Curiosity-16: A 354.8M Parameter Large Language Model

Arian Kharazmi *November 2025*

Abstract

Despite their age, GPT-2 models remain among the most downloaded open-source large language models. With a 2019 knowledge cutoff, and the tendency of GPT-2 models to hallucinate or misinterpret, these models face significant drawbacks. Using GPT-2 Medium (354.8m parameters) as the foundational model, we release Curiosity-16 (C16), a 354.8 million parameter large language model that utilizes a two-phase supervised fine-tuning (SFT) pipeline for increased domain-specific accuracy, reasoning, and recent knowledge injection. Using EleutherAI's LM Eval Harness, we evaluated Curiosity-16 and GPT-2 Medium on the HellaSwag and Massive Multitask Language Understanding (MMLU) benchmarks. For zero-shot HellaSwag, Curiosity-16 shows a +0.29-percentage point increase in normalized accuracy over GPT-2 Medium, and for MMLU, Curiosity-16 shows a +0.85-percentage point increase over GPT-2 Medium for normalized accuracy. However, subject-level gains are more pronounced. Our targeted fine-tuning pipeline affirms that model performance enhancements can be made with limited hardware and publicly available resources.

1. Introduction

Research in contemporary LLMs largely revolves around large frontier state-of-the-art models trained on massive high-end GPUs. These models are impressive but are not accessible from a research standpoint for most. The GPT-2 family of models, largely antiquated by today's standards, are still highly accessible models for most people who rely on consumer-grade hardware. First introduced by Radford et al. (2019), GPT-2 remains as a popular model with well-documented and understood architectures.

While fine-tuning GPT-2 Medium may be appealing, there are several limitations to consider when working with an older model. Base pre-trained GPT-2 models remain frozen in time with a knowledge cutoff from 2019, are more prone to hallucination, misinterpretation, and lack much of the sophistication of modern state-of-the-art LLMs, such as MoE (Mixture-of-Experts) architecture.

While there are a number of drawbacks to using GPT-2 Medium rather than a more modern, smaller, and quantized model such as LLaMa-3-8B or Qwen-3-4B, GPT-2 Medium provides a baseline that allows our supervised two-phase fine-tuning process to perform in isolation, rather than relying on modern architectural advantages for better evaluation results.

The purpose of our work was to ask a fundamental question: *Can a multi-phase supervised fine-tuning process on consumer-grade hardware and publicly available resources yield improved capabilities for a GPT-2-class model?*

To answer this, we present ***Curiosity-16***, a 354.8 million parameter model built upon GPT-2 Medium and a curated suite of HuggingFace datasets. Our contribution is a two-phase supervised fine-tuning (SFT) pipeline designed to inject recent world knowledge and specialization in reasoning capabilities.

Our aim with the development and evaluation of Curiosity-16 is to empirically demonstrate that quantifiable improvements to LLM performance are attainable with consumer hardware and publicly available resources without large compute budgets, infrastructure, and closed data sources.

2. Related Work

Our work uses the foundational Transformer architecture (Vaswani et al., 2017) and the GPT-2 series of models (Radford et al., 2019). Supervised fine-tuning has become a ubiquitous standard for creating custom LLM agents for specific tasks and uses (e.g., instruction-following). (Touvron et al., 2023). For evaluating models, we relied on EleutherAI's `lm-evaluation-harness`, an open-source framework for benchmarking and evaluating large language models, which supports many academic benchmarks, such as HellaSwag and the Massive Multitask Language Understanding (MMLU) benchmarks. (Gao et al., 2021).

Curiosity-16 relies on eleven curated datasets.

For Phase I (Knowledge Generalization), we utilized OpenAssistant OASST1 (Kopf et al., 2023) for conversational focus; FineWeb-BBC-News (Penedo et al., 2024) providing C16 with recent context; Wikipedia-2025 (Wikimedia Foundation, 2025) for modern world knowledge; ELI5 (Fan et al., 2019) for question answering; SQuAD v2 (Rajpurkar et al., 2018) for reading comprehension; and Dolly_15k (Databricks, 2023) for instruction following.

For Phase II (Reasoning and Task Specialization), we utilized multiple AGIEval (Zhong et al., 2023) datasets (LSAT-AR, LSAT-LR-LSAT-RC) for logic and reasoning capabilities; GSM8K (Cobbe et al., 2021) for mathematical reasoning; and Alpaca-Cleaned (Taori et al., 2023) for generalized instruction-following.

