

Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone

Microsoft

Abstract

We introduce **phi-3-mini**, a 3.8 billion parameter language model trained on 3.3 trillion tokens, whose overall performance, as measured by both academic benchmarks and internal testing, rivals that of models such as Mixtral 8x7B and GPT-3.5 (e.g., **phi-3-mini** achieves 69% on MMLU and 8.38 on MT-bench), despite being small enough to be deployed on a phone. Our training dataset is a scaled-up version of the one used for **phi-2**, composed of heavily filtered publicly available web data and synthetic data. The model is also further aligned for robustness, safety, and chat format. We also provide parameter-scaling results with a 7B, 14B models trained for 4.8T tokens, called **phi-3-small**, **phi-3-medium**, both significantly more capable than **phi-3-mini** (e.g., respectively 75%, 78% on MMLU, and 8.7, 8.9 on MT-bench). To enhance multilingual, multimodal, and long-context capabilities, we introduce three models in the **phi-3.5** series: **phi-3.5-mini**, **phi-3.5-MoE**, and **phi-3.5-Vision**. The **phi-3.5-MoE**, a 16 x 3.8B MoE model with 6.6 billion active parameters, achieves superior performance in language reasoning, math, and code tasks compared to other open-source models of similar scale, such as Llama 3.1 and the Mixtral series, and on par with Gemini-1.5-Flash and GPT-4o-mini. Meanwhile, **phi-3.5-Vision**, a 4.2 billion parameter model derived from **phi-3.5-mini**, excels in reasoning tasks and is adept at handling both single-image and text prompts, as well as multi-image and text prompts.

1 Introduction

The striking progress of AI in the last few years can be largely attributed to major efforts throughout the world towards *scaling-up* to ever-larger models and datasets. Large Language Models (LLMs) have steadily increased in size from a mere billion parameters just five years ago (GPT-2 had 1.5 billion parameters [RWC⁺19]) to trillion parameters today. The impetus for this effort originates in the seemingly predictable improvement one obtains by training large models, the so-called *scaling laws* [KMH⁺20, HBM⁺22, MRB⁺23]. However these laws assume a “fixed” data source. This assumption is now significantly disrupted by the existence of frontier LLMs themselves, which allow us to interact with data in novel ways. In our previous works on the phi models [GZA⁺23, LBE⁺23, JBA⁺23] it was shown that a combination of LLM-based filtering of publicly available web data, and LLM-created synthetic data, enable performance in smaller language models that were typically seen only in much larger models. For example our previous model trained on this data recipe, **phi-2** (2.7B parameters), matched the performance of models 25 times larger trained on regular data. In this report we present a new model, **phi-3-mini** (3.8B parameters), trained for 3.3T tokens on larger and more advanced versions of the datasets used in **phi-2**. With its small size, **phi-3-mini** can easily be inferred locally on a modern phone (see Figure 2), yet it achieves a quality that seems on-par with models such as Mixtral 8x7B [JSR⁺24] and GPT-3.5.

User: Explain why it is surprising that one can build a language model small enough to fit on a phone, yet almost as powerful as ChatGPT. Just use one funny sentence.

phi-3-mini: It's like fitting a supercomputer in a flip phone, but instead of breaking the phone, it just breaks the internet with its tiny, yet mighty, linguistic prowess!

User: Okay now more serious answer, and note that this was achieved solely by changing the training data.

phi-3-mini: The development of a compact language model that rivals the capabilities of ChatGPT, while fitting on a phone, is a testament to the power of data-driven machine learning. By meticulously curating and optimizing the training dataset, researchers can significantly reduce the model's size without compromising its performance. [...]

2 Technical Specifications

The **phi-3-mini** model is a transformer decoder architecture [VSP¹⁷], with default context length $4K$. We also introduce a long context version via LongRope [DZZ^{+24a}] that extends the context length to $128K$, called **phi-3-mini-128K**.

To best benefit the open source community, **phi-3-mini** is built upon a similar block structure as Llama-2 [TLI⁺²³] and uses the same tokenizer with vocabulary size of 32064¹. This means that all packages developed for Llama-2 family of models can be directly adapted to **phi-3-mini**. The model uses 3072 hidden dimension, 32 heads and 32 layers. We trained using bfloat16 for a total of 3.3T tokens. The model is already chat-finetuned, and the chat template is as follows:

```
<|user|>\n Question <|end|>\n <|assistant|>
```

