
A Multi-Aspect Evaluation of Dialogue in Pythia

Anonymous Author(s)

Affiliation

Address

email

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 [4], 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 would¹. 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 [15]. 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 [21]. Similarly, LLM performance on downstream tasks is largely determined by the amount of pre-training of the base model [6]. 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 [7].

¹<https://openai.com/index/instruction-following/>

37 In this work, we investigate the effects of conversational finetuning on an open source LLM Pythia
38 [3] of five different sizes through the lens of these fine-grained linguistic dimensions. We contribute
39 an empirical analysis of changes in model’s text generation behavior, as a result of finetuning on 3
40 chat datasets: Databricks Dolly², Open Assistant³, and ShareGPT⁴, with respect to dialogue-specific
41 metrics: UniEval [20], Themis [8], and a targeted GPT-4-based assessment.

42 2 Related Work

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

45 Alghishi et al. [1] evaluated different LLM adaptation techniques: in-context learning, fine-tuning
46 and RAG across open domain, knowledge-grounded and task-based dialogue. They find fine-tuning
47 to provide superior lift compared to ICL while retrieval augmentation moderately improves both ICL
48 and SFT. Notably, they used chat/instruct variants of Llama [18] and Mistral [10] without controlling
49 for model size. Whereas, we start with base models of different sizes and directly fine-tune them on
50 each dataset without relying on chat/instruct variants. Additionally the use of perplexity for evaluation
51 is limiting, as it gives an overall measure of gold text’s uncertainty under the model, whereas we look
52 at fine-grained linguistic dimensions with greater interpretability.

53 Mousavi et al. [12] also conducted dialogue fine-tuning experiments with T5 [17] and GPT-2
54 [16] in Italian. Their work emphasized human evaluation of dimensions such as appropriateness,
55 contextualization, and grammar correctness, which are more informative than perplexity but are quite
56 ad hoc (e.g., “Genericness”) and are not based on any concrete linguistic phenomena. Despite our
57 evaluation using automatic metrics, they are all based on models specifically trained to reproduce
58 human judgements on text annotation tasks.

59 DialogBench [14] is a synthetic dataset of dialogue-related tasks encompassing intent recognition,
60 knowledge grounded generation and coherent infilling, rendered as multiple-choice questions. By
61 contrast, we use organic conversations for which we simultaneously compute multiple metrics
62 each targeting a different aspect of dialogue. They evaluated a large number of pre-trained and
63 chat/instruction-tuned LLMs from different families and sizes and noted base LLMs do well on
64 correctness-related tasks but struggle with coherence and safety. Our evaluation additionally includes
65 analyses of metric reliability.

66 3 Methodology

67 We use the Pythia family of models [3] as the testbed for our fine-tuning experiments. These
68 decoder-only LLMs are pre-trained on 1 trillion tokens of the Pile corpus and come in 10 model
69 sizes, ranging from 14 million to 12 billion parameters — all trained on data presented in the same
70 order to facilitate reproducibility research. To reasonably cover the configuration space, we select the
71 following checkpoints: 140m, 410m, 1.3b, 2.8b, 6.9b (all dedup).

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

75 3.1 Evaluation Metrics

76 Evaluation of dialogue is inherently challenging due to a large space of valid responses to any
77 given utterance. Rather than capturing general quality of generated responses (which is expected to
78 improve through fine-tuning on any corpus), our goal is to disentangle specific linguistic dimensions
79 that contribute to high-quality dialogue. While human evaluation is considered a gold standard
80 for assessing LLMs, it comes with its own set of challenges, including bias, subjectivity, low
81 reproducibility, and high cost. After an extensive survey of automatic metrics, we adopt the following
82 3 model-based metrics that most closely reflect qualities characteristic of dialogue.

²<https://huggingface.co/datasets/databricks/databricks-dolly-15k>

³<https://huggingface.co/datasets/OpenAssistant>

⁴https://huggingface.co/datasets/anon8231489123/ShareGPT_Vicuna_unfiltered

83 **UniEval** [20] is based on the T5 model, finetuned in the to predict entailment between a text and a
 84 set of interpretable attributes, which for dialogue include: Naturalness, Coherence, Engagingness,
 85 Groundedness, and Understandability. Depending on a metric, it can be computed either on the evalu-
 86 ated model’s output alone (e.g., Naturalness and Understandability), model’s output in conjunction
 87 with either the prompt (e.g., Coherence), or an optional context (e.g., Groundedness), and finally all 3
 88 (e.g., Engagingness).

89 **Themis** [8] is a more recent model-based evaluation created by finetuning Llama3-8B [5] on
 90 58 datasets, encompassing 9 generative language tasks, with additional preference alignment and
 91 multi-perspective consistency validation. Themis evaluates the following 4 aspects for dialogue:
 92 Context Maintenance, Interestingness, Knowledge Use, Naturalness. In addition to scores on the
 93 5-point scale, Themis provides reviews to back up its ratings.

94 **GPT-4-as-a-judge.** Evaluation through prompting a strong LLM, such as GPT-4 [13] has emerged
 95 as a new paradigm in NLP. We use this as an opportunity to fine-tune evaluated dimensions to probe
 96 nuanced pragmatic aspects of dialogue not usually captured in generic evaluations. Specifically, we
 97 instructed GPT-4 to assess how well our LLMs are able to follow dialogue turn taking conventions,
 98 recognize user’s intention, generate coherent responses, and keep track of discourse referents. The
 99 complete prompt used in the FastChat format⁵ is included in the appendix.

100 We manually examined 5-10 lowest and highest scoring examples from the datasets for each metric,
 101 concluding they reasonably captured our intended evaluation dimensions (see Sections B.7–B.9).

