Improving Multilingual Capabilities with Cultural and Local Knowledge in Large Language Models While Enhancing Native Performance

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have shown remarkable capabilities, but their development has primarily focused on English and other high-resource languages, leaving many languages underserved. We present our latest Hindi-English bi-lingual LLM with 3% average improvement in benchmark scores over both languages, outperforming models twice its size. Using a curated dataset composed of English and Hindi instruction data of 485K samples, we instruction tuned models such as Qwen-2.5-14B-Instruct and Phi-4 to improve performance over both English and Hindi. Our experiments encompassing seven different LLMs of varying parameter sizes and over 140 training attempts with varying English-Hindi training data ratios demonstrated that it is possible to significantly improve multilingual performance without compromising native performance. Further, our approach avoids resource-intensive techniques like vocabulary expansion or architectural modifications, thus keeping the model size small. Our results indicate that modest fine-tuning with culturally and locally informed data can bridge performance gaps without incurring significant computational overhead. We release our training code, datasets, and models under mit and apache licenses to aid further research towards under-represented and low-resource languages.

1 Introduction

The rapid advancement of Large Language Models (LLMs) has led to great advances in various natural language processing tasks. However, the majority of research efforts have disproportionately focused on English and a select few high-resource languages. This disparity leaves a vast number of languages under-served, limiting the global accessibility and applicability of LLM technology. While the lack of readily available data for many languages is a contributing factor, it is not the sole reason. Economic factors and limited access to computational resources also play significant roles in accessibility to target audience. In this work, we address the gap by developing a bilingual LLM that performs well on English and Hindi tasks. We focused on maintaining relatively smaller model sizes and rather than resorting to resource-intensive methods such as vocabulary expansion, block expansion, or additional layers, we employ computationally efficient fine-tuning methods such as Supervised Fine-Tuning (SFT) (Face, 2025)(von Werra et al., 2020) with Low-Rank Adaptation (LoRA) (Hu et al., 2021) through Unsloth (Daniel Han and team, 2023). Our primary goal was to boost performance over Hindi tasks while retaining similar performance over English.

We demonstrate our method by fine-tuning Qwen-2.5-14B-Instruct (Qwen et al., 2025) and Phi-4 (Abdin et al., 2024) models on a mixed-language dataset. Moreover, our experiments extend to five other LLMs : Gemma 2 9B, Gemma 2 2B (Team, 2024a), Llama 3.1 8B, Llama 3.1 3B (Team, 2024b), Qwen 2.5 3B where over 140 fine-tuning attempts were conducted by varying the distribution ratios of Hindi and English samples of each domain in the training data. These experiments provide insights into how performance changes with varying dataset distributions over each domain. This can help in dataset curation to effectively balance bilingual performance. The promising results suggest that enhancing low-resource language capabilities doesn't necessarily require large-scale architectural changes but can be achieved through targeted, efficient fine-tuning of models with basic capabilities over a language.

2 Related Works

Prior studies have attempted to address this disparity through various techniques, including vocabulary expansion/modification (Tejaswi et al., 2024) (Csaki et al., 2023) (Shi et al., 2024) (Balachandran, 2023), modifications in architecture like block expansion and the addition of extra layers to accommodate linguistic diversity (Llama-Nanda, 2024), or continued pre-training followed by instruction tuning again (Mahdizadeh Sani et al., 2025) (Kuulmets et al., 2024) (Cui et al., 2023) (Vo et al., 2024) (Luukkonen et al., 2023) (Kallappa et al., 2025) (Toraman, 2024). However, such methods often incur substantial computational costs and lead to increase in model sizes. Prior works also include multi-lingual LLMs optimized for several languages including Hindi : Bloom-176B (BigScienceWorkshop, 2023), Aya-23B (Aryabumi et al., 2024), Aya-101 (Üstün et al., 2024) and Aya-expanse (Dang et al., 2024). Additionally we also have several other mono-lingual and bi-lingual LLMs focused on Hindi : llama-nanda-10B (Llama-Nanda, 2024), Airavata-7B (Gala et al., 2024), (BhabhaAI, 2024), Aryabhatta-8.5B (GenVRadmin, 2024), Sarvam-2B (Sarvamai, 2024), Krutrim-2-12B (Kallappa et al., 2025) and Nemotron-mini-Hindi (Joshi et al., 2024). The key differences can be seen in Table 1.

3 Datasets

Despite the existence of datasets to cover several domains for Hindi (Khan et al., 2024), (Ramesh et al., 2022), we decided to experiment primarily with translated / reformatted datasets which do not prohibit usage for research/commercial purpose. This was done so that the same work can be implemented/extended to low-resource languages. Also, fine-tuning on translated data is an efficient way to adapt mPLMs to new languages, leveraging their pre-trained multilingual knowledge. (Chen and Chen, 2024). For translation, we used GPT-4o-mini (OpenAI, 2024) through Microsoft Azure [1] to translate few datasets and benchmarks from English to Hindi : Big-Bench-Hard (Suzgun et al., 2022), XNLI (Conneau et al., 2018), Xl-Sum (Hasan et al., 2021). Some of the benchmarks which already have Hindi subsets were used directly : Global MMLU (Singh et al., 2024a), IndicXNLI (Aggarwal et al., 2022). Some of the publicly available datasets containing cultural and localized general knowledge like Indian legal FAQ (Aditya2411, 2024), UPSC FAQ (prnv19, 2024), IndianTAX FAQ (msinankhan1, 2024), IndianMedicines , IndiaCuisines and IndiaTravel Guide (cyberblip, 2024) were used to generate instruction-response pairs from the tabular format data using GPT-4o-mini as a part of our dataset collection. These were first translated to the other language from the original language then manually verified by multiple annotators to ensure quality in both languages. We also used a few subsets from the Aya collection (Singh et al., 2024b) i.e the translation, simplification and summarization subsets. In total the collected dataset had 3.12M samples with nearly 50:50 ratio of English and Hindi data. Around 90K samples from these cover localized and cultural knowledge. Among the rest, some domains and tasks had higher proportion in the collection. We used randomly selected subsets from those datasets while maintaining equal language ratios. After filtering the training data, we had around 485K samples of which 20% are of localized domain and cultural knowledge, while the rest are of generic tasks like math, MCQs, reasoning, summarization, rephrasing and translation.

[1] https://azure.microsoft.com/en-us/products/ai-services/openai-service/

4 Instruction Data Formatting

During Training we have appended the inputs with different strings based on the task at hand. The details of the appended strings for each task type can be seen in Table 2. The underlined portions were replaced with the corresponding texts for each sample. This modification helped in tuning the model to obey instructions well with less additional tokens needed for formatting instructions, while not compromising the performance on both the languages. The inputs were preprocessed to replace consecutive spaces with a single space, removal of leading and trailing spaces and replacement of double quotes with single quotes. Same chat templates were used as the original models with input portions processed into our format.

5 Initial Evaluation

Before proceeding to train over the full dataset, we have first experimented through several attempts by training on a subset of our data with/without including training data of benchmarks' domains and by varying ratio of each language in the dataset used. The subsets contain at most 2000 samples from

Prior Work	Model	VRAM Req	Approach	Status
(GenVRadmin, 2024)	Aryabhatta-8.5B	18 GB	Vocabulary Expansion	Open-Weights
(Gala et al., 2024)	Airavata-7B	14 GB	LoRA fine-tuning	Open-Weights Open-Dataset
(Sarvamai, 2024)	Sarvam-1-2B	6 GB	From Scratch	Open-Weights
(Kallappa et al., 2025)	Krutrim-2-12B-Instruct	50 GB	Vocabulary Expansion	Open-Weights
(Joshi et al., 2024)	Nemotron-mini-Hindi	9 GB	Continued-Pretraining	Open-Weights
(Llama-Nanda, 2024)	Llama-nanda-10B-chat	41 GB	Block Expansion Continued-Pretraining	Open-Weights
(Dang et al., 2024)	Aya-Expanse-8B	17 GB	From Scratch	Open-Weights
(Dang et al., 2024)	Aya-Expanse-35B	66 GB	From Scratch	Open-Weights
Our Work	placeholder-name	30 GB	LoRA with modified chat template	Open-Dataset Open-Weights

Table 1: Key differences between other works regarding Hindi LLMs

Task	Input Format
Natural Language Inference	"*Text1* ### *Text2* ### **NLI** ### :"
Multiple Choice Questions	"*Question* ### A) *a*, B) *b*,... ### **MCQ** ### :"
Numeric Questions	"*Question* ### **NUMERIC** ### :"
Boolean Questions	"*Question* ### **BOOLEAN** ### :"
Questions seeking Long responses	"*Question* ### **LONG RESPONSE** ### :"
Short responses (few words)	"*Input* ### **DIRECT RESPONSE** ### :"
Coding	"*Input* ### **CODE** ### :"
Text Summarization	"*Input* ### **SUMMARIZE** ### :"
Paraphrasing/Rephrasing	"*Input* ### **PARAPHRASE** ### :"
Translation to specified language	"*Input* ### **TRANSLATION** [*lang*] ### :"
Text Simplification/ELI5	"*Input* ### **SIMPLIFY** ### :"

Table 2: Formats of Input Texts used in training

each dataset source for both languages combined. We used normalized next-token log probabilities for MCQs and Boolean benchmarks during the initial evaluation stage to evaluate the models. We then compared how the scores changed with these variations and compared with the original models to gather insights into optimal final dataset sampling approaches. The results over Qwen-2.5-14B and Phi-4 can be seen below in Table 3 and Table 4 respectively. The results for the rest of the models can be found in Appendix C.

6 Dataset Distribution and Ordering

The performance of models from initial tests didn't vary significantly with/without being trained on math data. The performance on Math subsets of MMLU as well remained similar on both languages with/without being trained on math samples. Since we would be training on a large number of samples, we decided to still use a considerable amount of math samples. A significant performance gap was observed over boolean benchmarks with a nearly 3% increase in English and 5% increase in Hindi. Hence, we decided to use a slightly higher amount of boolean questions' samples in the final dataset. The language ratios for each domain in the final dataset were determined based on the initial training data ratios that gave the best results. The samples of the final dataset were sorted over input lengths in ascending order with a certain number of longest samples placed in the beginning, this approach could improve batch processing efficiency and training stability (Wang et al., 2024a). This number was set equal to the total effective batch size (i.e the product of batch size and gradient accumulation steps). The samples related to local and cultural knowledge were then placed such that they are evenly spread out in the dataset except the initial batch. More info on the dataset can be found in Appendix B. The training methods and details can be found in Appendix A.

7 End Evaluation

Apart from the benchmarks seen in Table 3 and Table 4, we perform evaluations over additional benchmarks like : MMLU-Pro (Wang et al., 2024b), BigBench-Hard (Suzgun et al., 2022), MuSR (Sprague et al., 2024), GPQA (Rein et al., 2023), MATH-Hard (Hendrycks et al., 2021). We used open-llm-leaderboard [2] (Fourrier et al., 2024) for evaluation over some of the benchmarks

[2]https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard

Benchmarks	Ratio of	ARC-Challenge		ARC-Easy		MMLU		BoolQ		Context-MCQ		Overall Average		
Domain data used?	Hindi	En	Hi	En	Hi	En	Hi	En	Hi	En	Hi	En	Hi	Tot
No	10%	90.61	73.21	94.82	80.05	75.74	53.60	84.16	77.24	91.4	79.7	87.34	72.76	80.05
No	20%	90.53	73.04	94.99	80.68	75.84	53.95	83.30	75.80	90.9	79.0	87.11	72.49	79.80
No	30%	90.78	73.55	95.16	80.89	75.67	54.00	81.22	74.03	91.2	78.5	86.80	72.19	79.50
No	40%	91.13	73.29	94.95	80.64	76.09	53.85	84.25	72.29	91.1	78.1	87.50	71.63	79.57
No	50%	91.30	73.38	94.99	81.19	75.63	54.21	81.53	73.63	91.0	79.0	86.89	72.28	79.59
No	60%	91.55	75.17	95.75	81.73	75.20	54.29	85.78	75.83	91.7	79.7	88.00	73.35	80.67
No	70%	91.38	74.91	95.71	82.28	75.52	54.32	85.08	80.82	90.7	79.7	87.68	74.41	81.04
No	80%	91.13	74.66	94.99	82.37	75.87	54.53	84.19	78.07	91.4	78.8	87.51	73.68	80.60
No	90%	91.47	75.09	95.50	82.83	75.59	54.69	84.19	79.44	91.2	79.5	87.59	74.30	80.95
No	100%	91.64	74.83	95.50	82.87	75.69	54.47	85.05	79.72	91.6	80.3	87.90	74.44	81.17
Yes	10%	90.96	72.70	94.74	80.26	75.90	53.78	88.47	81.12	90.4	77.3	88.09	73.03	80.56
Yes	20%	90.87	73.29	94.82	81.10	75.89	53.77	88.69	84.27	91.1	78.1	88.27	74.11	81.19
Yes	30%	91.04	73.63	94.91	81.40	75.74	54.24	88.07	81.95	90.8	78.6	88.11	73.96	81.04
Yes	40%	90.78	74.91	94.78	81.65	76.22	54.71	88.78	83.85	90.9	78.8	88.29	74.78	81.53
Yes	50%	91.04	74.74	94.78	81.86	76.34	54.80	88.69	84.61	91.1	78.5	88.39	74.90	81.64
Yes	60%	91.04	75.00	94.87	81.86	75.96	54.76	88.62	84.58	90.9	79.0	88.27	75.04	81.65
Yes	70%	90.87	74.15	94.53	82.11	75.46	54.91	87.86	84.06	91.2	79.7	87.98	74.98	81.48
Yes	80%	90.96	76.62	94.87	82.37	76.04	54.19	88.69	84.89	90.9	78.4	88.29	75.29	81.79
Yes	90%	91.47	75.60	94.74	82.53	75.84	54.77	87.79	84.89	90.8	79.7	88.15	75.50	81.82
Yes	100%	91.21	75.94	94.61	82.70	75.79	55.00	88.29	84.55	91.6	79.7	88.30	75.58	81.94
Original		90.87	69.62	95.45	78.49	74.37	52.16	86.09	78.89	91.2	77.4	87.60	71.31	79.46