The **phi-3-small** model (7B parameters) leverages the tiktoken tokenizer (for better multilingual tokenization) with a vocabulary size of 100352² and has default context length 8192. It follows the standard decoder architecture of a 7B model class, having 32 heads, 32 layers and a hidden size of 4096. We switched to GEGLU activation and used Maximal Update Parametrization (muP) [YHB⁺²²] to tune hyperparameters on a small proxy model and transfer them to the target 7B model. Those helped ensure better performance and training stability. Also, the model leverages a grouped-query attention, with 4 queries sharing 1 key. To optimize the training and inference speed, we design a novel blocksparse attention module. For each attention head, the blocksparse attention enforces different sparsity patterns over KV cache. This ensures that all tokens are attended to on different heads for the given choice of sparsity. As illustrated in Figure 1, the context is then efficiently divided and conquered among attention heads, with significant KV cache reduction. To achieve actual deployment speed-up from the blocksparse design, we implemented highly efficient, yet flexible kernels for both training and inference. For training, we build a triton kernel based on Flash Attention [DFE⁺²²]. For inference, we implemented a kernel for the prefilling phase and extended the paged attention kernel in vLLM for the decoding phase [KLZ⁺²³]. Lastly, in **phi-3-small** architecture, we alternate dense attention layers and blocksparse attention layers to optimize KV cache savings while maintaining long context retrieval performance. An additional 10% multilingual data was also used for this model.

The **phi-3.5-MoE** adopts an Mixture-of-Experts (MoE) architecture to selectively activate parts of modules on specific inputs to improve the model efficiency. It incorporates MoE layer as its feedforward models, employing the top2 routing among 16 expert networks. Particularly, each expert network is a separate GLU network and the routing module will selectively activate 2 expert networks out of the 16 expert networks for each token, leaving $16 \times 3.8B$ model to have 6.6B activated parameters with 42B

¹We remove BoS tokens and add some additional tokens for chat template.

²We remove unused tokens from the vocabulary.

	Block1	Block2	Block3	Block4	Block5	Block6	Block7	Block8
Head1	Yellow	Gray	Gray	Yellow	Gray	Gray	Blue	Blue
Head2	Gray	Yellow	Gray	Gray	Yellow	Gray	Blue	Blue
Head3	Gray	Gray	Yellow	Gray	Gray	Yellow	Blue	Blue
Head4	Yellow	Gray	Gray	Yellow	Gray	Gray	Blue	Blue

Figure 1: Toy illustration of the blocksparse attention in phi-3-small with 2 local blocks and vertical stride of 3. The table shows the Keys/values a query token in block 8 attended to. Blue=local blocks, orange=remote/vertical blocks, gray=blocks skipped.

total parameters. Additionally, we utilize the SparseMixer approach [LGC23, LDL⁺23] for training the sparse router in the MoE model. For comparison with other Phi series models, **phi-3.5-MoE** uses the same tokenizer as **phi-3-medium** and **phi-3-mini** with vocabulary size of 32064.

Highly capable language model running locally on a cell-phone. Thanks to its small size, **phi-3-mini** can be quantized to 4-bits so that it only occupies $\approx 1.8\text{GB}$ of memory. We tested the quantized model by deploying **phi-3-mini** on iPhone 14 with A16 Bionic chip running natively on-device and fully offline achieving more than 12 tokens per second.

Training Methodology. We follow the sequence of works initiated in “Textbooks Are All You Need” [GZA⁺23], which utilize high quality training data to improve the performance of small language models and deviate from the standard *scaling-laws*. In this work we show that such method allows to reach the level of highly capable models such as GPT-3.5 or Mixtral with only 3.8B total parameters (while Mixtral has 45B total parameters for example). Our training data consists of heavily filtered publicly available web data (according to the “educational level”) from various open internet sources, as well as synthetic LLM-generated data. Pre-training is performed in two disjoint and sequential phases; phase-1 comprises mostly of web sources aimed at teaching the model general knowledge and language understanding. Phase-2 merges even more heavily filtered webdata (a subset used in Phase-1) with some synthetic data that teach the model logical reasoning and various niche skills.

Data Optimal Regime. Unlike prior works that train language models in either “compute optimal regime” [HBM⁺22] or “over-train regime”, we mainly focus on the quality of data for a *given scale*³. We try to calibrate the training data to be closer to the “data optimal” regime for small models. In particular, we filter the publicly available web data to contain the correct level of “knowledge” and keep more web pages that could potentially improve the “reasoning ability” for the model. As an example, the result of a game in premier league in a particular day might be good training data for frontier models, but we need to remove such information to leave more model capacity for “reasoning” for the mini size models. We compare our approach with Llama-2 in Figure 3.

To test our data on larger size of models, we also trained **phi-3-medium**, a model with 14B parameters using the same tokenizer and architecture of **phi-3-mini**, and trained on the same data for slightly more epochs (4.8T tokens total as for **phi-3-small**). The model has 40 heads and 40 layers, with embedding dimension 5120. We observe that some benchmarks improve much less from 7B to 14B than they do from 3.8B to 7B, perhaps indicating that our data mixture needs further work to be in the “data optimal regime” for 14B parameters model.

³Just like for “compute optimal regime”, we use the term “optimal” in an aspirational sense for “data optimal regime”. We are not implying that we actually found the provably “optimal” data mixture for a given scale.