3. Methodology

Initial Script

We utilized Python with the PyTorch Machine Learning Framework and the HuggingFace Transformers and Dataset libraries for our baseline scripts.

Foundational Pre-trained Model

Curiosity-16 (354.8M parameters) uses `openai-community/gpt2-medium` (GPT-2 Medium) (354.8m parameters) as its pre-trained foundational model. Experiments were constrained to accessible consumer-grade hardware specifications (Apple Silicon M4 Mac, 16GB Unified Memory) to ensure broad accessibility and reproducibility on ubiquitous compute processing standards. We used eleven diverse and curated HuggingFace datasets arranged in a two-phase regimen. This resulted in ~153k training samples being used in the SFT pipeline.

Tokenizer

We opted to use AutoTokenizer from the HuggingFace Transformers library for choosing the fastest and most robust tokenizer for tokenization sequences. The selected tokenizer was GPT2Tokenizer.

Datasets

Phase I: Knowledge Generalization

[1] Phase I: For Phase I, six datasets were chosen for knowledge generalization, so C16 could have strong knowledge recall and access to recent facts and events. (~67,000 samples total.)

Dataset Name	Dataset Focus	Training Sample Size
oasst1	Open-ended conversation	12,000 samples
fineweb-bbc-news	Factual News Data	15,000 samples
wikipedia-20250620	Knowledge Generalization	12,000 samples
eli5	Long-form Q&A	8,000 samples
squad_v2	Reading Comprehension	8,000 samples
dolly-15k-instruction-alpaca-format	Instruction-following	12,000 samples

Phase II: Reasoning Capabilities

[2] Phase II: For Phase II, five datasets were chosen for task-focus, reasoning, and basic Chain-of-Thought capabilities for Curiosity-16. (~86,000 samples total.)

Dataset Name	Dataset Focus	Training Sample Size
agieval_lsat_ar	Analytical Reasoning	8,000 samples
agieval_lsat_lr	Logical Reasoning	8,000 samples
agieval_lsat_rc	Reading Comprehension	8,000 samples
gsm8k	Math World Problems	10,000 samples
alpaca-cleaned	Generalized Instruction-following	52,000 samples

Preprocessing

All data was tokenized via GPT-2 Tokenizer, truncated to 1024 tokens, right-padded, 90/10 train/validation split. Prompts formatted as *Instruction -> Response*.

A small subset of the preprocessing logic for the Dolly and SQuAD_v2 datasets relating to prompt formatting was assisted by an AI language model under strict supervision in order to standardize the training structure of the six Phase I datasets during the training process. This code was manually reviewed and refined prior to integration into the Phase I script.

All final scripts are publicly available in the Curiosity-16 GitHub repository.

Training

Parameter	Phase I	Phase II
Learning Rate	1e-5	1e-5
Epochs	3	2
Batch Size	4	4
Gradient Accumulation	4 (Effective Batch Size: 16)	2 (Effective Batch Size: 8)
LR Scheduler	Linear	Linear, 500 warmup steps
Optimizer	AdamW	AdamW
Weight Decay	0.01	0.01
Early Stopping	N/A	Threshold 0.01, Patience: 2
Precision	FP32	FP32
Epochs Completed	3.0	2.0
Training Loss	2.3217	1.6755
Total Steps	12,375	12,524
Evaluation Loss	2.8509	1.6144
Train Runtime	~17.3 hours	~13.0 hours
Throughput (Samples)	3.369 samples/second	2.149 samples/second
Throughput (Steps)	0.2111 steps/second	0.269 steps/second
Energy Consumption (Per Phase)	~0.74 kWh	~0.59 kWh

Training and Resource Dynamics

Phase I runtime completed in 17.3 hours. Phase II runtime completed in ~13 hours. Complete C16 runtime was 30.3 hours for both phases. Phase I runtime (~17.3) includes an initial ~1-hour partial run that was interrupted then restarted.

Phase II's eval loss rate represents a clear improvement over Phase I (43.37% drop in eval loss rate from Phase I to Phase II). Best model loaded from Phase I and Phase II checkpoints based on eval loss.

Training process was stable and did not encounter extreme values. (e.g., 3.0-5.0).

Weights were updated, but parameter count remained the same.

Broader architecture, such as layer count, hidden sizes, and head count remained the same between the two models. (24 layers, 1024 hidden size, 16 heads.)

