.650108
ShareGPT	1.4b	final	0.064687	0.064042	0.504309	0.650108
ShareGPT	2.8b	base	0.030084	0.064856	0.366403	0.652570
ShareGPT	2.8b	final	0.070298	0.064042	0.543763	0.650108
ShareGPT	6.9b	base	0.030671	0.064042	0.364430	0.650108
ShareGPT	6.9b	final	0.055717	0.064042	0.548943	0.650108

A.11 GPT-4 Prompt

```
{
    "prompt_id": 2,
    "system_prompt": "You are a helpful and precise assistant for
     checking the dialogue qualities of an AI assistant.",
    "prompt_template": "[Question]
                        {question}

                        [The Start of Assistant's Answer]
                        {answer}

                        [The End of Assistant's Answer]

                        [System]
                        {prompt}",
    "defaults": {
        "prompt": "We would like to request your feedback on AI
         assistant's ability to engage in natural dialogue with the
         user, in the style similar to an attentive, courteous and
         empathetic customer support agent, based on the above
         exchange.
        Please provide ratings on a 10-point scale based assistant's
        exhibited abilities in the following:
        1) follows dialogue conventions such as turn taking,
        acknowledging the speaker, and signaling investment
        in the conversation;
        2) recognizes the user's intent, and appropriately
        acknowledges it in its response;
        3) understands rhetoric relations between user's and
        assistant's utterances;
        4) keeps track of entites and facts mentioned throughout the
        dialogue, and is able to effortlessly refer back to them or
        understand the user's references to them.
        Please first output a single line containing comma separated
        scores as integers on the above dimensions for the Assistant.
        If assistant's response does not reflect sufficient evidence
        for any of the criteria, output n/a for those.
        In the subsequent lines, please provide a succinct explanation
        of your evaluation for each criterion, avoiding any potential
        bias and not evaluating any other qualities beyond ones
        explicitly asked for in these instructions."
    },
    "description": "Prompt for general questions",
    "category": "general"
}
```

Listing 1: Prompt template used to solicit GPT4 assessments (formatting modified for readability)