Table 3: Results (.2f) from each training attempt with 8% of our training data over Qwen 2.5 14B

Benchmarks	Ratio of	ARC-Challenge		ARC-Easy		MMLU		BoolQ		Context-MCQ		Overall Average		
Domain data used?	Hindi	En	Hi	En	Hi	En	Hi	En	Hi	En	Hi	En	Hi	Tot
No	10%	92.24	74.74	97.35	83.67	76.04	50.45	87.52	83.88	86.7	74.7	87.97	73.48	80.72
No	20%	92.06	75.77	97.39	84.18	76.01	51.61	87.13	83.33	87.0	75.0	87.91	73.97	80.94
No	30%	92.24	76.54	97.26	84.26	76.02	51.40	87.43	84.22	86.7	75.6	87.93	74.40	81.16
No	40%	92.15	77.30	97.35	84.97	76.08	51.76	87.16	83.79	87.2	76.1	87.98	74.78	81.38
No	50%	92.24	82.59	97.43	89.39	76.34	57.41	87.61	85.10	86.6	77.7	88.04	78.43	83.24
No	60%	92.24	77.39	97.26	84.76	75.82	51.72	87.46	83.91	86.8	75.5	87.91	74.65	81.28
No	70%	91.98	77.65	97.18	84.89	75.68	51.87	87.49	83.88	86.8	75.8	87.82	74.81	81.32
No	80%	91.21	77.30	97.31	84.64	75.75	51.59	87.31	84.34	86.2	76	87.55	74.77	81.16
No	90%	92.32	77.30	97.35	84.51	75.68	50.96	87.58	84.37	86.6	76.1	87.90	74.64	81.27
No	100%	92.41	78.16	97.39	85.35	75.87	52.12	87.58	83.88	86.1	76.4	87.87	75.18	81.52
Yes	10%	92.15	76.96	97.85	85.31	75.66	50.54	88.53	85.31	86.3	75.0	88.10	74.63	81.36
Yes	20%	92.49	77.05	97.56	85.69	75.49	50.06	88.87	85.29	86.4	74.5	88.16	74.52	81.34
Yes	30%	92.49	78.41	97.69	86.95	75.85	51.28	88.35	85.44	86.5	75.4	88.18	75.50	81.84
Yes	40%	92.66	82.25	97.77	90.36	75.86	56.32	88.65	85.92	86.7	78.3	88.33	82.25	83.48
Yes	50%	93.17	82.93	97.85	91.07	76.52	57.87	88.31	85.22	87.1	78.7	88.59	79.16	83.88
Yes	60%	92.49	78.83	97.51	87.07	75.91	52.04	88.07	84.21	86.6	75.9	88.11	75.61	81.86
Yes	70%	92.40	79.18	97.64	86.70	75.94	51.84	88.31	83.97	86.1	75.8	88.08	75.49	81.79
Yes	80%	92.66	79.35	97.56	87.75	76.04	52.05	88.13	84.34	85.9	76.6	88.06	76.02	82.04
Yes	90%	92.58	79.69	97.60	87.96	76.06	52.49	88.23	84.25	86.3	76.4	88.15	76.16	82.16
Yes	100%	92.49	80.12	97.69	87.58	75.95	52.55	88.32	84.52	86.0	76.2	88.09	76.19	82.14
Original		92.41	79.18	97.31	86.87	74.67	53.24	86.30	82.72	86.3	75.7	87.40	75.54	81.47

Table 4: Results (.2f) from each training attempt with 8% of our training data over Phi 4 14B

through eval-harness framework(Gao et al., 2021). Table 8 demonstrates The performance of our models in comparison with the original models over several benchmarks. We did observe variations in the scores from open-llm-leaderboard and the corresponding benchmark scores which were self reported for the original models. We used the scores from the leaderboard for all models over those benchmarks for reproducibility a fair comparison. The evaluation methods used can be seen in Table 5.

Benchmark	Eval Criteria	Eval Framework
ARC-C	0-Shot	log probabilities
ARC-E	0-Shot	log probabilities
BoolQ	0-Shot	log probabilities
CMCQ	0-Shot	log probabilities
MMLU	0-Shot	log probabilities
MMLU-Pro	5-Shot	eval-harness
BBH	3-Shot	eval-harness
GPQA	0-Shot	eval-harness
MATH Hard	4-Shot	eval-harness
MuSR	0-Shot	eval-harness

Table 5: Benchmarks used for evaluation and their details

8 Generative tasks evaluation

Scarcity of genuine and authentic multilingual benchmarks of broad range of topics has been a concern for many languages. Prior works in comparison like (Llama-Nanda, 2024) have not included generative evaluations over either languages. while (Joshi et al., 2024) utilized limited generative benchmarks using LLM-as-a-judge to score the responses, with only the MT-Bench a translation task undergoing human evaluation. Further, training on translated data to test over benchmarks translated from English defeats the purpose of building multilingual and multi-cultural LLMs. (Aryabumi et al., 2024) too utilizes translated benchmarks for multilingual generative task evaluation, with additional human evaluation without topic/domain restriction. We have performed human-evaluation in the same way over both languages. These results can be seen in Figure 1. We performed human evaluations though third party annotators over both languages over few of the models that achieved comparably good performance over non-english discriminative tasks. A total of 3217 comparisons were done primarily in Hindi (2097) and the rest over English (1120). For a fair comparison we utilized the default hyper-parameters of each of the models.

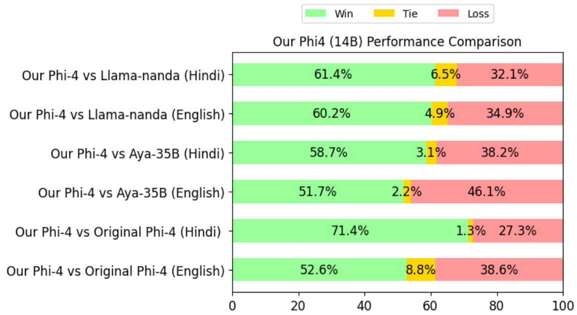

Figure 1: Win Rates with comparable models through human evaluation.

9 Comparisons

For additional comparisons, we compare the performance of our models with other Hindi bilingual LLMs and other open-source LLMs which are optimized for Hindi. Due to the large variations in number of parameters of our models and other comparable models, we compare average benchmark performance versus the model size in terms of VRAM requirement. The comparisons over English and Hindi benchmarks along side our Qwen and phi models can be seen in Table 6 and Table 7. Over the benchmarks of higher difficulty, our models have consistently outperformed models over twice their size as seen in Table 6.

9.1 Domain wise Performance change

The performance of our models compared to the original versions over MMLU-pro can be seen in Table 10. The type of questions the models faced through MMLU-Pro maybe of the same domain but were of different subdomains and task types compared to those in our datasets. For example, The CS benchmarks' questions were MCQs about various areas of computer science while our training data over CS was solely from MBPP (Austin et al., 2021) which consists of a text input and a python code as an output. Further the only source of training data we used for economics consist of TAX filing FAQs over Indian context and primarily in Hindi. Hence such domains' data usage was mentioned as N/A. The domains which had a performance boost in our models without being in training data had questions of the form of fill-mask or text completion which were similar to the training data from Winogrande-XL (Sakaguchi et al.,

Model ↓	ARC-C	ARC-E	BoolQ	CMCQ	MMLU	Average*	MMLU-Pro	GPQA	MuSR	BBH	MATH
AryaBhatta-GemmaUltra-8.5B	22.70	25.04	62.23	22.95	23.70	31.32	22.66	25.34	42.72	41.12	2.95
Airavata-7B	25.09	30.47	62.17	25.31	33.20	35.25	16.35	27.43	37.57	36.00	13.60
sarvam-1-2B	30.03	33.25	62.17	42.80	27.90	39.23	-	-	-	-	-
Nemotron-4-Mini-Hindi-Instruct	55.80	71.63	62.11	68.10	43.20	60.17	25.95	30.87	41.53	40.11	2.04
Llama-3-Nanda-10B-Chat	65.36	80.64	82.29	67.60	50.61	69.30	31.57	30.11	43.52	49.38	5.59
Krutrim-2-12b-instruct	67.32	81.10	84.74	76.30	56.10	73.11	-	-	-	-	-
aya-expanse-8b	74.06	87.08	86.45	83.30	56.89	77.56	30.04	30.29	37.17	49.42	7.02
aya-expanse-32B	85.41	**95.08**	**90.43**	89.80	69.71	86.08	41.30	32.55	38.62	56.29	13.37
Our Qwen Model (14B)	**90.61**	94.82	88.53	**90.70**	75.00	87.93	**52.63**	36.24	44.84	64.97	**25.08**
Our Phi Model (14B)	**97.39**	92.24	87.65	87.40	**75.59**	**88.05**	52.39	**39.77**	**49.07**	**66.97**	23.11

Table 6: Metrics (.2f) of our and other LLMs over several **English** benchmarks

*Averages for English were calculated using just the first 5 benchmarks for similar comparison with Hindi

The best and second best for each benchmark are highlighted as bold+underlined and underlined respectively

Model ↓	ARC-C	ARC-E	BoolQ	CMCQ	MMLU	Average
AryaBhatta-GemmaUltra-8.5B	22.70	25.08	62.17	22.95	23.80	31.34
Airavata-7B	22.87	25.13	62.17	23.28	33.20	33.33
sarvam-1-2B	32.76	35.06	62.16	47.10	24.22	40.26
Llama-3-Nanda-10B-Chat	45.99	60.56	71.96	54.70	36.35	53.91
Nemotron-4-Mini-Hindi-4B-Instruct	50.68	63.72	68.74	51.30	37.18	54.32
Krutrim-2-12b-instruct	56.83	70.66	78.86	64.10	46.51	63.39
aya-expanse-8b	57.42	72.90	80.42	69.00	43.39	64.63
aya-expanse-32B	73.29	85.48	**87.73**	**79.70**	**56.96**	76.63
Our Qwen Model (14B)	74.06	81.23	84.07	78.20	53.85	74.82
Our Phi Model (14B)	**81.74**	**89.06**	86.02	78.70	56.39	**78.38**

Table 7: Metrics (.2f) of our and other LLMs over several **Hindi** benchmarks

The best and second best for each benchmark are highlighted as bold+underlined and underlined respectively

Benchmark	Lang	Qwen-2.5-14B-Instruct	Our Qwen	Change	Phi-4	Our Phi-4	Change
ARC-Easy	En	95.45	94.82	▼ 0.63	97.31	97.39	▲ 0.08
	Hi	78.49	81.23	▲ 2.74	86.87	89.06	▲ 2.19
ARC-Challenge	En	90.87	90.61	▼ 0.26	92.41	92.24	▼ 0.17
	Hi	69.62	74.06	▲ 4.44	79.18	81.74	▲ 2.56
BoolQ	En	86.09	88.53	▲ 2.44	86.30	87.65	▲ 1.35
	Hi	78.89	84.07	▲ 5.18	82.72	86.02	▲ 3.30
Context-MCQ	En	91.20	90.70	▼ 0.50	86.30	87.40	▲ 1.10
	Hi	77.40	78.20	▲ 0.80	75.70	78.70	▲ 3.00
MMLU	En	74.37	75.00	▲ 0.63	74.67	75.59	▲ 0.92
	Hi	52.16	53.85	▲ 1.69	53.24	56.39	▲ 3.15
Average	**En**	**87.60**	**87.93**	**▲ 0.33**	**87.40**	**88.05**	**▲ 0.65**
	Hi	**71.31**	**74.82**	**▲ 3.51**	**75.54**	**78.38**	**▲ 2.84**
	Overall	**79.46**	**81.38**	**▲ 1.92**	**81.47**	**83.22**	**▲ 1.75**

Table 8: Performance of our models compared to originals over each benchmark : evals through log likelihoods

Benchmark	Lang	Qwen-2.5-14B-Instruct	Our Qwen	Change	Phi-4	Our Phi-4	Change
MMLU-Pro	En	49.04	52.63	▲ 3.59	53.78	52.39	▼ 1.39
MATH hard	En	00.00	25.08	▲ N/A	12.31	23.11	▲ 10.80
GPQA	En	32.21	36.24	▲ 4.03	33.72	39.77	▲ 6.05
MuSR	En	40.87	44.84	▲ 3.97	41.01	49.07	▲ 8.06
BigBench-Hard	En	63.74	64.97	▲ 1.23	68.60	66.97	▼ 1.63
Average		**37.17**	**44.75**	**▲ 7.58**	**41.88**	**46.26**	**▲ 4.38**

Table 9: Performance of our models compared to originals over each benchmark : evals through eval-harness

2021) and PIQA (Bisk et al., 2020) spanning several domains.