102 **Open LLM Leaderboard.** While probing the models for dialogue specific competencies, we also
 103 track the average of the common NLP benchmarks, such as MMLU, TruthfulQA, and Winogrande
 104 used in the Open LLM Leaderboard⁶ for ease of comparison — to serve as the base level of model’s
 105 task-specific performance (Figure 5 in appendix).

106 4 Results and Discussion

107 We have run finetuning experiments using the Lit-GPT library⁷ on two NVIDIA RTX A6000’s for
 108 model sizes from 160 million to 2.8 billion parameters, and four A6000’s for the 6.9 billion parameter
 109 model. PyTorch Fully-Sharded Data Parallelism [19] is enabled for the finetuning of 2.8b and 6.9b
 110 parameter models, allowing us to partition weights between GPUs. Depending on the base model
 111 size, we used a learning rate of 3×10^{-5} for the 160m model, 1×10^{-5} for the 410m model, 1×10^{-6}
 112 for all other models with the AdamW [11] optimizer, which we empirically verified to result in best
 113 validation set performance across all datasets. We used a batch size of 128 and fine-tuned for 10
 114 epochs for each experiment, keeping the checkpoint with lowest validation loss as well as the base
 115 and final checkpoints for verification. We initially considered parameter-efficient finetuning using the
 116 adapter approach [9], however it did not provide a tangible lift, therefore in the experiments reported
 117 we used full finetuning. All reported metrics have been computed on the held out test set of each
 118 respective dataset.

119 Quantitative results for each metric are shown in Table 1. For the reasons of space we report the
 120 scores averaged across all 3 datasets. The complete results for each dataset can be found in Section
 121 A.3 of the appendix. We exclude Engagingness from UniEval results as it is an additive quantity with
 122 unrestricted range, with full results in Appendix A.1. In addition we provide score histograms of
 123 each metric for the Pythia 1.4b on the Dolly dataset in Section A.2 of the appendix.

124 Overall, base models generally score the lowest across most metrics, with a slight upward trend with
 125 increasing model size. As could be reasonably expected, fine-tuning consistently provides substantial
 126 gains, moreover larger models benefit from it more, due to their increased capacity. Both Themis and
 127 GPT-4 score trends appear quite uniform, whereas UniEval scores are more irregular. Upon closer
 128 examination, UniEval’s naturalness and understandability follow the expected trend of improving
 129 with finetuning but gain less from increased model size. Average coherence of the base models
 130 starts at a relatively high level of 58% for the smallest model and shows a moderate upward trend,

⁵<https://github.com/lm-sys/FastChat>

⁶https://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

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, it accounts for most of the content in the entire prompt.

Space limitations prevent us from fully addressing all issues raised in the evaluation. High uniformity in scores of Themis and GPT-4 raises a question whether they actually distinguish between different aspects of dialogue. In appendix Section B 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 Appendix B.2) we find evidence for association of Themis metric scores and the most frequent n-grams used to describe them. Lastly, in Appendix B.5 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 Conclusion

In this study we conducted an extensive evaluation of dialogue abilities in the open source Pythia family of models⁸. 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 to 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.

⁸As control, we also perform an evaluation on common NLP benchmarks noting that conversational finetuning does not have a tangible impact on benchmark performance (Figure 5 in the appendix).

References

- [1] S. Alghisi, M. Rizzoli, G. Roccabruna, S. M. Mousavi, and G. Riccardi. Should we fine-tune or rag? evaluating different techniques to adapt llms for dialogue, 2024.
- [2] T. Althoff, K. Clark, and J. Leskovec. 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, 2016.
- [3] S. Biderman, H. Schoelkopf, Q. Anthony, H. Bradley, K. O’Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, A. Skowron, L. Sutawika, and O. van der Wal. Pythia: A suite for analyzing large language models across training and scaling, 2023.
- [4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, 2020.
- [5] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Roziere, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Wyatt, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Guzmán, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Thattai, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. Kloumann, I. Misra, I. Evtimov, J. Zhang, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. van der Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnston, J. Saxe, J. Jia, K. V. Alwala, K. Prasad, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, K. El-Arini, K. Iyer, K. Malik, K. Chiu, K. Bhalla, K. Lakhotia, L. Rantala-Young, L. van der Maaten, L. Chen, L. Tan, L. Jenkins, L. Martin, L. Madaan, L. Malo, L. Blecher, L. Landzaat, L. de Oliveira, M. Muzzi, M. Pasupuleti, M. Singh, M. Paluri, M. Kardas, M. Tsimpoukelli, M. Oldham, M. Rita, M. Pavlova, M. Kambadur, M. Lewis, M. Si, M. K. Singh, M. Hassan, N. Goyal, N. Torabi, N. Bashlykov, N. Bogoychev, N. Chatterji, N. Zhang, O. Duchenne, O. Çelebi, P. Alrassy, P. Zhang, P. Li, P. Vasic, P. Weng, P. Bhargava, P. Dubal, P. Krishnan, P. S. Koura, P. Xu, Q. He, Q. Dong, R. Srinivasan, R. Ganapathy, R. Calderer, R. S. Cabral, R. Stojnic, R. Raileanu, R. Maheswari, R. Girdhar, R. Patel, R. Sauvestre, R. Polidoro, R. Sumbaly, R. Taylor, R. Silva, R. Hou, R. Wang, S. Hosseini, S. Chennabasappa, S. Singh, S. Bell, S. S. Kim, S. Edunov, S. Nie, S. Narang, S. Raparthy, S. Shen, S. Wan, S. Bhosale, S. Zhang, S. Vandenhenne, S. Batra, S. Whitman, S. Sootla, S. Collot, S. Gururangan, S. Borodinsky, T. Herman, T. Fowler, T. Sheasha, T. Georgiou, T. Scialom, T. Speckbacher, T. Mihaylov, T. Xiao, U. Karn, V. Goswami, V. Gupta, V. Ramanathan, V. Kerkez, V. Gonguet, V. Do, V. Vogeti, V. Albiero, V. Petrovic, W. Chu, W. Xiong, W. Fu, W. Meers, X. Martinet, X. Wang, X. Wang, X. E. Tan, X. Xia, X. Xie, X. Jia, X. Wang, Y. Goldschlag, Y. Gaur, Y. Babaei, Y. Wen, Y. Song, Y. Zhang, Y. Li, Y. Mao, Z. D. Coudert, Z. Yan, Z. Chen, Z. Papakipos, A. Singh, A. Srivastava, A. Jain, A. Kelsey, A. Shajnfeld, A. Gangidi, A. Victoria, A. Goldstand, A. Menon, A. Sharma, A. Boesenberg, A. Baevski, A. Feinstein, A. Kallet, A. Sangani, A. Teo, A. Yunus, A. Lupu, A. Alvarado, A. Caples, A. Gu, A. Ho, A. Poulton, A. Ryan, A. Ramchandani, A. Dong, A. Franco, A. Goyal, A. Saraf, A. Chowdhury, A. Gabriel, A. Bharambe, A. Eisenman, A. Yazdan, B. James, B. Maurer, B. Leonhardi, B. Huang, B. Loyd, B. D. Paola, B. Paranjape, B. Liu, B. Wu, B. Ni, B. Hancock, B. Wasti, B. Spence, B. Stojkovic, B. Gamido, B. Montalvo, C. Parker, C. Burton, C. Mejia, C. Liu, C. Wang, C. Kim, C. Zhou, C. Hu, C.-H. Chu, C. Cai, C. Tindal, C. Feichtenhofer, C. Gao, D. Civin, D. Beaty, D. Kreymer, D. Li, D. Adkins, D. Xu, D. Testuggine, D. David, D. Parikh, D. Liskovich, D. Foss, D. Wang, D. Le, D. Holland, E. Dowling, E. Jamil, E. Montgomery, E. Presani, E. Hahn, E. Wood, E.-T. Le, E. Brinkman, E. Arcaute, E. Dunbar, E. Smothers, F. Sun, F. Kreuk, F. Tian, F. Kokkinos, F. Ozgenel, F. Caggioni, F. Kanayet, F. Seide, G. M. Florez, G. Schwarz, G. Badeer, G. Swee, G. Halpern, G. Herman, G. Sizov, Guangyi, Zhang, G. Lakshminarayanan, H. Inan, H. Shojanazeri, H. Zou,