CodeCarbon was used as a tool to measure energy usage during Phase I and Phase II training of Curiosity-16, resulting in ~1.33kWh of energy consumed during training runtime.

The Curiosity-16 model as released is based off of the best performing Phase II checkpoint.

4. Evaluation

In our experiments, we opted to use popular evaluation benchmarks MMLU (zero-shot) and HellaSwag (zero-shot) in our trials comparing Curiosity-16 to GPT-2 Medium via lm-evaluation-harness.

All accuracies and CIs have been rounded to four decimal places; **Δ** values have also been rounded to four decimal places.

For all reported confidence intervals, we use LM-Eval-Harness's built-in non-parametric bootstrap over question-level accuracies, assessed with 100,000 bootstrap iterations (`bootstrap_iters = 100000`). LM-Eval returns an estimated standard error (stderr) computed from this bootstrap distribution, for each metric and subject in both benchmarks. We then compute two-sided 95% confidence intervals as *mean ± 1.96 x stderr* for both models and each subject. A small Python script automates this calculation to the LM-Eval-Harness logs and formats the upper and lower bounds reported in the tables. Confidence intervals generally overlap due to sample size limitations in individual MMLU subjects.

MMLU and HellaSwag results both report 'acc' and 'acc_norm', which is interpreted as 'Raw Accuracy' and 'Normalized Accuracy'. **Δ** is the normalized accuracy values of Curiosity-16 minus the accuracy values of GPT-2 Medium. 'stderr' is interpreted as 'Standard Error'.

LM-Eval's "acc_norm" metric applies answer-choice normalization for reducing answer-choice bias. Gao et al. (2021).

Positive deltas indicate an increase in normalized accuracy from GPT-2 Medium to Curiosity-16, and negative deltas indicate a decrease in normalized accuracy from GPT-2 Medium to Curiosity-16.

Repeated LM-Eval evaluations showed variance of < 0.001 across both GPT-2 Medium and Curiosity-16, confirming stability between different seed tests.

4.1 HellaSwag (zero-shot):

Table 1: Overall HellaSwag Results

Metric	GPT-2 Medium	95% CI	Curiosity-16	95% CI	Δ
Normalized Accuracy	0.3938	[0.3843,0.4034]	0.3967	[0.3872,0.4063]	+0.0029

Table shows zero-shot HellaSwag results comparing Curiosity-16 to GPT-2 Medium. Normalized accuracy shows a +0.29-percentage point increase from GPT-2 Medium to Curiosity-16.

Interpretation: Small but consistent improvements indicate overall knowledge generalization without overfitting.

4.2 MMLU (zero-shot):

Table: Overall MMLU Results

Metric	GPT-2 Medium	95% CI	Curiosity-16	95% CI	Δ
Normalized Accuracy	0.2289	[0.2220,0.2359]	0.2375	[0.2305,0.2445]	+0.0085

Table: MMLU Groups Table

Subject Group	GPT-2 Medium	95% CI	Curiosity-16	95% CI	Δ
mmlu	0.2290	[0.2220,0.2359]	0.2375	[0.2305,0.2445]	+0.0085
mmlu_stem	0.2128	[0.1985,0.2271]	0.2173	[0.2029,0.2317]	+0.0044
mmlu_social_sciences	0.2184	[0.2038,0.2330]	0.2197	[0.2051,0.2343]	+0.0013
mmlu_humanities	0.2427	[0.2305,0.2550]	0.2557	[0.2432,0.2682]	+0.0130
mmlu_other	0.2350	[0.2201,0.2498]	0.2481	[0.2330,0.2633]	+0.0132

Table: Subject-level MMLU Increases

Subject	GPT-2 Medium	95% CI	Curiosity-16	95% CI	Δ
mmlu_college_physics	0.1765	[0.1021,0.2508]	0.2451	[0.1612,0.3289]	+0.0686
mmlu_professional_medicine	0.1838	[0.1377,0.2299]	0.2426	[0.1916,0.2936]	+0.0588
mmlu_medical_genetics	0.3	[0.2097,0.3902]	0.35	[0.2560,0.4439]	+0.05

mmlu_high_school_european_history	0.2182	[0.1549,0.2814]	0.2667	[0.1990,0.3343]	+0.0485
mmlu_anatomy	0.1926	[0.1258,0.2593]	0.2370	[0.1650,0.3090]	+0.0444
mmlu_college_computer_science	0.24	[0.1559,0.3241]	0.28	[0.1916,0.3684]	+0.0400
mmlu_logical_fallacies	0.2209	[0.1570,0.2847]	0.2577	[0.1903,0.3250]	+0.0368
mmlu_high_school_geography	0.1818	[0.1280,0.2357]	0.2171	[0.1596,0.2747]	+0.0354
mmlu_high_school_statistics	0.1574	[0.1087,0.2061]	0.1898	[0.1374,0.2422]	+0.0324
mmlu_nutrition	0.1863	[0.1426,0.2300]	0.2157	[0.1695,0.2618]	+0.0294

Table shows the top ten most dramatic increases in subject-level accuracy in specific subjects between Curiosity-16 and GPT-2 Medium.