9.2 Model biases over choices

The observations from domain wise performance changes by Phi and Qwen were significantly different. The domains which were well represented in our training data had a significant boost on both languages of MMLU. Despite training on MCQs which consist of 2-4 options, similar results of improvement were seen over MMLU-Pro which has upto 10 options. On the other hand, Phi-4 had a higher performance boost over MMLU which has the same number of options as the samples in the training data, but the performance over MMLU-Pro dropped irrespective of domain. The distribution of choices made by each of our LLMs and the corresponding original implementation can be seen in Figure 2. The instruction tuning dataset we used had an equal distribution of each of the choices among MCQ samples. The original Qwen model overwhelmingly chose from the final two options while our model was able to generalize well despite not being trained on MCQs with 10 choices. On the other hand the original phi-4 was able to perform better than its counterpart, but despite being fine-tuned with equal distribution of choices, the model displayed an inclination towards the first choice among the list of options. The extent of this bias varied between each domain significantly. More on this can be seen in Appendix D. As our phi model was fine-tuned from the original models' instruct variants, the biases were assumed to have been carried forward. Our models were able to respond well with less biases in choices over the domains whose samples are present in large quantities in our training data. To further look into this, we tried to fine-tune the base variant of qwen-2.5-14B rather than the instruct model to see the choice distribution over MMLU-Pro, while most of our dataset's samples of MCQs were having 4-5 samples, it was reflected in the choices made as seen in Figure 3 which demonstrates the issue within the original model similar to previous works demonstrating sensitivity on models' sensitivity to order of choices (Pezeshkpour and Hruschka, 2024). But a well balanced instruction tuning dataset can minimize this issue or an evaluation independent of order of choices (Zheng et al., 2023). A slight tilt from left to right in Figure 2 and Figure 3 can be expected as not all questions are accompanied by 10 options with a considerable amount having less.

10 Conclusion

We demonstrate that enhancing low-resource language capabilities in LLMs is possible through targeted fine-tuning rather than complex architectural changes. Our work shows that a 12-15B parameter LLM provides an effective balance between performance and accessibility, requiring just 30GB RAM. The performance analysis reveals that our Phi-4 model excels in general-purpose tasks, while the Qwen model shows stronger adaptation to specific domains, as evidenced by the domain-wise performance changes in Table 10. Our approach of using primarily translated datasets, except for culturally specific knowledge, makes this method readily adaptable to other low-resource languages. To further push the research in low-resource languages, we release our training code, datasets, and models under commercially permissible licenses.

10.1 Scalability to other languages

As not every language has readily available datasets of even a few domains, we took an approach of using just translated datasets for all domains other than those used for localized and cultural knowledge addition. This would enable reusing the approach to build bi-lingual LLMs optimized for

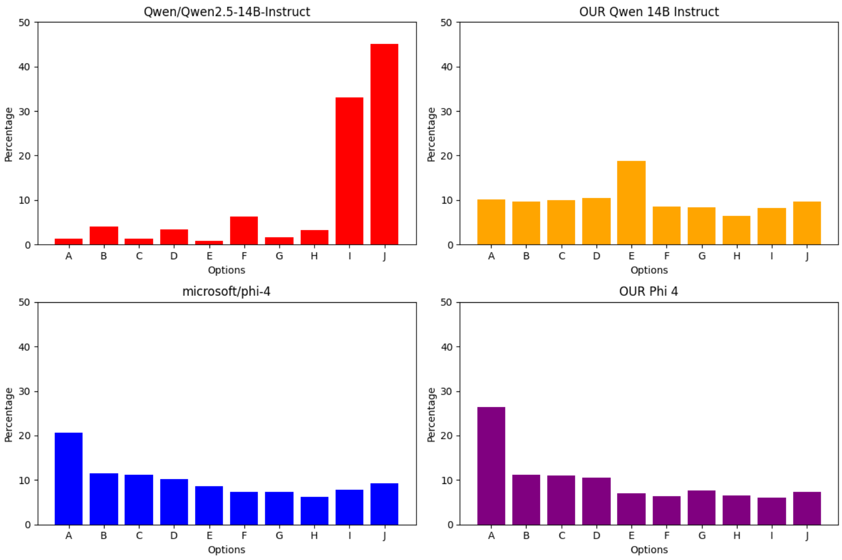

Figure 2: Distribution of each model's choices over MMLU-Pro

Model →	Qwen-2.5-14B		Change	Phi-4		Change	Training
Domain ↓	Original	Ours		Original	Ours		Data Used
Health	60.39	65.65	▲ 5.26	65.40	65.40	▲ 0.00	Yes
Biology	76.15	79.36	▲ 3.21	80.89	81.03	▲ 0.14	Yes
Engineering	38.08	46.85	▲ 8.77	47.06	44.17	▼ 2.89	Yes
Math	39.53	44.78	▲ 5.25	41.01	38.79	▼ 2.22	Yes
Physics	39.80	41.96	▲ 2.16	42.80	39.11	▼ 3.69	Yes
Chemistry	35.78	38.25	▲ 2.47	36.75	35.69	▼ 1.06	Yes
Law	37.78	41.42	▲ 3.64	48.14	47.14	▼ 1.00	Yes
Philosophy	53.51	57.92	▲ 4.41	62.32	59.72	▼ 2.60	N/A
Psychology	70.05	73.81	▲ 3.76	76.32	76.82	▲ 0.50	N/A
Business	37.90	45.63	▲ 7.73	40.94	38.91	▼ 2.03	N/A
CS	50.73	53.17	▲ 2.44	60.00	58.78	▼ 1.22	N/A
Economics	66.71	66.47	▼ 0.24	68.84	69.08	▲ 0.26	No
History	58.01	57.74	▼ 0.27	63.78	62.73	▼ 1.05	No
Other	54.44	53.68	▼ 0.76	57.47	56.71	▼ 0.76	No

Table 10: Domain wise performance changes over MMLU-Pro (English) with our models

other languages as long as a proficient LLM supports the language to translate the texts fluently.

10.2 Model Efficiency

Unsloth's version of phi-4 (Unsloth AI, 2023) with llama architecture led to an improved performance but increased emissions. Our model resulted in lesser emissions during evaluation over the open-llm-leaderboard, while improving the model's performance. A comparison of our model to the original and unsloth's phi-4 can be seen in Figure 4.

11 License

Our Qwen and Phi models are available through the same licenses as the models we used as a base i.e apache-2.0 and mit respectively. the models can be accessed here [3]. The training datasets are publicly available here[4]. Most datasets used for training the models have a copyleft license, with the rest having no license specified and are publicly available on huggingface.

[3]Our Phi-4 model :https://huggingface.co/

[4]Datasets : https://huggingface.co/

Limitations

Our models, although demonstrating robust performance across multiple benchmarks, may produce inaccurate, incomplete, or irrelevant outputs due to knowledge cutoffs in its training data. The models although working well directly with the original chat template are better optimized for our prompt formats. The approach presented has been tested in several attempts with Hindi, we believe a similar boost can be obtained over other languages as well, but has not been tested yet.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. 2024. Phi-4 technical report.

Aditya2411. 2024. Law india dataset. `https://huggingface.co/datasets/Aditya2411/law_india`. Accessed: 2024-10-29.

Divyanshu Aggarwal, Vivek Gupta, and Anoop Kunchukuttan. 2022. Indicxnli: Evaluating multilingual inference for indian languages. *arXiv preprint arXiv:2204.08776*.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, et al. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

Abhinand Balachandran. 2023. Tamil-llama: A new tamil language model based on llama 2. *arXiv preprint arXiv:2311.05845*.

BhabhaAI. 2024. Gajendra-v0.1. `https://huggingface.co/BhabhaAI/Gajendra-v0.1`. Accessed: 2024-10-29.

BigScienceWorkshop. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Po-Heng Chen and Yun-Nung Chen. 2024. Efficient unseen language adaptation for multilingual pre-trained language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18983–18994, Miami, Florida, USA. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Zoltan Csaki, Pian Pawakapan, Urmish Thakker, and Qiantong Xu. 2023. Efficiently adapting pretrained language models to new languages. *arXiv preprint arXiv:2311.05741*.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

cyberblip. 2024. Travel india dataset. `https://huggingface.co/datasets/cyberblip/Travel_india`. Accessed: 2024-10-29.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, et al. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

Hugging Face. 2025. Supervised fine-tuning trainer. https://github.com/huggingface/trl/blob/main/trl/trainer/sft_trainer.py. Accessed: 2025-02-05.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. 2024. Open llm leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard.

Jay Gala, Thanmay Jayakumar, Jaavid Aktar Husain, Mohammed Safi Ur Rahman Khan, Diptesh Kanojia, Ratish Puduppully, Mitesh M Khapra, Raj Dabre, Rudra Murthy, Anoop Kunchukuttan, et al. 2024. Airavata: Introducing hindi instruction-tuned llm. *arXiv preprint arXiv:2401.15006*.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

GenVRadmin. 2024. Aryabhatta-gemmaorca-merged. https://huggingface.co/GenVRadmin/AryaBhatta-GemmaOrca-Merged. Accessed: 2024-10-29.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Raviraj Joshi, Kanishk Singla, Anusha Kamath, Raunak Kalani, Rakesh Paul, Utkarsh Vaidya, Sanjay Singh Chauhan, Niranjan Wartikar, and Eileen Long. 2024. Adapting multilingual llms to low-resource languages using continued pre-training and synthetic corpus. *arXiv preprint arXiv:2410.14815*.

Aditya Kallappa, Palash Kamble, Abhinav Ravi, Akshat Patidar, Vinayak Dhruv, Deepak Kumar, Raghav Awasthi, Arveti Manjunath, Himanshu Gupta, Shubham Agarwal, et al. 2025. Krutrim llm: Multilingual foundational model for over a billion people. *arXiv preprint arXiv:2502.09642*.

Mohammed Safi Ur Rahman Khan, Priyam Mehta, Ananth Sankar, Umashankar Kumaravelan, Sumanth Doddapaneni, Sparsh Jain, Anoop Kunchukuttan, Pratyush Kumar, Raj Dabre, Mitesh M Khapra, et al. 2024. Indicllmsuite: A blueprint for creating pre-training and fine-tuning datasets for indian languages. *arXiv preprint arXiv:2403.06350*.

Hele-Andra Kuulmets, Taido Purason, Agnes Luhtaru, and Mark Fishel. 2024. Teaching llama a new language through cross-lingual knowledge transfer. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3309–3325, Mexico City, Mexico. Association for Computational Linguistics.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

Llama-Nanda. 2024. Llama-3-nanda-10b-chat. https://github.com/mbzuai-nlp/Llama-3-Nanda-10B-Chat/blob/main/Llama-3-Nanda-10B-Chat-Paper.pdf.

Risto Luukkonen, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, et al. 2023. Fingpt: Large generative models for a small language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2710–2726.

Samin Mahdizadeh Sani, Pouya Sadeghi, Thuy-Trang Vu, Yadollah Yaghoobzadeh, and Gholamreza Haffari. 2025. Extending LLMs to new languages: A case study of llama and Persian adaptation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8868–8884, Abu Dhabi, UAE. Association for Computational Linguistics.