- 218 H. Wang, H. Zha, H. Habeeb, H. Rudolph, H. Suk, H. Aspegren, H. Goldman, H. Zhan,
219 I. Damlaj, I. Molybog, I. Tufanov, I. Leontiadis, I.-E. Veliche, I. Gat, J. Weissman, J. Geboski,
220 J. Kohli, J. Lam, J. Asher, J.-B. Gaya, J. Marcus, J. Tang, J. Chan, J. Zhen, J. Reizenstein,
221 J. Teboul, J. Zhong, J. Jin, J. Yang, J. Cummings, J. Carvill, J. Shepard, J. McPhie, J. Torres,
222 J. Ginsburg, J. Wang, K. Wu, K. H. U, K. Saxena, K. Khandelwal, K. Zand, K. Matosich,
223 K. Veeraraghavan, K. Michelena, K. Li, K. Jagadeesh, K. Huang, K. Chawla, K. Huang,
224 L. Chen, L. Garg, L. A. L. Silva, L. Bell, L. Zhang, L. Guo, L. Yu, L. Moshkovich, L. Wehrstedt,
225 M. Khabsa, M. Avalani, M. Bhatt, M. Mankus, M. Hasson, M. Lennie, M. Reso, M. Groshev,
226 M. Naumov, M. Lathi, M. Keneally, M. Liu, M. L. Seltzer, M. Valko, M. Restrepo, M. Patel,
227 M. Vyatskov, M. Samvelyan, M. Clark, M. Macey, M. Wang, M. J. Hermoso, M. Metanat,
228 M. Rastegari, M. Bansal, N. Santhanam, N. Parks, N. White, N. Bawa, N. Singhal, N. Egebo,
229 N. Usunier, N. Mehta, N. P. Laptev, N. Dong, N. Cheng, O. Chernoguz, O. Hart, O. Salpekar,
230 O. Kalinli, P. Kent, P. Parekh, P. Saab, P. Balaji, P. Rittner, P. Bontrager, P. Roux, P. Dollar,
231 P. Zvyagina, P. Ratanchandani, P. Yuvraj, Q. Liang, R. Alao, R. Rodriguez, R. Ayub, R. Murthy,
232 R. Nayani, R. Mitra, R. Parthasarathy, R. Li, R. Hogan, R. Battey, R. Wang, R. Howes, R. Rinott,
233 S. Mehta, S. Siby, S. J. Bondu, S. Datta, S. Chugh, S. Hunt, S. Dhillon, S. Sidorov, S. Pan,
234 S. Mahajan, S. Verma, S. Yamamoto, S. Ramaswamy, S. Lindsay, S. Lindsay, S. Feng, S. Lin,
235 S. C. Zha, S. Patil, S. Shankar, S. Zhang, S. Zhang, S. Wang, S. Agarwal, S. Sajuyigbe,
236 S. Chintala, S. Max, S. Chen, S. Kehoe, S. Satterfield, S. Govindaprasad, S. Gupta, S. Deng,
237 S. Cho, S. Virk, S. Subramanian, S. Choudhury, S. Goldman, T. Remez, T. Glaser, T. Best,
238 T. Koehler, T. Robinson, T. Li, T. Zhang, T. Matthews, T. Chou, T. Shaked, V. Vontimitta,
239 V. Ajayi, V. Montanez, V. Mohan, V. S. Kumar, V. Mangla, V. Ionescu, V. Poenaru, V. T.
240 Mihailescu, V. Ivanov, W. Li, W. Wang, W. Jiang, W. Bouaziz, W. Constable, X. Tang, X. Wu,
241 X. Wang, X. Wu, X. Gao, Y. Kleinman, Y. Chen, Y. Hu, Y. Jia, Y. Qi, Y. Li, Y. Zhang,
242 Y. Zhang, Y. Adi, Y. Nam, Yu, Wang, Y. Zhao, Y. Hao, Y. Qian, Y. Li, Y. He, Z. Rait, Z. DeVito,
243 Z. Rosnbrick, Z. Wen, Z. Yang, Z. Zhao, and Z. Ma. The llama 3 herd of models, 2024.
- 244 [6] A. Gudibande, E. Wallace, C. Snell, X. Geng, H. Liu, P. Abbeel, S. Levine, and D. Song. The
245 false promise of imitating proprietary llms, 2023.
- 246 [7] L. Horn and G. Ward. *The Handbook of Pragmatics*. Basil Blackwell, 2004.
- 247 [8] X. Hu, L. Lin, M. Gao, X. Yin, and X. Wan. Themis: A reference-free NLG evaluation
248 language model with flexibility and interpretability. In Y. Al-Onaizan, M. Bansal, and Y.-
249 N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural
250 Language Processing*, pages 15924–15951, Miami, Florida, USA, Nov. 2024. Association for
251 Computational Linguistics.
- 252 [9] Z. Hu, L. Wang, Y. Lan, W. Xu, E.-P. Lim, L. Bing, X. Xu, S. Poria, and R. K.-W. Lee.
253 Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models,
254 2023.
- 255 [10] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand,
256 G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril,
257 T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7b, 2023.
- 258 [11] I. Loshchilov and F. Hutter. Decoupled weight decay regularization, 2019.
- 259 [12] S. M. Mousavi, G. Roccabruna, M. Lorandi, S. Caldarella, and G. Riccardi. Evaluation
260 of response generation models: Shouldn’t it be shareable and replicable? In A. Bosselut,
261 K. Chandu, K. Dhole, V. Gangal, S. Gehrmann, Y. Jernite, J. Novikova, and L. Perez-Beltrachini,
262 editors, *Proceedings of the Second Workshop on Natural Language Generation, Evaluation,
263 and Metrics (GEM)*, pages 136–147, Abu Dhabi, United Arab Emirates (Hybrid), Dec. 2022.
264 Association for Computational Linguistics.
- 265 [13] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida,
266 J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Bal-
267 tescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner,
268 L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage,
269 K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan,
270 C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho,
271 C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch,