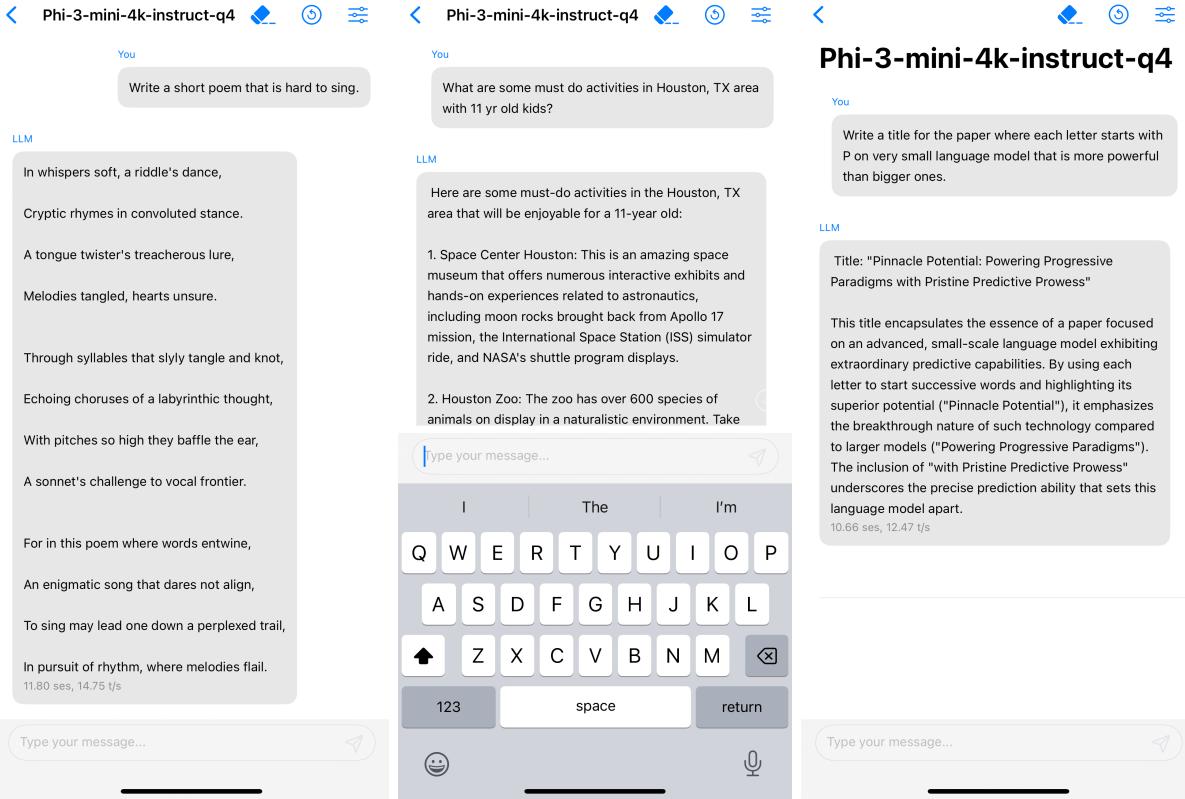


Figure 2: 4-bit quantized **phi-3-mini** running natively on an iPhone with A16 Bionic chip, generating over 12 tokens per second.

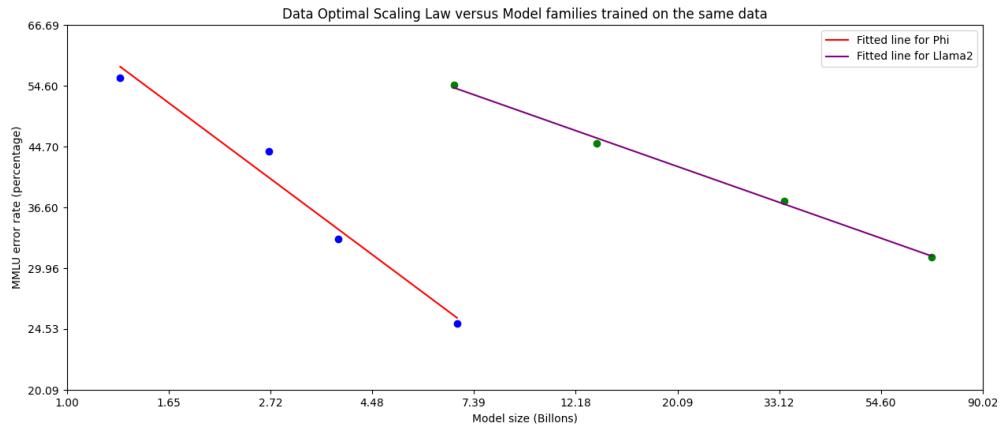


Figure 3: Scaling law close to the “Data Optimal Regime” (from left to right: phi-1.5, phi-2, phi-3-mini, phi-3-small) versus Llama-2 family of models (7B, 13B, 34B, 70B) that were trained on the same fixed data. We plot the log of MMLU error versus the log of model size.