Arindam Mitra, Hamed Khanpour, Corby Rosset, and Ahmed Awadallah. 2024. Orca-math: Unlocking the potential of slms in grade school math. *arXiv preprint arXiv:2402.14830*.

msinankhan1. 2024. India tax faqs dataset. https://huggingface.co/datasets/msinankhan1/India_Tax_FAQs. Accessed: 2024-10-29.

OpenAI. 2024. Gpt-4o system card.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017.

prnv19. 2024. Upsc faq dataset. https://huggingface.co/datasets/prnv19/UPSC_FAQ. Accessed: 2024-10-29.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan Ak, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, et al. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Sarvamai. 2024. sarvam-2b-v0.5. https://huggingface.co/sarvamai/sarvam-2b-v0.5. Accessed: 2024-10-29.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. 2024. Continual learning of large language models: A comprehensive survey. *arXiv preprint arXiv:2404.16789*.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, et al. 2024a. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *arXiv preprint arXiv:2412.03304*.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024b. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. 2024. Musr: Testing the limits of chain-of-thought with multistep soft reasoning.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Gemma Team. 2024a. Gemma 2: Improving open language models at a practical size.

Llama Team. 2024b. The llama 3 herd of models.

Atula Tejaswi, Nilesh Gupta, and Eunsol Choi. 2024. Exploring design choices for building language-specific llms. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10485–10500.

Cagri Toraman. 2024. Llamaturk: Adapting open-source generative large language models for low-resource language.

Hieu Tran, Zhichao Yang, Zonghai Yao, and Hong Yu. 2024. Bioinstruct: instruction tuning of large language models for biomedical natural language processing. *Journal of the American Medical Informatics Association*, 31(9):1821–1832.

Unsloth AI. 2023. Phi-4. https://unsloth.ai/blog/phi4. Accessed: 2025-02-08.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Anh-Dung Vo, Minseong Jung, Wonbeen Lee, and Daewoo Choi. 2024. Redwhale: An adapted korean llm through efficient continual pretraining.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl.

Jiachen T. Wang, Tong Wu, Dawn Song, Prateek Mittal, and Ruoxi Jia. 2024a. GREATS: Online selection of high-quality data for LLM training in every iteration. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017a. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017b. Crowdsourcing multiple choice science questions.

Longhui Yu, Weisen Jiang, Han Shi, YU Jincheng, Zhengying Liu, Yu Zhang, James Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. Metamath: Bootstrap your own mathematical questions for large language models. In *The Twelfth International Conference on Learning Representations*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models.

A Model Replication

The hyper-parameters used for training can be seen below in Table 11. The initial training attempts using a portion of the data (i.e 8% samples) were done on various different devices, the final models were trained on a single H200 SXM for 55,56,54 hours each using Qwen2.5-14B-Instruct, Phi-4, Qwen2.5-14B-base respectively.

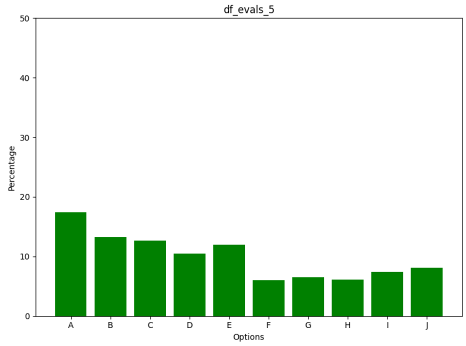

Figure 3: Distribution of respnse choices of our model training from qwen-base variant over MMLU-Pro

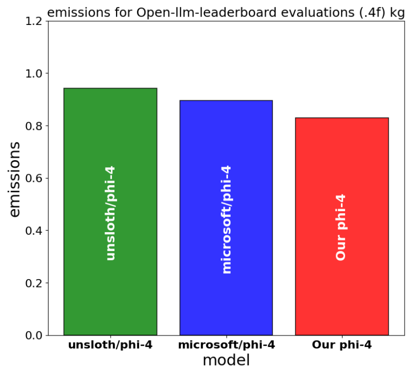

Figure 4: Emissions : open-llm-leaderboard evaluation

Hyperparameter		Value
Seed	Row Shuffling	1024
	Dataset Sampling	1024
	Training	1024
	Random State	1024
Epochs		1
Total Batch Size		600
	Batch Size	40
	Gradient Accumulation	15
Learning Rate		2e-5
Weight Decay		1e-2
Warmup Steps		0

Table 11: Training hyper-parameters used

The initially collected dataset sources, sample sizes and the later used sample counts can be seen in Table 12 along with the ratios of each language. The sampling within each dataset is done at random

Domain	Dataset	Total Samples	Used Samples	Hindi Ratio	Original Source
Legal FAQ	India Law	51,210	51,210	N/A	(Aditya2411, 2024)
Cooking Recipes	India Recipe	13,742	13,742	*	**
Travel FAQ	India Travel	2,000	2,000	N/A	(cyberblip, 2024)
Tax FAQ	India TAX	2,235	2,235	N/A	(msinankhan1, 2024)
General Knowledge	India UPSC	620	620	N/A	(prnv19, 2024)
General	BoolQ	18,799	18,799	N/A	(Clark et al., 2019)
General	Context MCQs	18,505	18,505	N/A	(Lai et al., 2017) (Welbl et al., 2017b)
General	ARC challenge	2,835	2,835	N/A	(Clark et al., 2018)
General	ARC Easy	5,637	5,637	N/A	(Clark et al., 2018)
General	Winogrande XL	82,973	10,000	85	(Sakaguchi et al., 2021)
Biology	Camel Biology	39,990	39,990	N/A	(Li et al., 2023)
Biology	Bio Instruct	49,956	49,956	N/A	(Tran et al., 2024)
Coding	MBPP	928	928	N/A	(Austin et al., 2021)
Chemistry	Camel Chemistry	39,975	39,975	N/A	(Li et al., 2023)
NLI	XNLI/IndicXNLI	395,192	20,000	80	(Conneau et al., 2018) (Aggarwal et al., 2022)
Math	MATH QA	68,583	10,000	50	(Amini et al., 2019)
Math	Math Hard	4,593	4,593	N/A	(Hendrycks et al., 2021)
Math	Math Easy	14,953	14,953	N/A	(Hendrycks et al., 2021)
Math	GSM8K	14,937	14,973	N/A	(Cobbe et al., 2021)
Math	Camel Math	99,626	10,000	50	(Li et al., 2023)
Math	META Math	199,782	20,000	80	(Yu et al., 2023)
Math	Orca Math	399,847	10,000	50	(Mitra et al., 2024)
Medical	MedMCQA	372,779	20,000	70	(Pal et al., 2022)
Paraphrasing	Aya Paraphrase	1,001	1,001	N/A	(Singh et al., 2024b)
Physics	Camel Physics	39,995	39,995	N/A	(Li et al., 2023)
Reasoning	PIQA	35,396	35,396	N/A	(Bisk et al., 2020)
Reasoning	SIQA	65,630	20,000	80	(Sap et al., 2019)
Simplification	Aya Simplify	994,944	10,000	60	(Singh et al., 2024b)
Summarization	XLSum	79,625	10,000	50	(Hasan et al., 2021)
Translation	Aya Translate	1,156	1,156	N/A	(Singh et al., 2024b)
		3,117,450	**485,469**		

Table 12: Sources of our training dataset's samples and their distributions

* indicates that the original dataset had a language mix of English and Hindi. Among the rest, initial sample counts were 50:50 for each language and were later individually sampled based on the ratios mentioned for each dataset.

** The dataset at the time of data collection was publicly available on hf without a restrictive license, but is currently made private.

using the seed specified in Table 11. The samples were sorted in ascending order based on input size and the longest 600 samples in terms of input token count were added in the beginning of the training data.

B Datasets and Benchmarks Info

The benchmarks used can be seen in Table 13 along their features like domain, original source, total number of samples, number of samples used and the ratio of Hindi samples among those used.

Benchmark	Source
ARC Easy	(Clark et al., 2018)
ARC Challenge	(Clark et al., 2018)
Context MCQs	(Lai et al., 2017), (Welbl et al., 2017b)
BoolQ	(Clark et al., 2019)
MMLU	(Hendrycks et al., 2020), (Singh et al., 2024a)
MMLU-Pro	(Wang et al., 2024b)
MATH-HARD	(Hendrycks et al., 2021)
GPQA	(Rein et al., 2023)
MuSR	(Sprague et al., 2024)
Bigbench-Hard	(Suzgun et al., 2022)

Table 13: Benchmarks used and their corresponding sources

C Results from other attempts

The results from other attempts with a smaller sized LLMs can be seen in Llama-3.1-8B: Table 15, Llama-3.2-3B: Table 16, Gemma-2-9B: Table 17, Gemma-2-2B: Table 18, Qwen-2.5-3B: Table 14.

D Model Choices

The choices selected by each of the models over each domain of MMLU-Pro can be seen in the below images Figure 5 to Figure 18.

Benchmarks	Ratio of	ARC-Challenge		ARC-Easy		MMLU		BoolQ		Context-MCQ		Overall Average		
Data used?	Hindi	En	Hi	En	Hi	En	Hi	En	Hi	En	Hi	En	Hi	Tot
No	10%	78.07	39.51	88.97	47.98	59.42	35.44	62.26	62.25	82.0	56.4	74.14	48.31	61.23
No	20%	77.65	40.19	88.72	50.00	59.92	34.63	62.35	62.28	75.9	53.2	72.91	48.06	60.48
No	30%	77.65	39.51	88.51	49.79	59.33	34.76	62.32	62.16	76.9	55.5	72.94	48.34	60.64
No	40%	77.56	40.44	88.59	50.63	59.92	34.38	62.39	63.35	76.1	52.5	72.91	48.04	60.48
No	50%	78.16	41.89	88.72	50.55	60.97	35.23	62.35	62.31	77.5	54.2	73.54	48.83	61.18
No	60%	78.50	41.81	88.72	50.46	61.00	35.40	62.35	62.31	78.2	54.7	73.75	48.93	61.34
No	70%	78.33	42.06	88.89	50.46	60.85	35.37	62.35	62.31	78.1	54.9	73.70	49.02	61.36
No	80%	78.24	42.32	88.59	50.55	60.86	35.36	62.35	62.31	78.1	55.3	73.62	49.16	61.39
No	90%	76.79	39.76	88.34	45.92	57.91	32.35	62.23	62.19	77.9	50.6	72.63	46.16	59.39
No	100%	75.77	38.91	87.88	45.54	57.76	31.98	62.26	62.19	76.7	50.8	72.07	45.88	58.97
Yes	10%	78.50	42.32	89.86	50.93	60.03	35.39	71.25	62.74	80.6	56.3	76.04	49.53	62.79
Yes	20%	77.99	39.93	88.80	50.25	59.74	34.51	62.54	62.07	74.5	53.2	72.71	47.99	60.35
Yes	30%	77.82	40.53	88.76	50.42	59.47	34.57	62.75	62.19	74.0	50.9	72.56	47.72	60.14
Yes	40%	77.82	40.53	88.64	50.38	59.67	34.09	62.72	62.22	71.3	49.3	72.03	47.30	59.67
Yes	50%	78.16	41.13	88.59	51.18	60.72	34.95	62.66	62.28	75.2	52.3	73.06	48.36	60.71
Yes	60%	78.50	41.47	88.72	50.42	60.68	35.17	62.45	62.34	76.3	53.1	73.33	48.50	60.91
Yes	70%	78.50	42.06	88.68	50.51	60.71	35.12	62.45	62.37	76.2	53.5	73.30	48.71	61.01
Yes	80%	78.58	42.24	88.72	50.51	60.76	35.24	62.42	62.37	76.6	53.6	73.41	48.79	61.10
Yes	90%	77.22	42.15	88.85	49.87	57.39	30.28	64.86	64.03	69.0	43.7	71.46	46.00	58.73
Yes	100%	75.77	38.91	87.88	45.54	57.76	31.98	63.79	62.80	72.1	43.7	71.46	44.58	58.02
Original		77.73	41.21	88.26	49.20	60.25	34.26	62.20	62.25	76.3	52.7	72.94	47.92	60.43

Table 14: Results (.2f) from each training attempt with 5% of our training data over Qwen2.5-3B-Instruct