- 272 D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou,
273 D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson,
274 V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene,
275 J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse,
276 A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain,
277 S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, Łukasz
278 Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim,
279 Y. Kim, J. H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Łukasz Kondraciuk, A. Kondrich,
280 A. Konstantinidis, K. Kosc, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung,
281 D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju,
282 K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. Mc-
283 Grew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick,
284 L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati,
285 O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang,
286 C. O’Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish,
287 E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. de Avila Belbute Peres, M. Petrov,
288 H. P. de Oliveira Pinto, Michael, Pokorny, M. Pokrass, V. H. Pong, T. Powell, A. Power,
289 B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach,
290 C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry,
291 H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker,
292 P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song,
293 N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. B. Thompson,
294 P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone,
295 A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei,
296 C. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter,
297 S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu,
298 Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk,
299 and B. Zoph. Gpt-4 technical report, 2024.
- 300 [14] J. Ou, J. Lu, C. Liu, Y. Tang, F. Zhang, D. Zhang, and K. Gai. Dialogbench: Evaluating llms as
301 human-like dialogue systems, 2024.
- 302 [15] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal,
303 K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder,
304 P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with
305 human feedback, 2022.
- 306 [16] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are
307 unsupervised multitask learners. 2019.
- 308 [17] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu.
309 Exploring the limits of transfer learning with a unified text-to-text transformer, 2023.
- 310 [18] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra,
311 P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu,
312 J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini,
313 R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A.
314 Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra,
315 I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M.
316 Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan,
317 I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and
318 T. Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.
- 319 [19] Y. Zhao, A. Gu, R. Varma, L. Luo, C.-C. Huang, M. Xu, L. Wright, H. Shojanazeri, M. Ott,
320 S. Shleifer, A. Desmaison, C. Balioglu, P. Damania, B. Nguyen, G. Chauhan, Y. Hao, A. Math-
321 ews, and S. Li. Pytorch fsdp: Experiences on scaling fully sharded data parallel, 2023.
- 322 [20] M. Zhong, Y. Liu, D. Yin, Y. Mao, Y. Jiao, P. Liu, C. Zhu, H. Ji, and J. Han. Towards a unified
323 multi-dimensional evaluator for text generation, 2022.
- 324 [21] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, S. Zhang, G. Ghosh,
325 M. Lewis, L. Zettlemoyer, and O. Levy. Lima: Less is more for alignment, 2023.

326 6 Limitations

327 This study has a number of limitations due to a combination of practical constraints and a limited
328 computational budget.

- 329 • Pythia as the only evaluated family of models raises a question whether different architec-
330 tures, training data and recipes could affect dialogue abilities in different ways.
- 331 • Our focus on model sizes under 8B parameters could be providing an incomplete picture
332 e.g., emergence of dialogue abilities in sufficiently large models entirely without finetuning.
- 333 • Supervised finetuning is only one of the options available, including RLHF, DPO, and
334 others.
- 335 • To simplify the logistics of training and evaluation data generation and to facilitate the use
336 of metrics, we consider single dialogue turns, whereas true multi-turn and interactive human
337 evaluation could provide deeper insights.