Post-training. Post-training of **phi-3** went through two stages, including supervised finetuning (SFT) and direct preference optimization (DPO). SFT leverages highly curated high-quality data across diverse domains, e.g., math, coding, reasoning, conversation, model identity, and safety. The SFT data mix starts with using English-only examples. DPO data covers chat format data, reasoning, and responsible AI (RAI) efforts. We use DPO to steer the model away from unwanted behavior, by using those outputs as “rejected” responses. Besides improvement in math, coding, reasoning, robustness, and safety, post-training transforms a language model to an AI assistant that users can efficiently and safely interact with.

3 Academic benchmarks

On the next page we report the results for **phi-3** on standard open-source benchmarks measuring the model’s reasoning ability (both common sense reasoning and logical reasoning). We compare to phi-2 [JBA⁺23], Mistral-7b-v0.1 [JSM⁺23], Mixtral-8x7b [JSR⁺24], Gemma 7B [TMH⁺24], Llama-3-instruct-8b [AI23], and GPT-3.5. All the reported numbers are produced with the exact same pipeline to ensure that the numbers are comparable. These numbers might differ from other published numbers due to slightly different choices in the evaluation. As is now standard, we use few-shot prompts to evaluate the models, at temperature 0. The prompts and number of shots are part of a Microsoft internal tool to evaluate language models, and in particular we did no optimization to the pipeline for the **phi-3** models⁴. The number of k -shot examples is listed per-benchmark. An example of a 2-shot prompt is described in Appendix A.

⁴For example, we found that using `##` before the Question can lead to a noticeable improvement to **phi-3-mini**’s results across many benchmarks, but we did not do such changes in the prompts.

	Phi-3-mini 3.8b	Phi-3-small 7b	Phi-3-medium 14b	Phi-2 2.7b	Mistral 7b	Gemma 7b	Llama-3-In 8b	Mixtral 8x7b	GPT-3.5 version 1106
MMLU (5-Shot) HBK⁺21a	68.8	75.7	78.0	56.3	61.7	63.6	66.5	70.5	71.4
HellaSwag (5-Shot) ZHB⁺19	76.7	77.0	82.4	53.6	58.5	49.8	71.1	70.4	78.8
ANLI (7-Shot) NWP⁺20	52.8	58.1	55.8	42.5	47.1	48.7	57.3	55.2	58.1
GSM-8K (8-Shot; CoT) CKB⁺21	82.5	89.6	91.0	61.1	46.4	59.8	77.4	64.7	78.1
MATH (0-Shot; CoT) HKB⁺21b	41.3	34.6	53.1	—	15.0	13.6	28.2	11.1	45.3
MedQA (2-Shot) JPO⁺20	53.8	65.4	69.9	40.9	50.0	49.6	60.5	62.2	63.4
AGIEval (0-Shot) ZCG⁺23	37.5	45.1	50.2	29.8	35.1	42.1	42.0	45.2	48.4
TriviaQA (5-Shot) ICWZ17	64.0	58.1	73.9	45.2	75.2	72.3	67.7	82.2	85.8
Arc-C (10-Shot) CCE⁺18	84.9	90.7	91.6	75.9	78.6	78.3	82.8	87.3	87.4
Arc-E (10-Shot) CCE⁺18	94.6	97.0	97.7	88.5	90.6	91.4	93.4	95.6	96.3
PIQQA (5-Shot) BZGC19	84.2	86.9	87.9	60.2	77.7	78.1	75.7	86.0	86.6
SociQA (5-Shot) BZGC19	76.6	79.2	80.2	68.3	74.6	65.5	73.9	75.9	68.3
BigBench-Hard (3-Shot; CoT) SRR⁺22 SSS⁺22	71.7	79.1	81.4	59.4	57.3	59.6	51.5	69.7	68.32
WinoGrande (5-Shot) SLBBC19	70.8	81.5	81.5	54.7	54.2	55.6	65.0	62.0	68.8
OpenBookQA (10-Shot) MCKS18	83.2	88.0	87.4	73.6	79.8	78.6	82.6	85.8	86.0
BoolQ (2-Shot) CLC⁺19	77.2	84.8	86.5	—	72.2	66.0	80.9	77.6	79.1
CommonSenseQA (10-Shot) THLB19	80.2	80.0	82.8	69.3	72.6	76.2	79.0	78.1	79.6
TruthfulQA (10-Shot; MC2) LHE22	65.0	70.2	75.1	—	53.0	52.1	63.2	60.1	85.8
HumanEval (0-Shot) CTJ⁺21	58.5	61.0	62.2	59.0	28.0	34.1	60.4	37.8	62.2
MBPP (3-Shot) AON⁺21	70.0	71.7	75.2	60.6	50.8	51.5	67.7	60.2	77.8
Average	69.7	73.6	76.7	—	58.9	59.3	67.3	66.8	72.8
GPQA (2-Shot; CoT) RHS⁺23	32.8	34.3	—	—	—	—	—	—	29.0
MT Bench (2 round ave.) ZCS⁺23	8.38	8.70	8.91	—	—	—	—	—	8.35

4 Multilingual and Long Context

To enhance the Phi-3 models with multilingual and long-context capabilities, we developed the versions **phi-3.5-mini** and **phi-3.5-MoE**, which incorporate more multilingual and long-text data during mid-training. Specifically, we employed the long-rope method [DZZ⁺24a](#) and a mixed context window approach to expand the context length limit from 4K to 128K without compromising performance on 4K-context tasks.