Benchmarks	Ratio of	ARC-Challenge		ARC-Easy		MMLU		BoolQ		Context-MCQ		Overall Average		
Data used?	Hindi	En	Hi	En	Hi	En	Hi	En	Hi	En	Hi	En	Hi	Tot
No	10%	73.89	61.06	85.94	66.66	62.30	42.11	64.13	61.06	82.8	64.4	73.81	57.52	65.67
No	20%	75.43	55.72	87.37	69.40	63.09	42.95	63.94	61.49	83.2	65.3	74.60	58.97	66.78
No	30%	75.40	55.97	87.04	69.95	62.98	43.03	62.69	59.90	83.2	65.8	74.26	58.93	66.60
No	40%	73.63	54.86	86.66	68.56	62.34	42.25	63.91	61.76	82.2	65.2	73.74	58.52	66.13
No	50%	74.23	55.89	86.66	70.12	62.60	42.35	64.80	61.79	82.4	65.0	74.13	59.02	66.58
No	60%	72.70	54.86	84.81	67.97	60.65	42.06	64.46	60.97	82.1	65.2	72.94	58.21	65.58
No	70%	75.26	56.23	88.80	69.82	62.53	42.27	65.72	60.14	82.2	64.9	74.90	58.67	66.79
No	80%	74.23	54.69	86.24	68.10	62.18	42.62	64.53	61.27	81.5	64.9	73.73	58.31	66.02
No	90%	73.81	54.95	85.90	67.89	61.81	42.33	63.88	61.39	81.3	63.5	73.34	58.01	65.68
No	100%	73.81	55.03	86.07	68.64	61.57	42.30	63.88	57.48	80.8	64.3	73.22	57.55	65.38
Yes	10%	79.27	59.13	91.50	75.59	63.91	42.49	83.98	74.49	83.5	66.0	80.43	63.54	71.98
Yes	20%	79.35	58.79	91.41	76.47	64.01	43.65	85.96	79.66	84.5	66.6	81.05	65.03	73.04
Yes	30%	79.01	61.69	92.47	76.43	64.04	43.17	84.95	77.82	83.4	66.8	80.77	65.18	72.98
Yes	40%	79.18	61.35	91.62	76.68	63.62	43.27	84.98	74.79	83.7	65.6	80.62	64.34	72.48
Yes	50%	78.92	60.92	91.67	76.18	62.95	43.15	85.26	78.19	83.8	67.5	80.52	65.19	72.85
Yes	60%	77.39	60.07	92.00	75.97	63.44	43.43	85.02	78.37	82.2	66.5	80.01	64.87	72.44
Yes	70%	78.33	61.35	91.71	76.09	63.67	43.41	83.36	75.28	82.7	66.0	79.95	64.45	72.20
Yes	80%	76.79	58.79	89.73	75.42	62.84	42.91	83.27	74.27	82.2	66.4	78.97	63.56	71.26
Yes	90%	76.88	59.81	90.40	75.00	62.69	43.06	83.03	73.97	82.0	65.7	79.00	63.51	71.25
Yes	100%	76.54	59.81	89.73	75.72	62.54	43.70	82.35	77.00	81.2	67.5	78.47	64.74	71.61
Original		75.34	53.92	84.76	65.78	61.69	43.32	65.17	62.16	78.4	67.1	73.07	58.45	65.76

Table 15: Results (.2f) from each training attempt with 5% of our training data over LLama 3.1 8B

Benchmarks	Ratio of	ARC-Challenge		ARC-Easy		MMLU		BoolQ		Context-MCQ		Overall Average		
Data used?	**Hindi**	**En**	**Hi**	**En**	**Hi**	**En**	**Hi**	**En**	**Hi**	**En**	**Hi**	**En**	**Hi**	**Tot**
No	10%	60.83	41.97	75.71	55.47	51.60	33.69	65.44	62.71	68.6	49.1	64.44	48.59	56.51
No	20%	60.75	43.60	76.85	55.80	52.79	33.86	65.01	62.55	69.2	51.1	64.92	49.38	57.15
No	30%	60.66	42.32	76.26	55.13	53.28	33.84	64.64	62.19	68.4	51.0	64.65	48.89	56.77
No	40%	60.49	41.97	75.46	55.13	52.28	33.67	64.46	62.61	69.7	50.9	64.48	48.86	56.67
No	50%	60.41	44.28	76.09	55.51	51.71	31.63	65.20	62.77	68.0	52.3	64.28	49.30	56.79
No	60%	60.49	45.56	76.34	56.43	51.24	32.36	65.29	62.98	68.7	51.8	64.41	49.82	57.12
No	70%	62.20	45.64	77.31	57.23	52.50	32.01	64.98	62.49	68.9	51.5	65.18	49.78	57.48
No	80%	61.94	44.88	76.85	56.18	52.48	33.06	65.56	61.76	70.4	53.7	61.94	49.91	57.68
No	90%	63.31	46.84	77.99	58.21	49.12	30.54	63.70	62.28	68.6	52.8	64.54	50.13	57.34
No	100%	62.71	45.98	77.98	58.83	52.07	33.01	65.38	62.09	70.4	54.3	65.71	50.84	58.28
Yes	10%	69.45	48.37	84.34	62.03	55.20	33.56	72.75	72.52	72.0	53.1	70.75	53.92	62.33
Yes	20%	68.08	47.01	84.13	61.32	54.30	33.34	70.15	69.65	72.3	52.8	69.79	52.82	61.31
Yes	30%	67.91	47.52	84.13	62.28	54.46	34.80	72.47	73.17	71.8	55.5	70.15	54.65	62.40
Yes	40%	68.08	47.44	83.58	62.41	53.88	33.69	70.36	71.67	72.6	53.8	69.70	53.80	61.75
Yes	50%	69.11	48.38	83.88	63.26	54.00	34.05	73.58	74.30	71.1	54.0	70.33	54.80	62.57
Yes	60%	67.15	47.86	83.37	62.92	53.61	33.34	75.16	75.55	70.9	53.0	70.04	54.53	62.28
Yes	70%	67.15	47.95	83.16	62.75	53.55	34.17	73.57	72.77	71.6	54.3	69.80	54.39	62.10
Yes	80%	67.58	46.08	82.95	62.54	51.69	32.10	73.12	73.66	70.0	51.7	69.06	53.21	61.14
Yes	90%	63.91	47.18	79.88	60.35	48.89	31.31	69.51	62.96	68.7	54.0	66.18	51.16	58.70
Yes	100%	68.00	48.63	83.12	62.96	52.87	35.91	70.06	67.85	71.8	55.8	69.17	54.23	61.70
Original		62.12	40.70	74.12	52.48	50.37	31.30	62.72	62.22	68.6	41.2	63.58	45.58	54.58

Table 16: Results (.2f) from each training attempt with 5% of our training data over Llama 3.2 3B

Benchmarks	Ratio of	ARC-Challenge		ARC-Easy		MMLU		BoolQ		Context-MCQ		Overall Average		
Data used?	**Hindi**	**En**	**Hi**	**En**	**Hi**	**En**	**Hi**	**En**	**Hi**	**En**	**Hi**	**En**	**Hi**	**Tot**
No	10%	86.52	75.25	94.52	87.24	68.53	53.93	86.82	83.69	86.7	79.0	84.62	75.82	80.22
No	20%	87.11	75.68	94.57	87.11	68.46	53.89	86.66	83.42	86.9	78.6	84.74	75.80	80.27
No	30%	86.34	75.42	94.86	87.28	68.74	53.85	86.91	83.94	87.2	78.4	84.81	75.42	80.29
No	40%	86.86	75.85	95.32	87.45	68.88	54.36	86.60	83.76	86.8	78.1	84.89	75.91	80.40
No	50%	86.86	75.51	95.11	87.41	68.49	53.96	86.82	84.06	87.1	77.8	84.88	75.75	80.31
No	60%	87.11	76.62	95.70	87.83	68.43	53.73	86.60	84.15	87.2	78.3	85.01	76.12	80.57
No	70%	88.65	78.07	95.16	89.27	71.32	56.13	87.76	85.01	88.3	79.1	86.24	77.51	81.88
No	80%	88.22	77.47	95.24	88.93	70.00	55.06	87.19	85.13	87.1	85.13	85.55	77.04	81.30
No	90%	86.94	76.00	95.28	87.58	69.42	54.61	86.48	84.12	87.0	79.2	85.02	76.30	80.66
No	100%	88.48	76.36	95.37	89.10	70.00	54.36	86.64	84.34	87.1	79.1	85.52	76.65	81.08
Yes	10%	87.79	78.24	95.70	90.27	68.87	54.18	86.85	84.91	87.2	79.1	85.28	77.34	81.31
Yes	20%	87.54	77.81	95.45	90.31	68.76	53.99	86.85	84.91	87.5	79.8	85.22	77.36	81.29
Yes	30%	87.88	78.41	95.87	90.10	68.87	54.60	86.81	85.19	87.4	79.3	85.37	77.50	81.44
Yes	40%	87.80	77.38	94.91	89.86	68.25	53.56	86.85	84.83	87.5	79.3	85.06	77.39	81.02
Yes	50%	87.46	77.73	95.37	90.28	68.25	53.57	86.97	84.89	87.2	79.7	85.05	77.23	81.14
Yes	60%	88.31	78.41	95.74	90.65	68.62	54.18	86.81	85.19	88.0	78.9	85.50	77.47	81.48
Yes	70%	89.16	78.84	95.20	89.56	71.17	56.20	88.04	85.56	88.5	78.4	86.42	77.71	82.06
Yes	80%	87.62	78.58	95.45	89.94	67.91	52.55	86.88	84.12	87.6	78.1	85.09	76.66	80.87
Yes	90%	88.22	78.66	95.37	90.19	68.59	53.70	86.85	84.30	87.5	79.8	85.30	77.33	81.32
Yes	100%	87.88	78.24	95.03	90.02	69.21	53.31	87.00	85.44	87.7	79.4	85.37	77.28	81.32
Original		88.74	79.18	95.33	88.76	71.00	56.14	87.89	84.67	88.2	77.3	86.23	77.21	81.72

Table 17: Results (.2f) from each training attempt with 5% of our training data over Gemma 2 9B

Benchmarks	Ratio of	ARC-Challenge		ARC-Easy		MMLU		BoolQ		Context-MCQ		Overall Average		
Data used?	Hindi	En	Hi	En	Hi	En	Hi	En	Hi	En	Hi	En	Hi	Tot
No	10%	65.36	45.39	80.26	58.96	49.54	35.22	77.22	75.19	64.7	54.6	67.42	53.87	60.64
No	20%	64.93	45.31	80.01	58.80	49.20	35.08	76.64	74.89	64.4	54.0	67.04	53.61	60.32
No	30%	64.68	46.67	80.35	59.43	49.53	35.17	76.06	74.92	65.0	54.6	67.12	54.16	60.64
No	40%	70.22	49.66	83.63	63.97	52.08	36.83	81.83	76.48	68.0	57.6	71.15	56.91	64.03
No	50%	61.86	45.81	79.04	57.99	48.09	34.49	76.54	75.34	63.7	54.0	65.85	53.52	59.69
No	60%	61.60	45.56	79.58	58.58	47.99	34.39	75.65	75.71	64.6	54.0	65.88	53.65	59.77
No	70%	63.22	47.78	63.22	59.42	48.26	34.33	76.97	76.13	62.9	52.9	66.33	54.11	60.22
No	80%	65.53	46.50	81.73	61.03	50.29	35.40	76.79	75.80	64.6	55.3	67.79	54.81	61.30
No	90%	65.10	46.59	81.73	60.19	50.14	35.41	76.64	75.01	65.0	54.1	67.72	54.26	60.99
No	100%	67.92	48.81	82.79	62.33	51.42	36.02	80.24	76.14	67.6	56.9	69.99	56.04	63.01
Yes	10%	66.38	48.12	82.24	62.33	49.00	34.76	75.35	72.56	64.2	54.4	67.43	54.43	60.93
Yes	20%	66.13	48.89	82.24	62.67	48.85	34.84	74.92	71.86	63.8	53.0	67.19	54.25	60.72
Yes	30%	65.53	48.46	82.15	62.25	49.11	34.87	73.91	71.03	64.2	53.1	66.98	53.94	60.46
Yes	40%	67.92	48.04	82.45	62.42	50.67	36.23	77.00	75.19	65.4	55.6	68.69	55.49	62.09
Yes	50%	68.08	51.02	83.96	64.05	47.99	34.64	76.66	74.30	63.9	54.7	68.12	55.74	61.93
Yes	60%	68.08	50.34	84.21	64.52	47.76	34.62	72.75	70.32	63.5	53.7	67.26	54.70	60.98
Yes	70%	68.25	51.45	84.55	64.73	48.31	34.78	75.87	73.35	64.6	54.3	68.31	55.72	62.02
Yes	80%	66.47	49.83	83.50	63.55	48.70	34.62	73.67	69.90	63.4	53.9	67.15	54.36	60.75
Yes	90%	67.06	49.74	83.42	63.76	49.44	35.32	73.49	69.50	64.2	53.3	67.52	54.32	60.92
Yes	100%	67.58	49.40	83.00	63.09	50.93	36.01	75.75	73.72	66.0	54.6	68.65	55.36	62.00
Original		71.50	51.62	84.05	64.31	51.13	36.49	82.69	77.12	70.9	59.2	72.05	57.74	64.90