338 A Appendix

339 A.1 Full UniEval Results

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 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 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

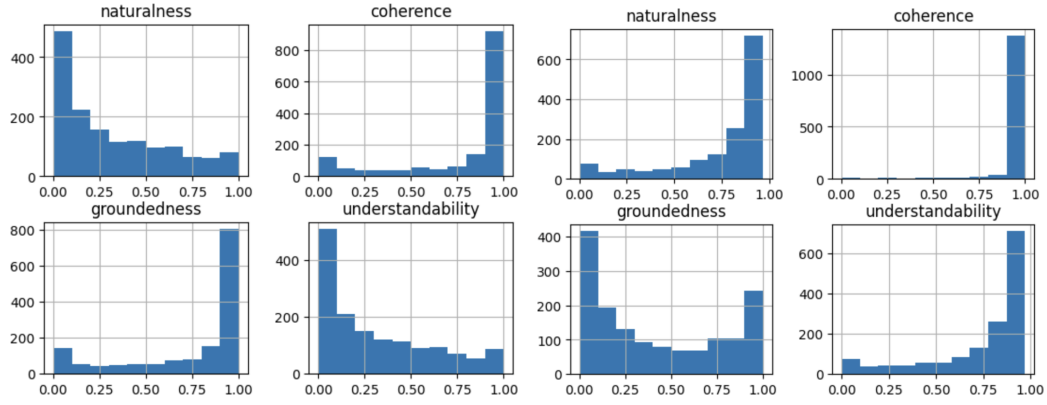


Figure 1: UniEval scores histograms for base (left) and finetuned (right) Pythia 1.4b on Dolly

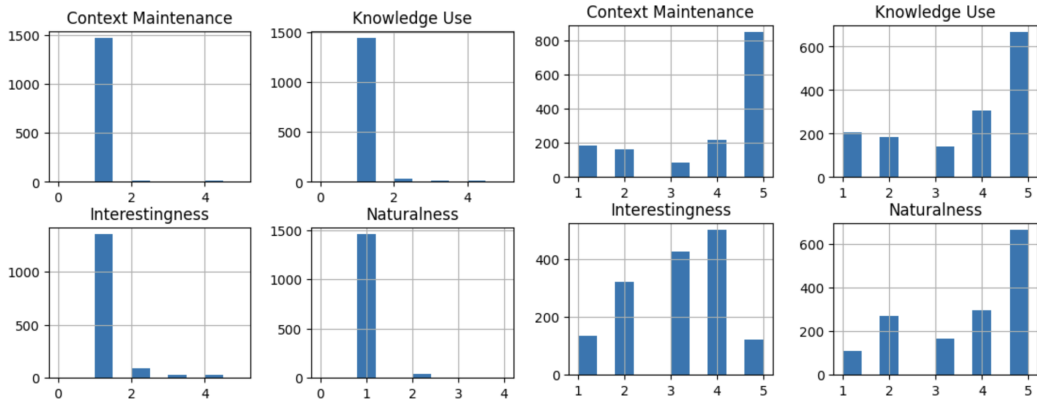


Figure 2: Themis scores histograms for base (left) and finetuned (right) Pythia 1.4b on Dolly

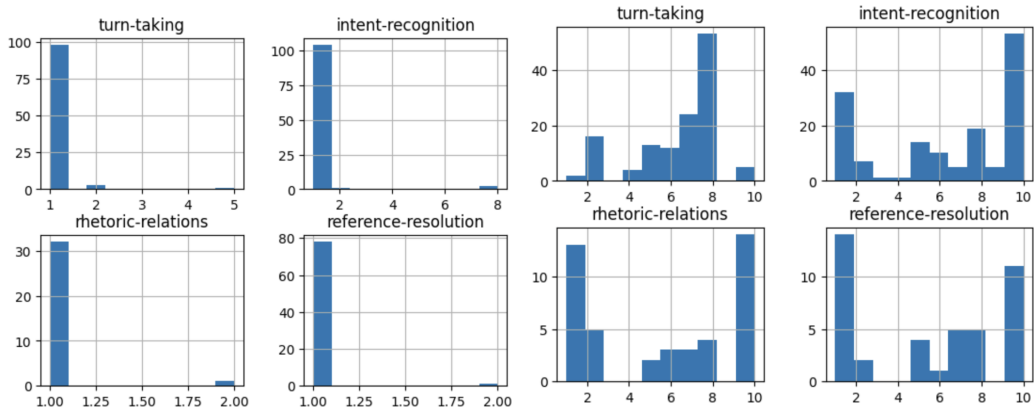
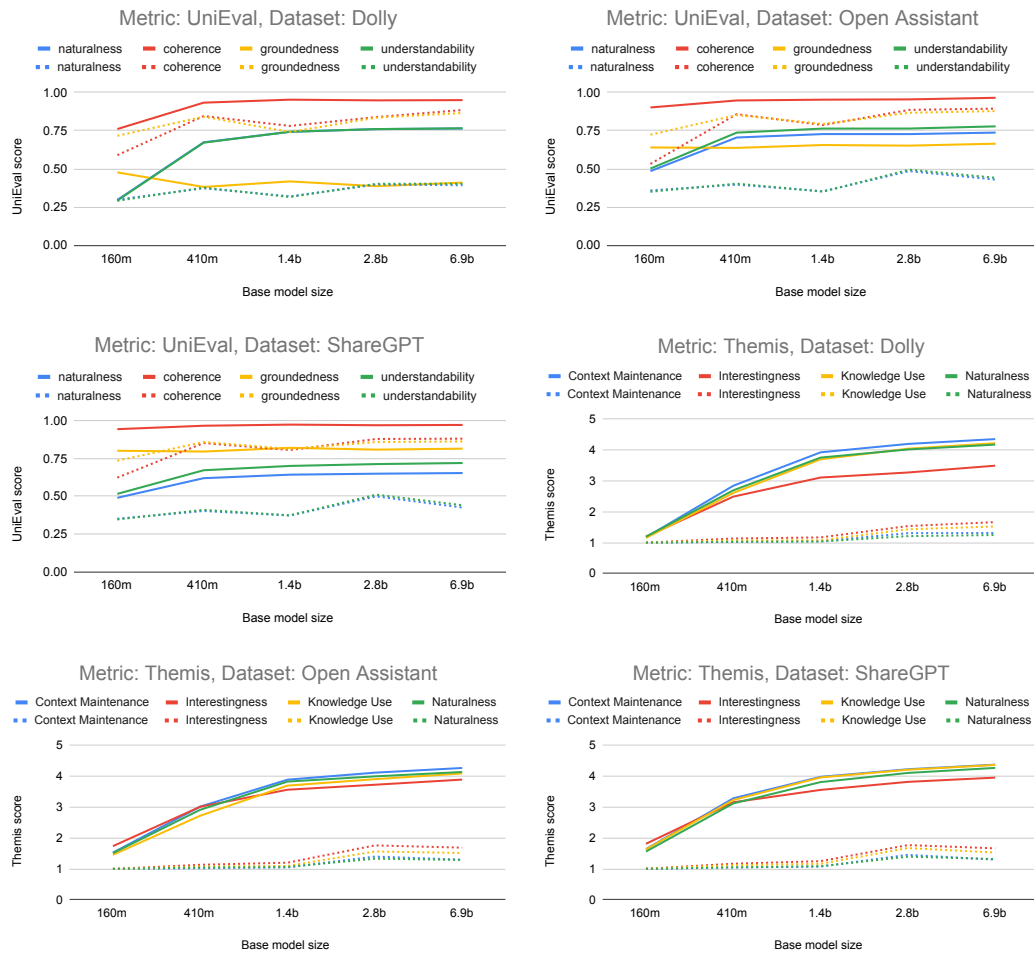