Table: Subject-level MMLU Decreases

Subject	GPT-2 Medium	95% CI	Curiosity-16	95% CI	**Δ**
mmlu_abstract_algebra	0.2	[0.1212,0.2788]	0.15	[0.0797,0.2203]	−0.05
mmlu_high_school_computer_science	0.27	[0.1825,0.3575]	0.23	[0.1471,0.3129]	−0.04
mmlu_world_religions	0.3158	[0.2459,0.3857]	0.2807	[0.2132,0.3482]	−0.035
mmlu_college_biology	0.2569	[0.1853,0.3286]	0.2222	[0.1541,0.2904]	−0.0347
mmlu_professional_psychology	0.2549	[0.2203,0.2895]	0.2304	[0.1970,0.2638]	−0.0245
mmlu_high_school_microeconomics	0.2101	[0.1582,0.2619]	0.1891	[0.1392,0.2389]	−0.0210
mmlu_electrical_engineering	0.2483	[0.1777,0.3188]	0.2276	[0.1591,0.2961]	−0.0207
mmlu_machine_learning	0.1964	[0.1225,0.2703]	0.1786	[0.1073,0.2498]	−0.0179
mmlu_conceptual_physics	0.2681	[0.2113,0.3248]	0.2511	[0.1955,0.3066]	−0.0170
mmlu_high_school_world_history	0.2658	[0.2095,0.3222]	0.2532	[0.1977,0.3086]	−0.0127

Table shows the top ten most dramatic decreases in subject-level accuracy in specific subjects between Curiosity-16 and GPT-2 Medium.

Interpretation: Across MMLU's 57 subjects, most per-subject confidence intervals generally overlap. This behavior is expected; individual MMLU subjects assign a limited number of questions, widening the CIs of both models. However, there are consistent directional gains in Humanities and STEM subjects, indicating that the targeted two-phase fine-tuning pipeline for Curiosity-16 yielded tangible subject-specific improvements in normalized accuracy. This pattern aligns with other legacy architectures evaluated on multitask benchmarks with fewer sample sizes per subject.

5. Results

These findings show modest but statistically stable gains in overall normalized accuracy, with clearest increases appearing in domain-specific MMLU subject-level tests. Results show regression across abstract reasoning and mathematics-heavy subjects, while showing small accuracy gains in STEM and Humanities subjects, suggesting an uneven generalization result stemming from the choice of datasets used. Zero-shot HellaSwag and MMLU evaluations were conducted twice in order to assess run-to-run variance, with both evaluations showing repeated, remarkably similar results and trends between GPT-2 Medium and Curiosity-16. Four MMLU subjects (High School US History, Public Relations, Computer Security, and High School Physics) showed no changes in normalized accuracy between GPT-2 Medium and Curiosity-16.

6. Limitations and Future Work

Our study was constrained by the use of a single consumer-grade machine, which limited batch sizes and constrained use of any larger models. While our two-phase SFT strategy shows promise, the performance decreases in certain STEM subjects indicate that our dataset curation could be further improved. Future work could involve more targeted dataset selection for abstract reasoning, more "focused" phases, and applying this SFT pipeline to more modern LLM architectures.

The inclusion of the Wikipedia-2025 dataset may also have led to limited indirect contamination of evaluation questions; we affirm that this is a necessary tradeoff for the knowledge injection that results in C16 possessing more recent world knowledge.

We did not apply RLHF methods during training and testing, providing a new opportunity to assess Curiosity-16 or future model performance gains and decreases.

Comparing the results of both Phase I and Phase II of C16 consistently yielded a larger gain in normalized accuracy towards the Phase II model, which represents the finalized Curiosity-16 model. Therefore, evaluation reporting excluded weaker Phase I-only results and focused on the most capable iteration of the C16 model (Phase II).