Figure 4 compares the performance of **phi-3.5-mini**, **phi-3.5-MoE**, and **phi-3.5-MoE** on MMLU multilingual tasks. **phi-3.5-mini** demonstrates significant improvement over **phi-3.5-mini** in languages

Model	Ctx Size	Python	C++	Rust	Java	TypeScript	Average
gpt-4O-2024-05-13	128k	95	80	85	96	97	90.6
gemini-1.5-flash-latest	1000k	93	79	87	94	97	90
Phi-3.5-MoE	128k	89	74	81	88	95	85
Phi-3.5-Mini	128k	86	67	73	77	82	77
Llama-3.1-8B-Instruct	128k	80	65	73	76	63	71
Mixtral-8x7B-Instruct-v0.1	32k	66	65	64	71	74	68
Mixtral-8x22B-Instruct-v0.1	64k	60	67	74	83	55	67.8

Table 1: Comparison results on RepoQA benchmark.

such as Arabic, Chinese, Russian, Ukrainian, and Vietnamese, with average MMLU-multilingual scores of 55.4 and 47.3, respectively. Due to its larger model capacity, **phi-3.5-MoE** achieves a significantly higher average score of 69.9, outperforming **phi-3.5-mini**.

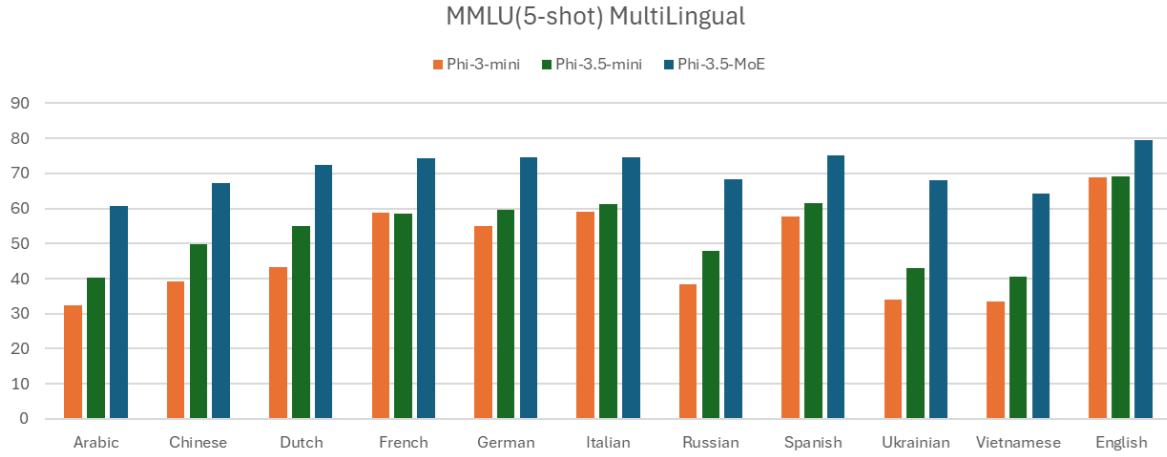


Figure 4: Comparison of **phi-3-mini**, **phi-3.5-mini** and **phi-3.5-MoE** on MMLU-Multilingual tasks

We evaluate the **phi-3.5-mini** and **phi-3.5-MoE** models on two long-context understanding tasks: RULER [HSK⁺24] and RepoQA [LTD⁺24]. As shown in Tables 1 and 2 both **phi-3.5-MoE** and **phi-3.5-mini** outperform other open-source models with larger sizes, such as Llama-3.1-8B, Mixtral-8x7B, and Mixtral-8x22B, on the RepoQA task, and achieve comparable performance to Llama-3.1-8B on the RULER task. However, we observe a significant performance drop when testing the 128K context window on the RULER task. We suspect this is due to the lack of high-quality long-context data in mid-training, an issue we plan to address in the next version of the model release.

In the table 3, we present a detailed evaluation of the **phi-3.5-mini** and **phi-3.5-MoE** models compared with recent SoTA pretrained language models, such as GPT-4o-mini, Gemini-1.5 Flash, and open-source models like Llama-3.1-8B and the Mistral models. The results show that **phi-3.5-mini** achieves performance comparable to much larger models like Mistral-Nemo-12B and Llama-3.1-8B, while **phi-3.5-MoE** significantly outperforms other open-source models, offers performance comparable to Gemini-1.5 Flash, and achieves above 90% of the average performance of GPT-4o-mini across various language benchmarks.