Figure 3: GPT-4 scores histograms for base (left) and finetuned (right) Pythia 1.4b on Dolly

341 **A.3 Complete Results by Dataset**



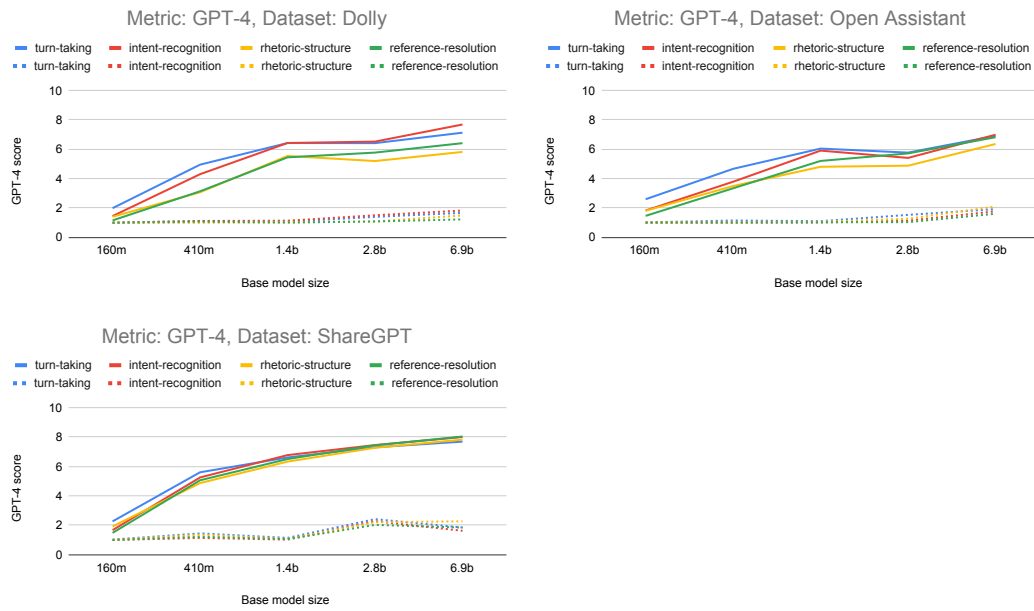


Figure 4: Dialogue metrics vs base model size after finetuning on each dataset. Dotted lines correspond to measurements of the base models prior to finetuning. Best viewed in color.