7. Conclusion

While Curiosity-16 remains far behind contemporary frontier large language models, C16 demonstrates that a strategic, two-phase SFT pipeline can yield directionally positive and measurable performance gains on a GPT-2-class model using only public resources and consumer-grade hardware. Results demonstrate that impactful LLM design and development is feasible without massive processing power; empowering independent researchers to fine-tune, scale, and deploy their own capable models. By effectively sequencing knowledge generalization and reasoning specialization into a focused two-phase SFT pipeline, we achieved clear improvements in MMLU and HellaSwag benchmarks, particularly for STEM and Humanities subjects. While results are modest and substantially lower than modern standards set by larger frontier models, these findings reveal the clear limitations and potential of fine-tuning legacy LLM architectures, and provide a structured foundation for fine-tuning modern, successor LLMs.

8. Reproducibility and Availability

Model: https://huggingface.co/ariankharazmi/Curiosity-16

GitHub: https://github.com/ariankharazmi/Curiosity-16-LLM

Dataset Suite: 11 HuggingFace Datasets (listed above)

Inference app: app.py (Curiosity-16 interactive demo available via HuggingFace Spaces)

LM-Eval-Harness HellaSwag and MMLU Commands Used:

```
lm_eval --model hf --model_args pretrained=ariankharazmi/Curiosity-16 \
  --tasks hellaswag --device mps --batch_size auto \
  --output_path results/C16_hellaswag_logs
```

```
lm_eval --model hf --model_args pretrained=openai-community/gpt2-medium \
  --tasks hellaswag --device mps --batch_size auto \
  --output_path results/gpt2m_hellaswag_logs
```

```
lm_eval --model hf --model_args pretrained=ariankharazmi/Curiosity-16 \
  --tasks mmlu --device mps --batch_size 2 \
  --output_path results/C16_mmlu_logs
```

```
lm_eval --model hf --model_args pretrained=openai-community/gpt2-medium \
  --tasks mmlu --device mps --batch_size 2 \
  --output_path results/gpt2m_mmlu_logs
```

All LM-Eval runs used default settings, bootstrap iters = 100,000, fixed seeds (random/numpy/torch/fewshot = 1234).

9. References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *arXiv preprint arXiv:1706.03762*.
2. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. OpenAI Technical Report. https://cdn.openai.com/better-languagemodels/language_models_are_unsupervised_multitask_learners.pdf
3. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.
4. Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., ... & Zou, A. (2021). A Framework for Few-Shot Language Model Evaluation. *Version v0.4.0*. Zenodo. https://doi.org/10.5281/zenodo.5371629
5. Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., & Choi, Y. (2019). HellaSwag: Can a Machine Really Finish Your Sentence? *arXiv preprint arXiv:1905.07830*.
6. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring Massive Multitask Language Understanding. *arXiv preprint arXiv:2009.03300*.
7. Köpf, A., Kilcher, Y., von Rütte, D., Anagnostidis, S., Tam, Z. R., Stevens, K., ... & Nagyfi, R. (2023). OpenAssistant Conversations – Democratizing Large Language Model Alignment. *arXiv preprint arXiv:2304.07327*.
8. Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., ... & Launay, J. (2024). The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. *arXiv preprint arXiv:2406.17557*.
9. Wikimedia Foundation. (2025). Wikimedia Downloads. Retrieved from https://dumps.wikimedia.org/
10. Rajpurkar, P., Jia, R., & Liang, P. (2018). Know What You Don't Know: Unanswerable Questions for SQuAD. *arXiv preprint arXiv:1806.03822*.
11. Fan, A., Jernite, Y., Perez, E., Grangier, D., Weston, J., & Auli, M. (2019). ELI5: Long Form Question Answering. *arXiv preprint arXiv:1907.09190*.
12. Conover, M., Hayes, M., Mathur, A., Meng, X., Xie, J., Wan, J., ... & Zaharia, M. (2023). Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM. *Databricks Blog*.
13. Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., & Schulman, J. (2021). Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*.
14. Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., & Hashimoto, T. B. (2023). Stanford Alpaca: An Instruction-following LLaMA model. *GitHub repository*.
15. Zhong, W., Cui, R., Guo, Y., Liang, Y., Lu, S., Wang, Y., ... & Duan, N. (2023). AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. *arXiv preprint arXiv:2304.06364*.

Appendix

Full Confidence Intervals available on Curiosity-16 GitHub and HuggingFace repositories.