Table 18: Results (.2f) from each training attempt with 5% of our training data over Gemma 2 2B

Figure 5: Each model's choice distribution over MMLU-Pro : Biology

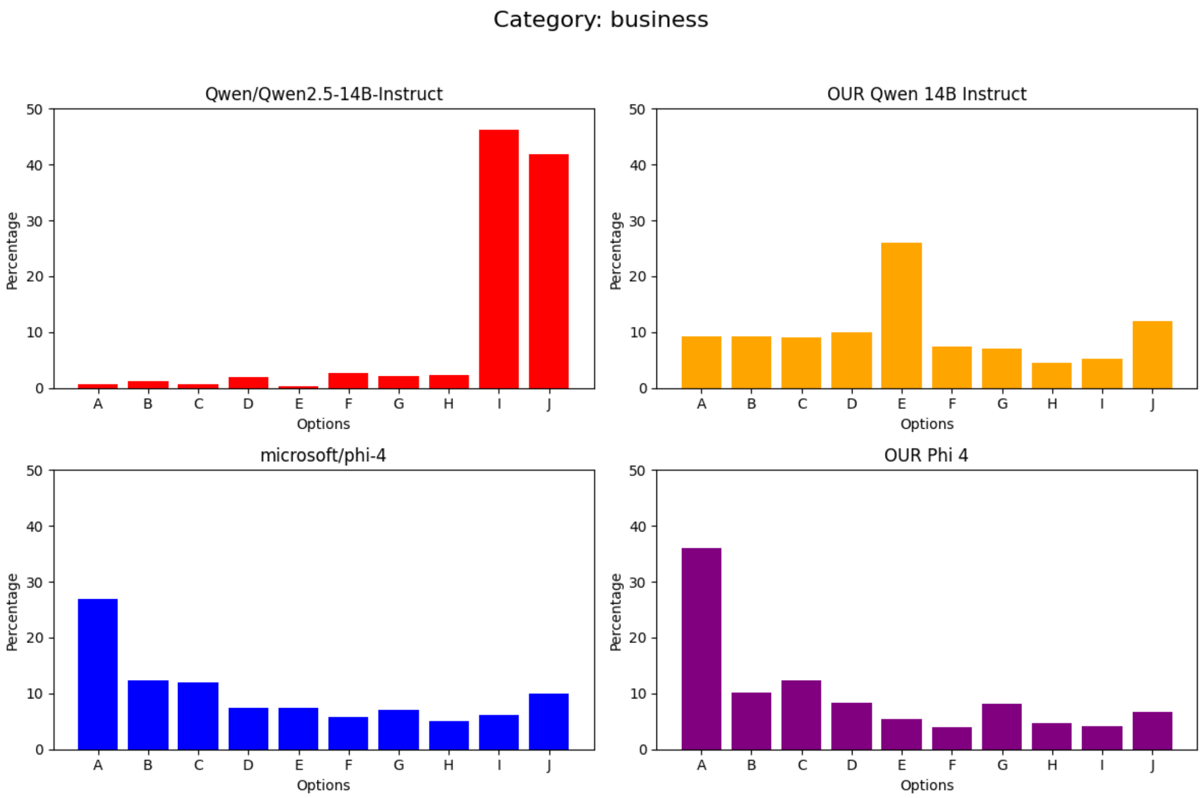

Figure 6: Each model's choice distribution over MMLU-Pro : Business

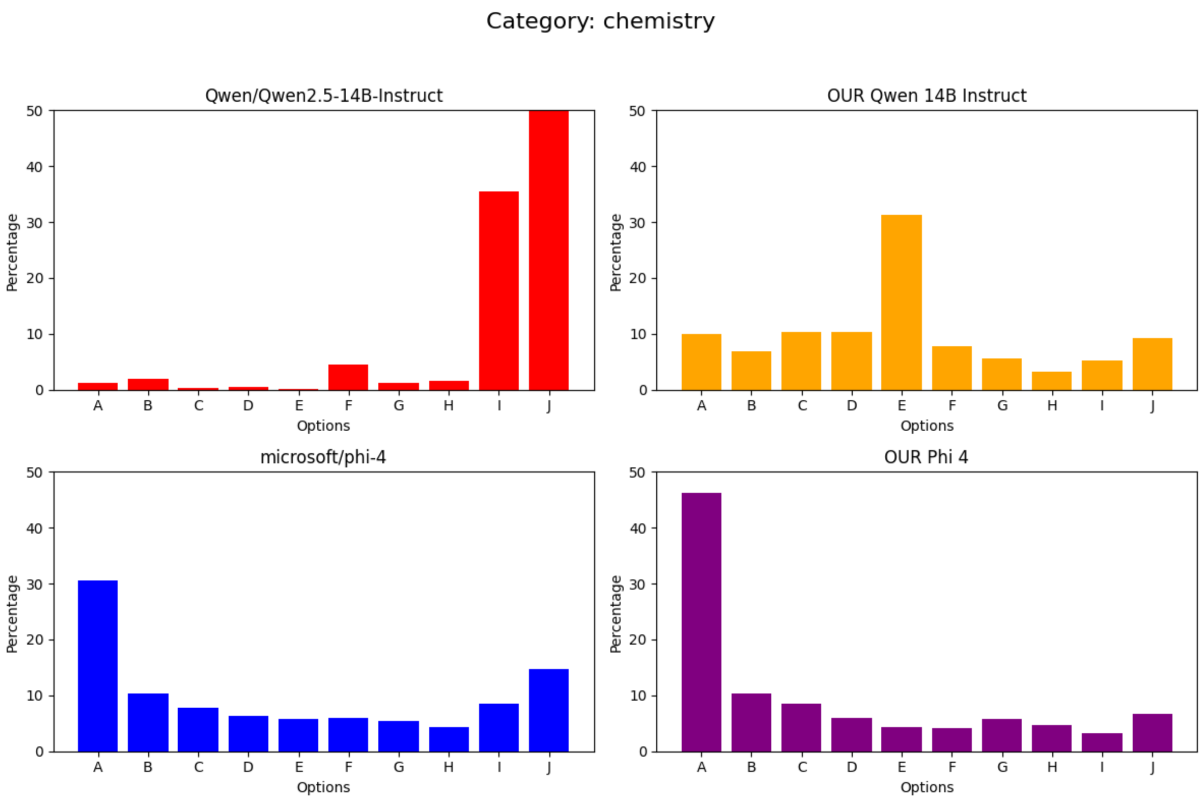

Figure 7: Each model's choice distribution over MMLU-Pro : Chemistry

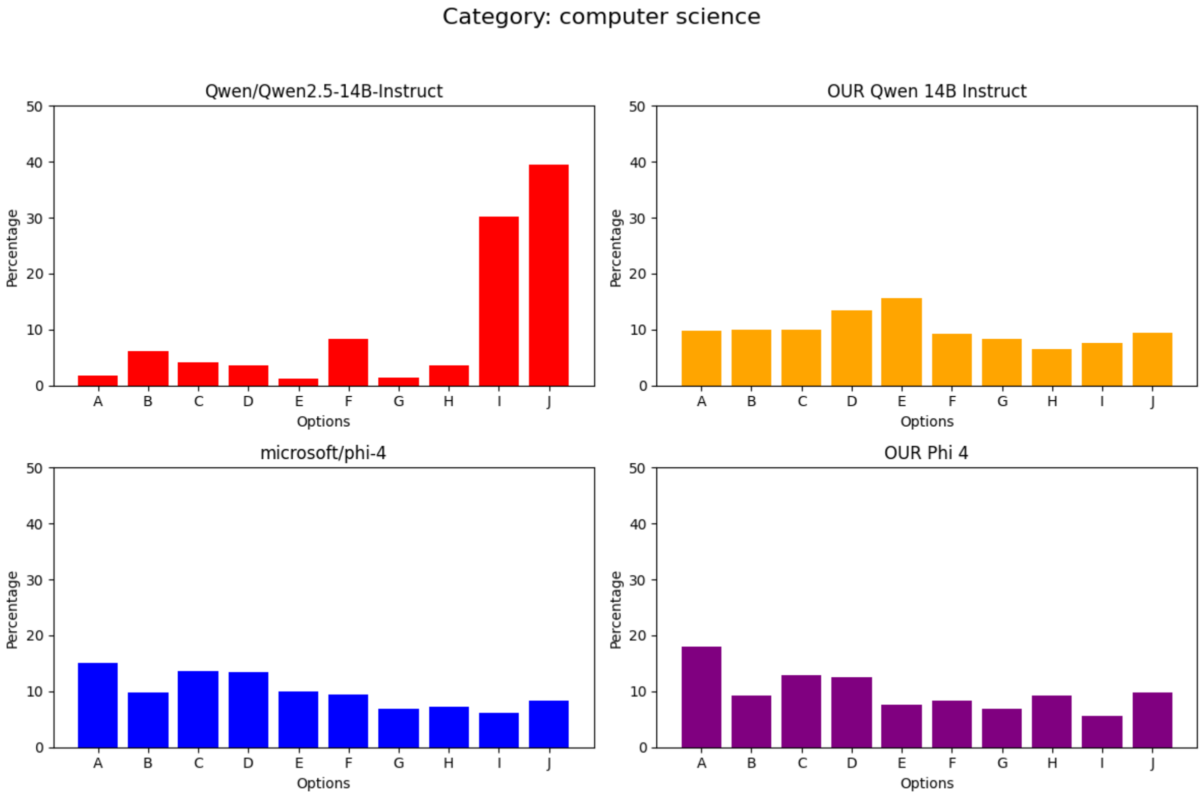

Figure 8: Each model's choice distribution over MMLU-Pro : CS

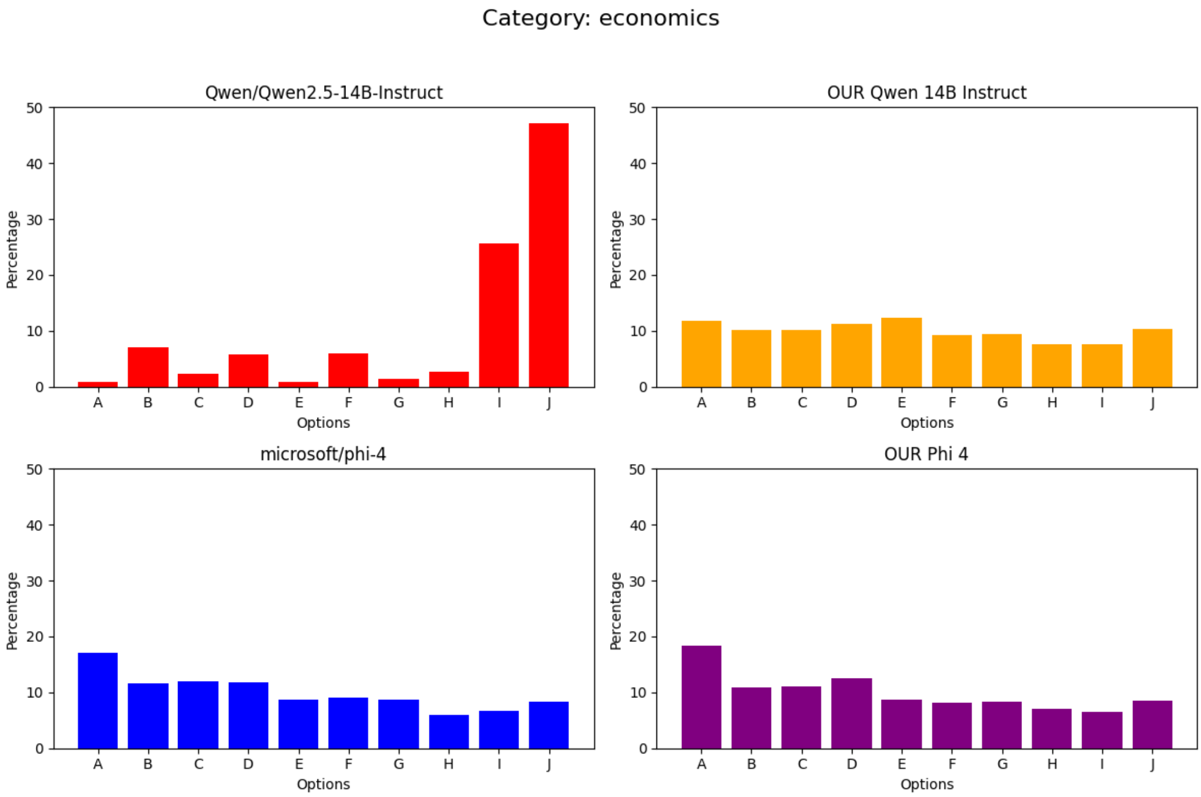