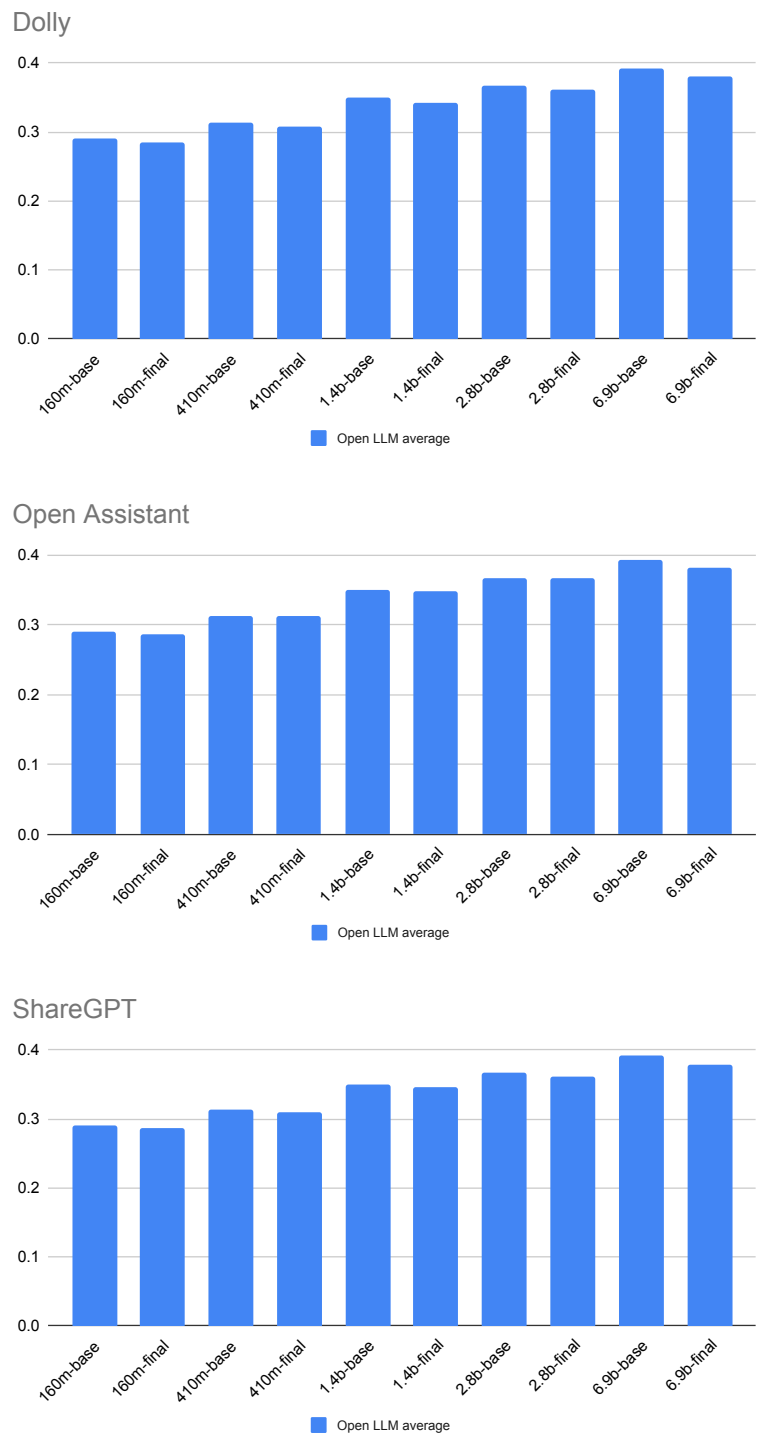


Figure 5: Open LLM Leaderboard Evaluation

		UniEval						Themis				GPT-4			
		naturalness	coherence	engagingness	groundedness	understandability	overall	Context Maintenance	Interestingness	Knowledge Use	Naturalness	turn-taking	intent-recognition	rhetoric-relations	reference-resolution
UniEval	naturalness	1.00000	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	1.00000	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	1.00000	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	1.00000	-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	1.00000	-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	1.00000	-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	1.00000	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	1.00000	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	1.00000	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	1.00000	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	1.00000	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	1.00000	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	1.00000	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	1.00000

Figure 6: Pearson Correlation Measured among Metrics across all Examples

B Analysis

B.1 Metric Correlation

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 6. 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.

B.2 Mining Rating Explanations

Motivated by the findings from correlation analysis, we leverage explanations provided by Themis in 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 ngrams that are associated with either high or low ratings for each metric. The complete list is provided in Section B.4 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.,

376 “does not recognize the user’s [intent]” comes up when turn taking receives a low rating, or “follows
377 dialogue conventions” is spuriously associated with a high score in reference resolution.

378 It should be noted that in both scenarios the explanation, whether or not it agrees with the rating
379 in the corresponding dimension, cannot be regarded as a real a rationale nor a cause for that rating,
380 given the generative nature of this assessment.

381 B.3 Word Overlap and Diversity

382 As we aimed to motivate in Figure 1, base model responses (even if fluent) tend to be of low diversity
383 and repetitive, without providing new content beyond what is already contained in the prompt.
384 Additionally, repetitiveness surfaced in our analysis of rating explanations in Section B.2 as resulting
385 in low interestingness. On the other hand, responses that correctly recognize the user’s intent, exhibit
386 some degree of overlap with the user’s utterance through acknowledging it in the response. While
387 high overlap by itself does not automatically imply fluent dialogue, some overlap is a necessary
indication of mirroring in conversation (cf. — [2]). For example:

PROMPTER: “Write a C++ function that takes a reference to a std::string containing markdown
formatted text and returns a std::string containing html formatted text.”

ASSISTANT: “Sure, here’s a C++ function that takes a reference to a std::string . . . , etc.”

388

389 We operationalize diversity as the ratio of vocabulary size to the length of the response, and overlap
390 as the length of the longest common subsequence of tokens shared with the prompt, normalized by
391 the length of the response. Appendix section B.5 contains these simple heuristics computed for each
392 model size and dataset, across base and finetuned checkpoints. As can be noted already from the
393 smallest models e.g., Pythia-160m response overlap of the base model tends to be low e.g., 0.0104,
394 whereas in the finetuned model of the same size it reaches 0.1424, which nearly matches the level of
395 the gold responses from the dataset (e.g., 0.1525 for Dolly). For the other two datasets, the overlap
396 of the gold responses is lower, at 0.087 and 0.064 for Open Assistant and ShareGPT, respectively.
397 However, the overlap of the LLM responses follows the same trend: starting significantly lower with
398 the base model, and reaching the dataset’s gold level after finetuning.

399 Response diversity of the base models lies in the range 0.25 – 0.35, whereas after finetuning it more
400 than doubles, approaching the level of gold responses (0.65 – 0.82), even though not quite matching
401 them. Taken together, these trends provide a plausible explanation of the previously observed steady
402 increases in metrics such as coherence, intent recognition.