Model	Ctx Size	4k	8k	16k	32k	64k	128k	Average
Llama-3.1-8B-Instruct	128k	95.5	93.8	91.6	87.4	84.7	77.0	88.3
Phi-3.5-MoE	128k	94.8	93.0	93.2	91.6	85.7	64.2	87.1
Phi-3.5-Mini	128k	94.3	91.1	90.7	87.1	78.0	63.6	84.1
Mixtral-8x22B-Instruct-v0.1	64k	95.6	94.9	93.4	90.9	84.7	31.7	81.9
Mixtral-8x7B-Instruct-v0.1	32k	94.9	92.1	92.5	85.9	72.4	44.5	80.4

Table 2: Comparison results on RULER benchmark.

5 Safety

Phi-3-mini was developed in accordance with Microsoft’s responsible AI principles. The overall approach consisted of safety alignment in post-training, red-teaming, automated testing and evaluations across dozens of RAI harm categories. Helpfulness and harmlessness preference datasets [BJN⁺22, JLD⁺23] with modifications inspired by [BSA⁺24] and multiple in-house generated datasets were leveraged to address the RAI harm categories in safety post-training. An independent red team at Microsoft iteratively examined **phi-3-mini** to further identify areas of improvement during the post-training process. Based on their feedback, we curated additional datasets tailored to address their insights, thereby refining the post-training dataset. This process resulted in significant decrease of harmful response rates, as shown in Figure 5.

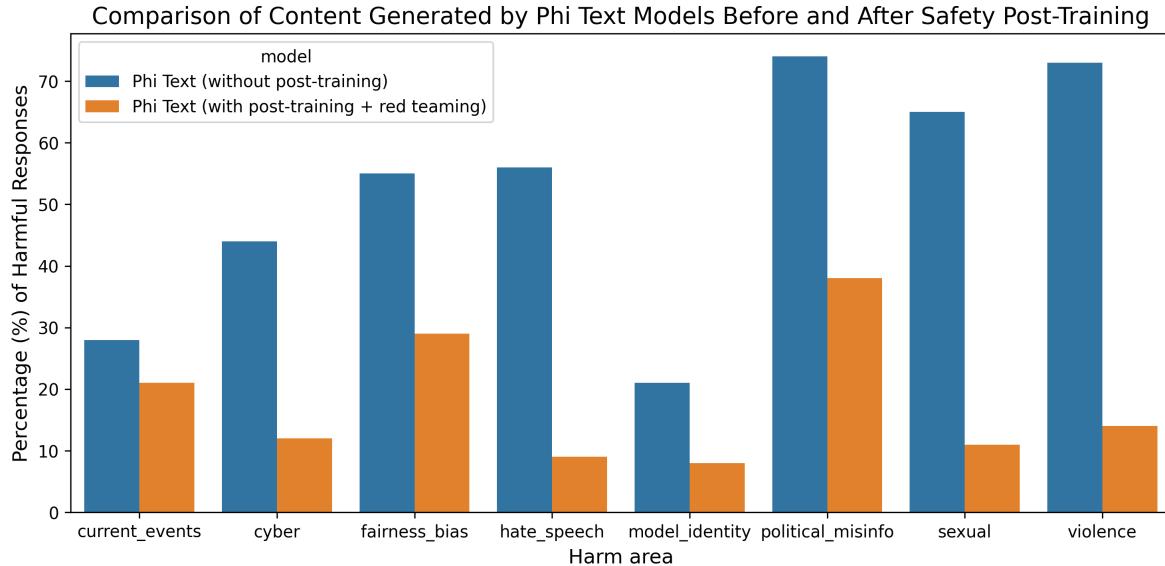


Figure 5: Comparison of harmful response percentages by Microsoft AI Red Team between **phi-3-mini** before and after the safety alignment. Note that the harmful response percentages in this chart are inflated numbers as the red team tried to induce **phi-3-mini** in an adversarial way to generate harmful responses through multi-turn conversations.

The safety alignment of **phi-3-small**, **phi-3-medium** and **phi-3.5-MoE** was conducted by undergoing the same red-teaming process, utilizing identical datasets, and incorporating a slightly larger number of samples. Table 4 shows the results of in-house RAI benchmarks [MHJ⁺23] for **phi-3** models compared to phi-2 [JBA⁺23], Mistral-7b-v0.1 [JSM⁺23], Gemma 7b [TMH⁺24], and Llama-3-instruct-8b [AI23]. This benchmark utilized GPT-4 to simulate multi-turn conversations in five different categories