HellaSwag

Metric	GPT-2 Medium	Curiosity-16	Δ
Raw Accuracy	0.3331	0.3385	0.0054
Normalized Accuracy	0.3938	0.3967	0.0029

Massive Multitask Language Understanding

Subject	GPT-2 Medium	Curiosity-16	Δ
mmlu	0.2290	0.2375	+0.0085
mmlu_humanities	0.2427	0.2557	+0.0130
mmlu_formal_logic	0.2937	0.3175	+0.0238
mmlu_high_school_european_history	0.2182	0.2667	+0.0485
mmlu_high_school_us_history	0.2451	0.2451	0
mmlu_high_school_world_history	0.2658	0.2532	−0.0127
mmlu_international_law	0.2397	0.2314	−0.0083
mmlu_jurisprudence	0.25	0.2685	+0.0185
mmlu_logical_fallacies	0.2209	0.2577	+0.0368
mmlu_moral_disputes	0.2486	0.2399	−0.0087
mmlu_moral_scenarios	0.2436	0.2726	+0.0291
mmlu_philosophy	0.1865	0.1929	+0.0064
mmlu_prehistory	0.2222	0.2438	+0.0216
mmlu_professional_law	0.2451	0.2581	+0.0130
mmlu_world_religions	0.3158	0.2807	−0.0351

mmlu_other	0.2350	0.2481	+0.0132
mmlu_business_ethics	0.3	0.31	+0.0100
mmlu_clinical_knowledge	0.2038	0.2302	+0.0264
mmlu_college_medicine	0.2139	0.2428	+0.0289
mmlu_global_facts	0.19	0.18	−0.01
mmlu_human_aging	0.3184	0.3094	−0.0090
mmlu_management	0.1748	0.2039	+0.0291
mmlu_marketing	0.2949	0.2906	−0.0043
mmlu_medical_genetics	0.3	0.35	+0.05
mmlu_miscellaneous	0.2337	0.2286	−0.0051
mmlu_nutrition	0.1863	0.2157	+0.0294
mmlu_professional_accounting	0.2305	0.2482	+0.0177
mmlu_professional_medicine	0.1838	0.2426	+0.0588
mmlu_virology	0.2831	0.2711	−0.0120
mmlu_social_sciences	0.2184	0.2197	+0.0013
mmlu_econometrics	0.2368	0.2281	−0.0088
mmlu_high_school_geography	0.1818	0.2172	+0.0354
mmlu_high_school_government_and_politics	0.1969	0.2124	+0.0155
mmlu_high_school_macroeconomics	0.2051	0.2077	+0.0026
mmlu_high_school_microeconomics	0.2101	0.1891	−0.0210
mmlu_high_school_psychology	0.1945	0.2147	+0.0202
mmlu_human_sexuality	0.2519	0.2595	+0.0076
mmlu_professional_psychology	0.2549	0.2304	−0.0245

mmlu_public_relations	0.2182	0.2182	0
mmlu_security_studies	0.1878	0.2	+0.0122
mmlu_sociology	0.2438	0.2338	+0.0122
mmlu_us_foreign_policy	0.27	0.28	+0.0100
mmlu_stem	0.2128	0.2173	+0.0044
mmlu_abstract_algebra	0.2	0.15	−0.05
mmlu_anatomy	0.1926	0.2370	+0.0444
mmlu_astronomy	0.1908	0.2105	+0.0197
mmlu_college_biology	0.2569	0.2222	−0.0347
mmlu_college_chemistry	0.16	0.15	−0.01
mmlu_college_computer_science	0.24	0.28	+0.0400
mmlu_college_mathematics	0.23	0.22	−0.01
mmlu_college_physics	0.1765	0.2451	+0.0687
mmlu_computer_security	0.28	0.28	0
mmlu_conceptual_physics	0.2681	0.2511	−0.0170
mmlu_electrical_engineering	0.2483	0.2276	−0.0207
mmlu_elementary_mathematics	0.2169	0.2222	+0.0053
mmlu_high_school_biology	0.1774	0.1839	+0.0065
mmlu_high_school_chemistry	0.2118	0.2365	+0.0246
mmlu_high_school_computer_science	0.27	0.23	−0.04
mmlu_high_school_mathematics	0.2148	0.2259	+0.0111
mmlu_high_school_physics	0.1987	0.1987	0
mmlu_high_school_statistics	0.1574	0.1898	+0.0324

mmlu_machine_learning	0.1964	0.1786	−0.0179