403 B.4 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 2: 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 3: Top Frequent 5-grams for High and Low Rating along each Dimension Measured by GPT-4

404 B.5 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

405 B.6 Qualitative Observations

406 To help validate the quantitative findings we manually reviewed a small number of example genera-
407 tions scoring high and low on each of the reported metrics. We specifically focused on high-scoring
408 examples from the finetuned smaller 410m model, and contrast them with low-scoring examples
409 obtained for the 6.9b base model. Space limitations prevent us from citing extensive examples in
410 the main text, and instead we provide a comprehensive set of examples in Sections B.7 through B.9
411 of the appendix. By and large UniEval, Themis, and GPT-4-based metrics are reliable indicators of
412 generation quality. Still, a high score in any of these metrics does not guarantee the generations to
413 be completely sensible or free of obvious hallucinations. In addition, we have come across a small
414 number of clearly incoherent examples that nonetheless obtained high scores in UniEval.

415 When it comes to differentiating among each metric’s fine grained dimensions, while in some cases
416 different dimensions clearly signal distinct linguistic phenomena (e.g., GPT-4 based turn taking
417 and intent recognition), there does appear to be a correlation among many of them. For example,
418 responses scoring highly on naturalness also exhibited high coherence and vice versa. This is not
419 surprising, given it is not always possible even for a human to pinpoint the exact quality that a certain
420 example exhibits. Similarly, low scores do not necessarily reflect a deficiency in a particular measured
421 aspect but rather a response’s overall low quality. Additional observations include metrics such as
422 Engagingness, Interestingness, and Knowledge Use could be prone to heuristics, such as number of
423 facts mentioned rather than their relevance to the prompt.

424 In the next section we aim to confirm our intuitions by conducting two additional quantitative
425 analyses.

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 4: 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 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 5: Examples of base Pythia-6.9b scoring low on UniEval

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

dimension	prompt	response
Context Maintenance	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?	$= -2*y - 2*p + p + 7,$ $0 = -5*y + 2*p + 11$ 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 So 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 $0 = -5*a + 4*o - 17, -5*a + 0*a - 4*o = -1$ [...]
Interestingness	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. '	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
Knowledge Use	Classify each of these car manufacturers as either French, German, American, Japanese, or other: Audi, Peugeot, Aston Martin, Ssangyong, Chevrolet, BMW	French: Aston Martin, Peugeot Danish: Audi, Ssangyong German: Aston Martin, BMW American: Chevrolet, Ssangyong Japanese: Aston Martin, Ssangyong: Peugeot
Naturalness	Who is considered the bachelor president of the United States?	In the United States of America, the president of the United States is the president.

Table 7: Examples of low-scoring responses from finetuned 410m Pythia according to Themis

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 8: 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 America. 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 America. [...]
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 9: Examples of low-scoring responses of a base Pythia-6.9b according to GPT4 assessment

```

1  {
2      "prompt_id": 2,
3      "system_prompt": "You are a helpful and precise assistant for
4      checking the dialogue qualities of an AI assistant.",
5      "prompt_template": "[Question]
6                          {question}
7
8                          [The Start of Assistant's Answer]
9                          {answer}
10
11                         [The End of Assistant's Answer]
12
13                         [System]
14                         {prompt}]",
15      "defaults": {
16          "prompt": "We would like to request your feedback on AI
17          assistant's ability to engage in natural dialogue with the
18          user, in the style similar to an attentive, courteous and
19          empathetic customer support agent, based on the above
20          exchange.
21          Please provide ratings on a 10-point scale based assistant's
22          exhibited abilities in the following:
23          1) follows dialogue conventions such as turn taking,
24          acknowledging the speaker, and signaling investment
25          in the conversation;
26          2) recognizes the user's intent, and appropriately
27          acknowledges it in its response;
28          3) understands rhetoric relations between user's and
29          assistant's utterances;
30          4) keeps track of entites and facts mentioned throughout the
31          dialogue, and is able to effortlessly refer back to them or
32          understand the user's references to them.
33          Please first output a single line containing comma separated
34          scores as integers on the above dimensions for the Assistant.
35          If assistant's response does not reflect sufficient evidence
36          for any of the criteria, output n/a for those.
37          In the subsequent lines, please provide a succinct explanation
38          of your evaluation for each criterion, avoiding any potential
39          bias and not evaluating any other qualities beyond ones
40          explicitly asked for in these instructions."
41      },
42      "description": "Prompt for general questions",
43      "category": "general"
44  }

```

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

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: Claims made in abstract are supported in main body of the paper as well as appendices.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: Limitations are included after references.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[NA\]](#)

Justification: No theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: Model, datasets, fine-tuning hyper parameter details are shared.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Due to employer's legal limitations code and data can be only provided upon request.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We include all hyperparameters.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Mean values are reported.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide detailed information about infrastructure used for model training.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Research conforms to code of ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: Due to space limitations.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [No]

Justification: Paper contributes an evaluation study.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: Common open source models and data are used and cites them.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: References to model paper and links to dataset cards provided.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Crowdsourcing not used in paper.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Crowdsourcing not used in paper.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

739 Question: Does the paper describe the usage of LLMs if it is an important, original, or
740 non-standard component of the core methods in this research? Note that if the LLM is used
741 only for writing, editing, or formatting purposes and does not impact the core methodology,
742 scientific rigorousness, or originality of the research, declaration is not required.

743 Answer: [NA]

744 Justification: No LLM use during research or paper preparation.

745 Guidelines:

- 746 • The answer NA means that the core method development in this research does not
- 747 involve LLMs as any important, original, or non-standard components.
- 748 • Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for
- 749 what should or should not be described.