Category	Benchmark	Phi-3.5-mini 3.8B	Phi-3.5-MoE 16x3.8B	Mistral 7B	Mistral-Nemo 12B	Llama-3.1-In 8B	Gemma-2 9B	Gemini-1.5 Flash	GPT-4o-mini
Popular	Arena Hard	37	37.9	18.1	39.4	25.7	42	55.2	75
	BigBench Hard CoT (0-shot)	69	79.1	33.4	60.2	63.4	63.5	66.7	80.4
MMLU	MMLU (5-shot)	69	78.9	60.3	67.2	68.1	71.3	78.7	77.2
	MMLU-Pro (0-shot, CoT)	47.5	54.3	18	40.7	44	50.1	57.2	62.8
Reasoning	ARC Challenge (10-shot)	84.6	91.0	77.9	84.8	83.1	89.8	92.8	93.5
	BoolQ (2-shot)	78	84.6	80.5	82.5	82.8	85.7	85.8	88.7
	GPQA (0-shot, CoT)	27.2	36.8	15.6	28.6	26.3	29.2	37.5	41.1
	HellaSwag (5-shot)	69.4	83.8	71.6	76.7	73.5	80.9	67.5	87.1
	OpenBookQA (10-shot)	79.2	89.6	78	84.4	84.8	89.6	89	90
	PIQA (5-shot)	81	88.6	73.4	83.5	81.2	83.7	87.5	88.7
	Social IQA (5-shot)	74.7	78.0	73	75.3	71.8	74.7	77.8	82.9
	TruthfulQA (10-shot, MC2)	64	77.5	64.7	68.1	69.2	76.6	76.6	78.2
	WinoGrande (5-shot)	68.5	81.3	58.1	70.4	64.7	74	74.7	76.9
Multilingual	Ml MMLU (5-shot)	55.4	69.9	47.4	58.9	56.2	63.8	77.2	72.9
	MGSM (0-shot CoT)	47.9	58.7	31.8	63.3	56.7	76.4	75.8	81.7
Math	GSM8K (8-shot, CoT)	86.2	88.7	54.4	84.2	82.4	84.9	82.4	91.3
	MATH (0-shot, CoT)	48.5	59.5	19	31.2	47.6	50.9	38	70.2
Long context	Qasper	41.9	40.0	31.4	30.7	37.2	13.9	43.5	39.8
	SQuALITY	24.3	24.1	25.9	25.8	26.2	0	23.5	23.8
Code	HumanEval (0-shot)	61.5	70.7	35.4	63.4	66.5	61	74.4	86.6
	MBPP (3-shot)	68.6	80.8	50.4	68.1	69.4	69.3	77.5	84.1
Average		61.1	69.2	48.5	61.3	61.0	63.3	68.5	74.9

Table 3: Model quality on representative benchmarks

and to evaluate the model responses. Ungroundedness between 0 (fully grounded) and 4 (not grounded) measures if the information in a response is based on a given prompt. In other categories, responses were evaluated in terms of the severity of harmfulness from 0 (no harm) to 7 (extreme harm) and the defect rates (DR- x) were computed as the percentage of samples with the severity score being greater than or equal to x .

6 Weakness

In terms of LLM capabilities, while **phi-3-mini** model achieves similar level of language understanding and reasoning ability as much larger models, it is still fundamentally limited by its size for certain tasks. The model simply does not have the capacity to store too much “factual knowledge”, which can be seen for example with low performance on TriviaQA. However, we believe such weakness can be resolved by augmentation with a search engine. We show an example using the HuggingFace default Chat-UI with

	Phi-3-mini 3.8b	Phi-3-small 7b	Phi-3-medium 14b	Phi-3.5-MoE 16x3.8b	Phi-2 2.7b	Mistral 7b	Gemma 7b	Llama-3-In 8b
Ungroundedness	0.603	0.299	0.213	0.228	1.481	0.935	0.679	0.328
Third Party Harm (DR-1)	0.240	0.253	0.251	0.105	0.240	0.562	0.383	0.373
Harmful Content Continuation (DR-3)	0.007	0.003	0.010	0.005	0.029	0.026	0.013	0.013
Harmful Content Summarization (DR-3)	0.100	0.110	0.112	0.12	0.144	0.223	0.103	0.082
Jailbreak (DR-1)	0.123	0.107	0.111	0.106	0.150	0.156	0.114	0.130

Table 4: Comparison of Microsoft internal multi-turn conversation RAI benchmark results of **phi-3** models and other models. Note that a lower value indicates a better performance for all metrics in the table.

phi-3-mini in Figure 6. Another weakness related to model’s capacity is that we mostly restricted the language to English. Exploring multilingual capabilities for Small Language Models is an important next step, with some initial promising results on **phi-3-small** by including more multilingual data.

Despite our diligent RAI efforts, as with most LLMs, there remains challenges around factual inaccuracies (or hallucinations), reproduction or amplification of biases, inappropriate content generation, and safety issues. The use of carefully curated training data, and targeted post-training, and improvements from red-teaming insights significantly mitigates these issues across all dimensions. However, there is significant work ahead to fully address these challenges, and downstream use of the models should be evaluated for the specific use cases and safety considerations for that context.