Figure 9: Each model's choice distribution over MMLU-Pro : Economics

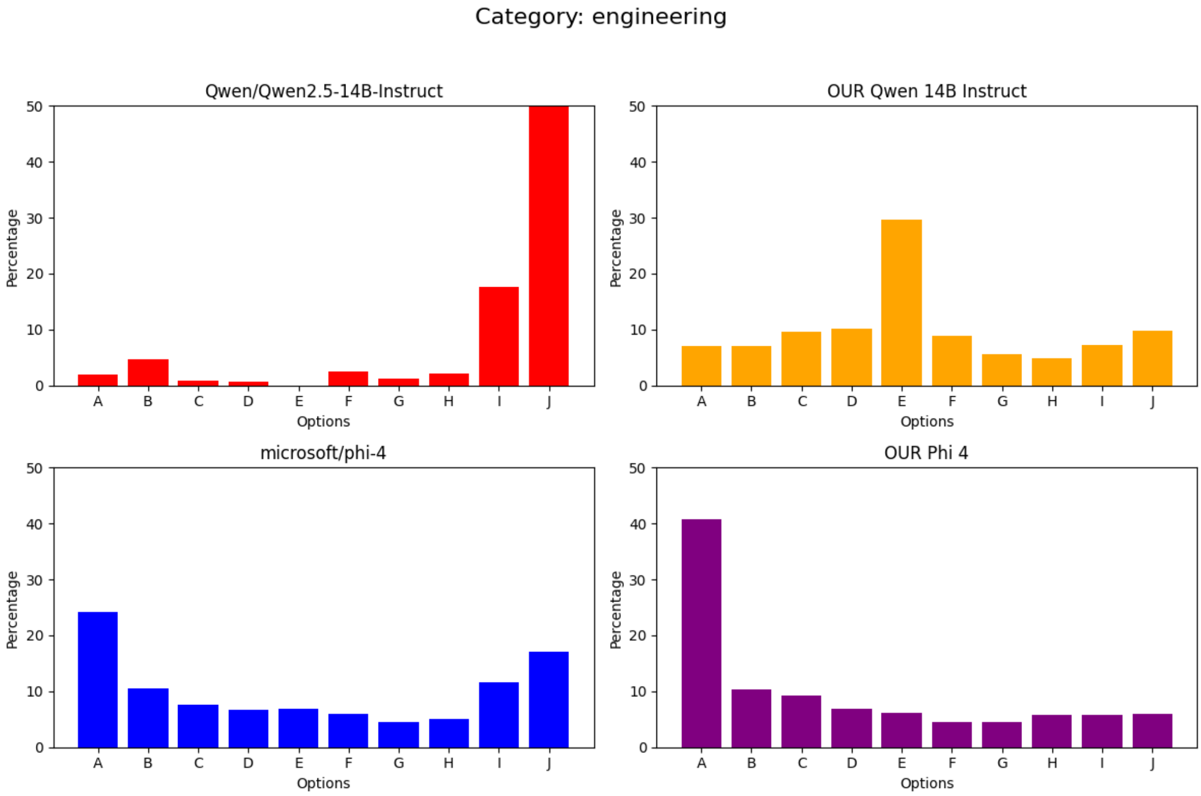

Figure 10: Each model's choice distribution over MMLU-Pro : Engineering

Figure 11: Each model's choice distribution over MMLU-Pro : Health

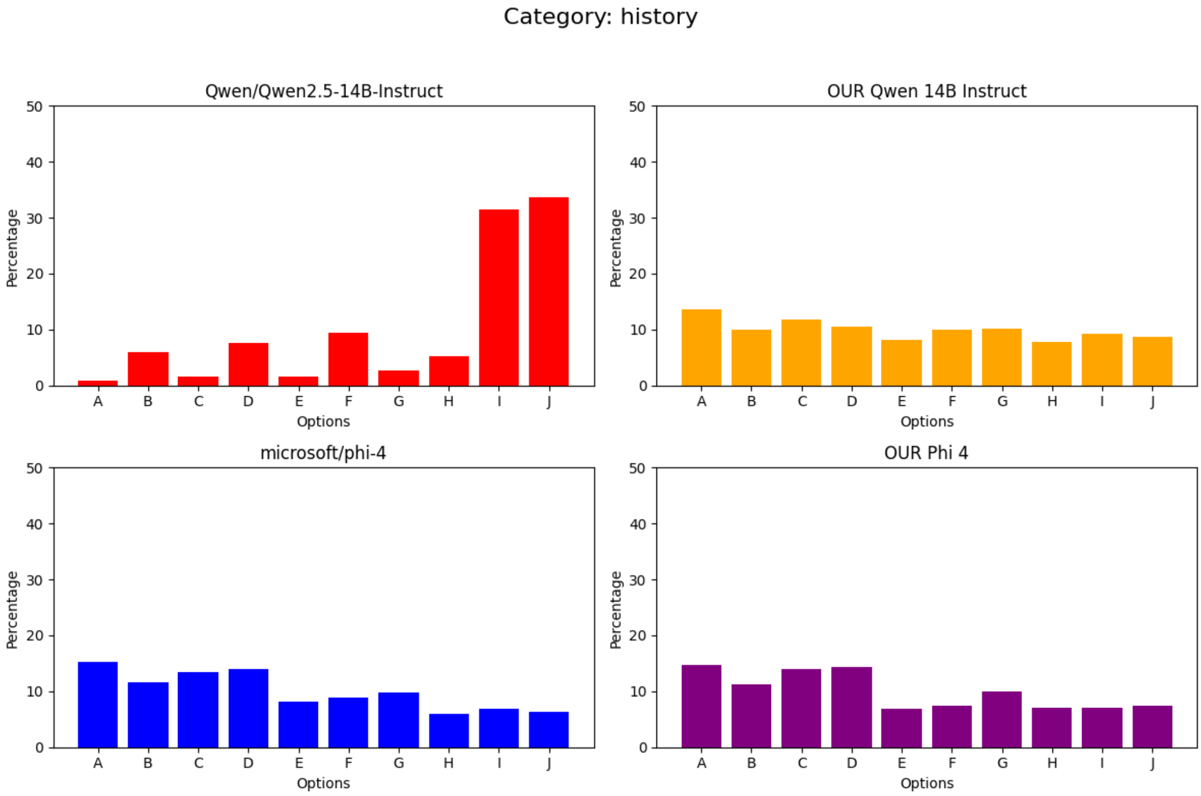

Figure 12: Each model's choice distribution over MMLU-Pro : History

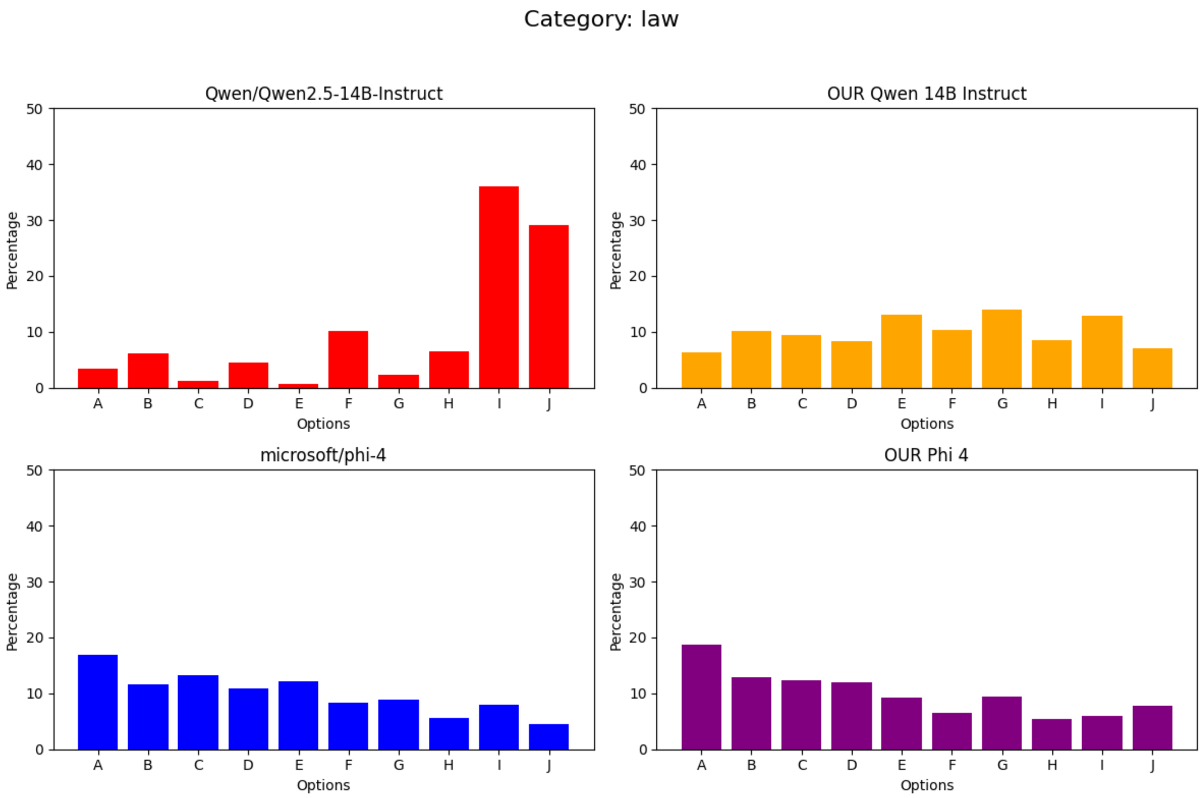

Figure 13: Each model's choice distribution over MMLU-Pro : Law

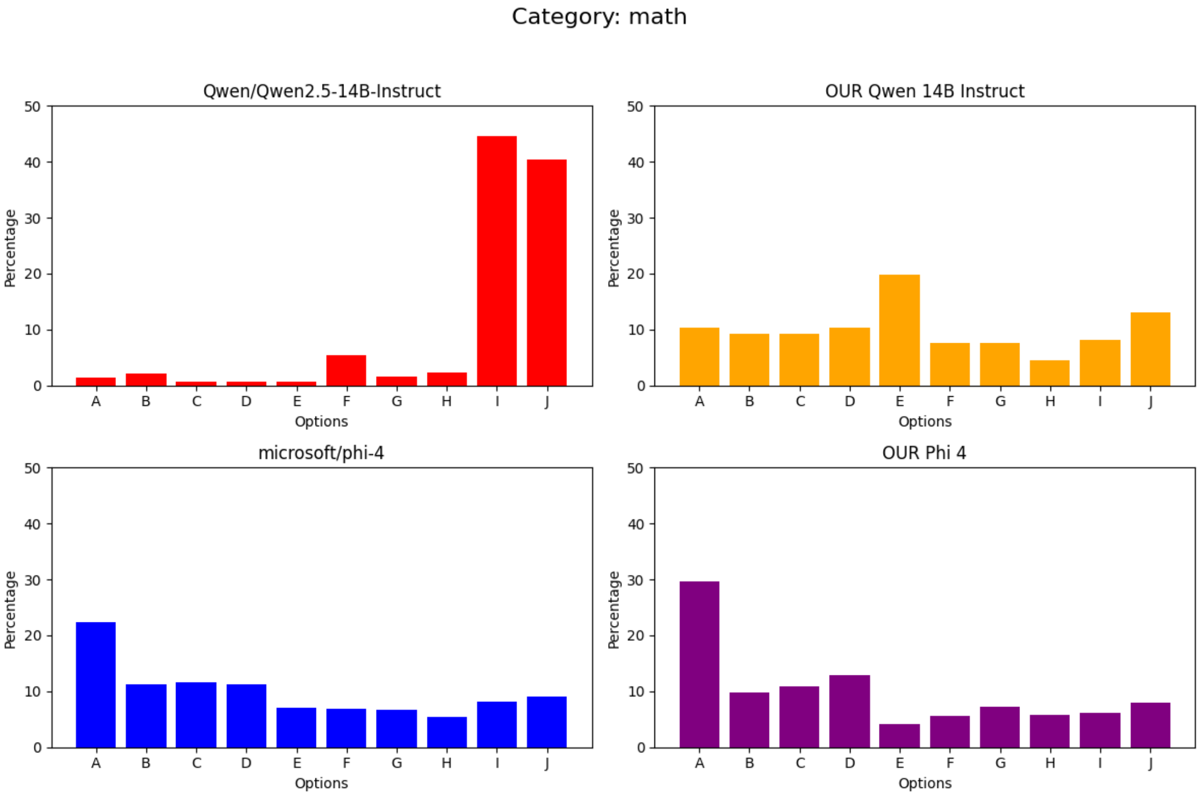

Figure 14: Each model's choice distribution over MMLU-Pro : Math

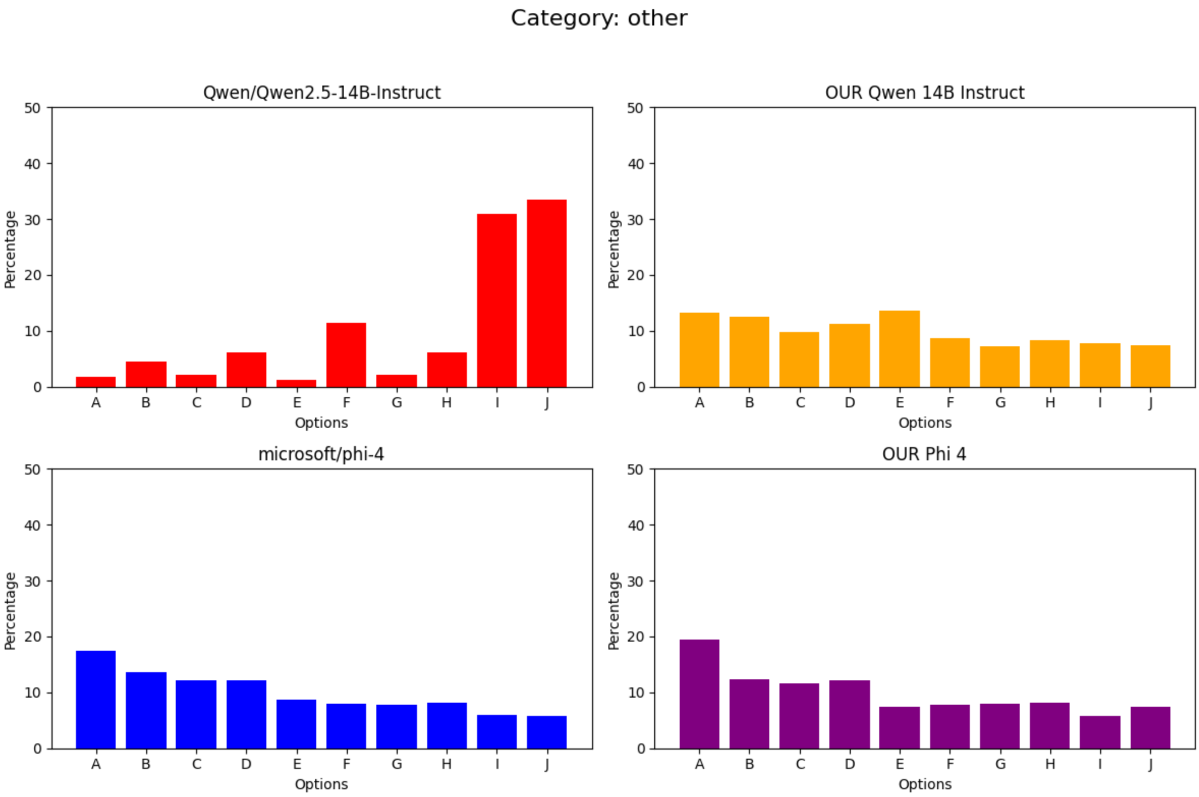

Figure 15: Each model's choice distribution over MMLU-Pro : Other

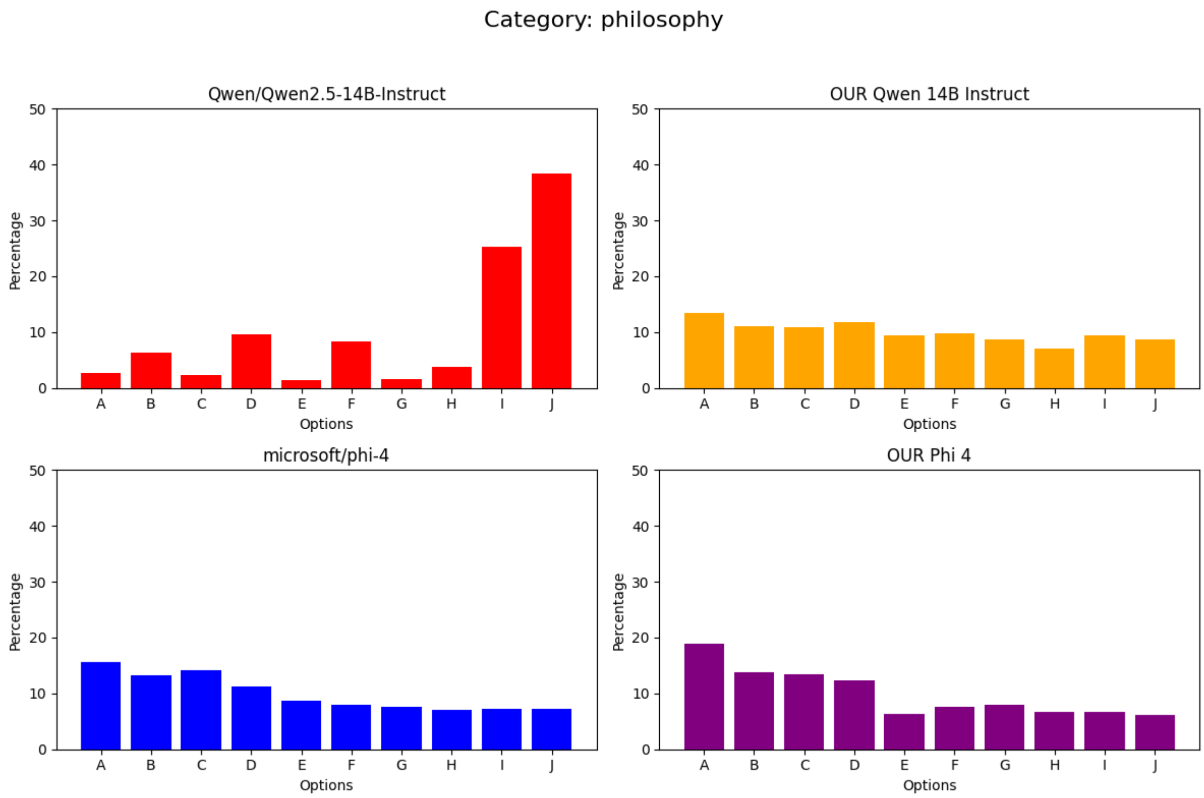

Figure 16: Each model's choice distribution over MMLU-Pro : Philosophy

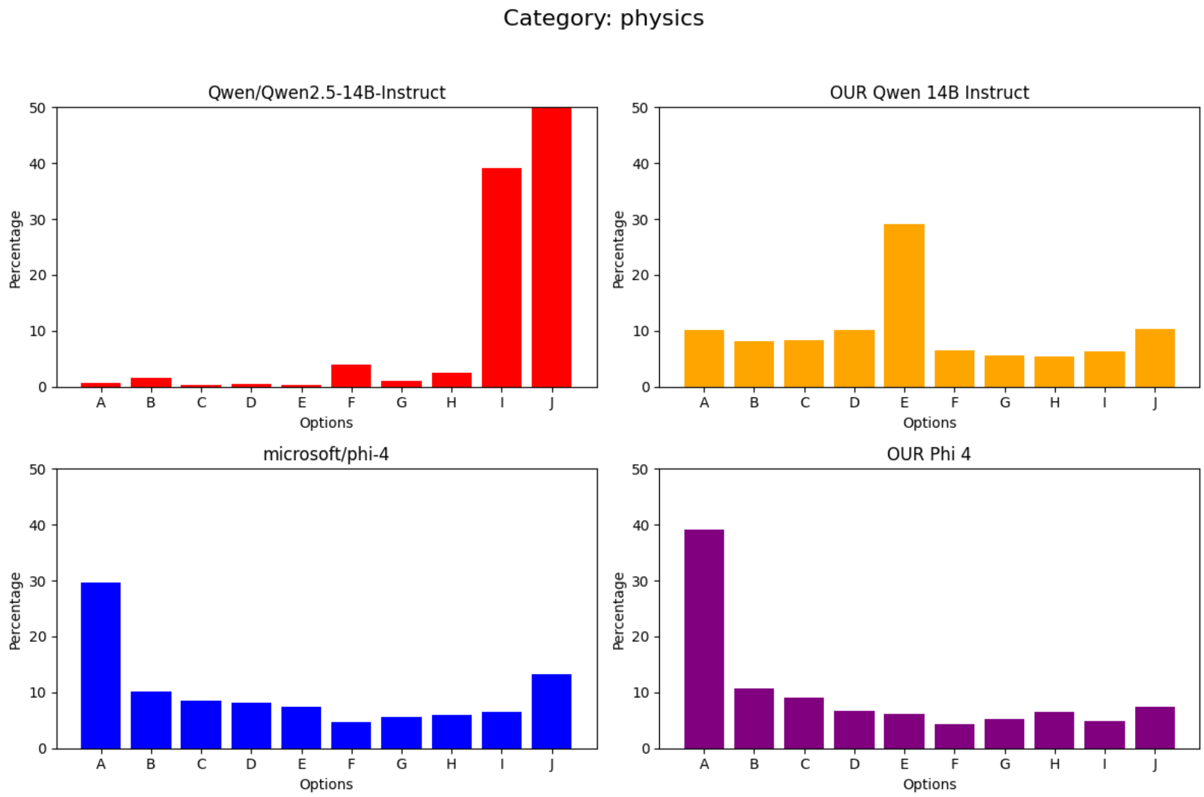

Figure 17: Each model's choice distribution over MMLU-Pro : Physics

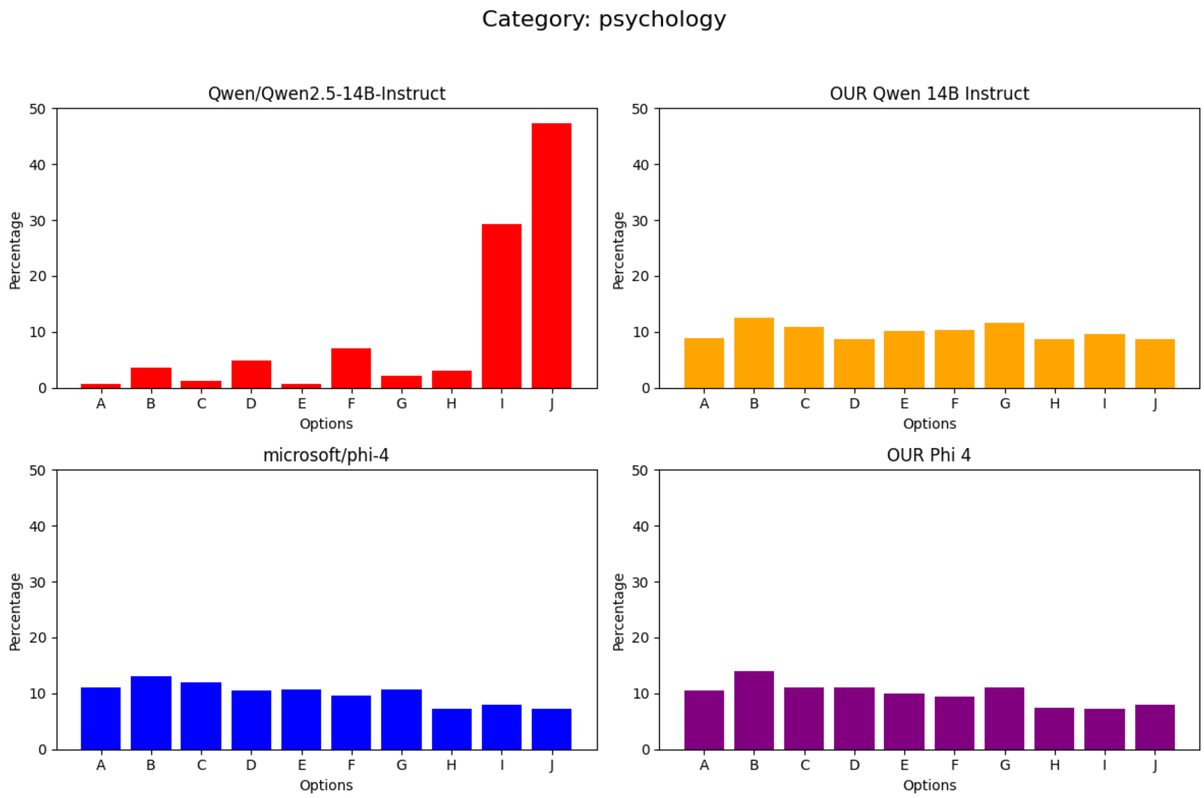

Figure 18: Each model's choice distribution over MMLU-Pro : Psychology