7 Phi-3.5-Vision

7.1 Technical Specifications

Architecture The **Phi-3.5-Vision** (4.2B parameters) is a multimodal model designed to process an image/multi-image and a textual prompt as inputs, and subsequently generate textual outputs. This model is composed of two primary components: an image encoder, *i.e.*, CLIP ViT-L/14 [RKH⁺21] and a transformer decoder, *i.e.*, phi-3.5-mini. The visual tokens, once extracted by the image encoder, are then combined with text tokens in an interleaved way (no particular order for image and text tokens). To accommodate high-resolution images and various aspect ratios, a dynamic cropping strategy [DZZ⁺24b] is utilized to split the input image into a 2d array of blocks, where the tokens of the blocks are concatenated to represent the whole image. For multi-image input, we simply concatenated tokens from each images together.

Pre-training The **Phi-3.5-Vision** model undergoes a pre-training phase using a diverse dataset, which consists of a combination of interleaved image-text documents (*e.g.*, [LST⁺24]), image-text pairs from FLD-5B [XWX⁺24], synthetic data derived from Optical Character Recognition (OCR) of PDF files, datasets for chart/table comprehension, and text-only data. The objective of predicting the next token is employed specifically on text tokens, while any loss associated with image tokens is disregarded during this phase. The pre-training process involves a total of $0.5T$ tokens that encompass both visual and text elements. During the pre-training phase, the maximum image resolution is capped at 1344×1344 as the majority of the training images are smaller than this resolution.

Post-training. The **Phi-3.5-Vision** model contains two post-training stages: supervised finetuning (SFT) and direct preference optimization (DPO). For SFT, we leveraged text SFT dataset, public multi-modal instruct tuning datasets along with large-scale multimodal instruct tuning datasets that we built

Suggest an itinerary for a 3-day trip during the 2026 Winter Olympic games

Day 1: Arrival and Surrounding Neighborhoods

Morning:

- Arrive in Pyeongchang, South Korea, and check-in to your accommodations.
- Enjoy a light breakfast at your hotel.

Mid-morning:

- Take a leisurely walk around the town and explore the local shops and cafes.

Lunch:

- Try local specialties such as kimbap (Korean rice rolls) or mandu (Korean dumplings) at a nearby restaurant.

Afternoon:

- Visit Pyeongchang Olympic Cauldron Park, the site of the 2018 Winter Olympics, to take photos and admire the beautiful scenery.
- Rent a bike and explore the surrounding neighborhoods, such as Pyeongchang Village, where you can find souvenir shops, outdoor activities, and various dining options.

Evening:

- Dine at a local restaurant and enjoy traditional Korean cuisine such as samgyeopsal (grilled pork belly) or bulgogi (marinated beef).

Day 2: Olympic Sports and Cultural Activities

Morning:

- Visit the Alpensia Sports Park, where the majority of the 2026 Winter Olympic events will be held.
- See the ski slopes, ice rinks, and indoor arenas, and watch a glimpse of the opening and closing ceremonies.

Mid-morning:

- Participate in a guided tour to learn about the history of the 2026 Winter Olympics and the unique features of the Pyeongchang venues.

Search web ⓘ

Ask anything ⌂

Suggest an itinerary for a 3-day trip during the 2026 Winter Olympic games

Web Search
Completed

Day 1: Arrival and Exploring Milan

- Arrive in Milan on February 6, 2026, the day of the Opening Ceremony.
- Spend the day exploring Milan's city center, including Piazza del Duomo, Galleria Vittorio Emanuele II, and the Sforza Castle.
- Enjoy the vibrant Milanese nightlife, with options like the city's many trendy cafes and restaurants.
- Visit the new Fiera Milano expo center for a unique and inspiring experience.

Day 2: Cortina d'Ampezzo and Outdoor Activities

- Take the train from Milan to Cortina d'Ampezzo, arriving in the afternoon.
- Check into your hotel and spend the evening relaxing and enjoying the local cuisine, such as risotto and polenta, at traditional pizzerias or trattorias.
- On Day 2, take part in various outdoor activities, such as skiing, snowboarding, or tubing. Many of these activities are available at nearby resorts like Madonna di Campiglio, Kronplatz, or Passo San Pellegrino.
- In the evening, explore Cortina d'Ampezzo's charming streets, visit the historic St. Felix Castle, and enjoy the beautiful mountain scenery.

Day 3: Paralympic Games and Departure

- On Day 3, attend the Paralympic Games, held in Turin on March 6, 2026.
- Take the train from Cortina d'Ampezzo to Turin, arriving in time for the Paralympic events.
- Explore Turin's attractions, such as the Mole Antonelliana, the Egyptian Museum, and Piazza Vittorio Veneto.
- Return to Milan in the evening for your departure, concluding your 3-day itinerary during the Milano Cortina 2026 Winter Olympic Games.

Throughout your trip, make sure to book your accommodations, transportation, and tickets in advance, as these events are highly popular and may sell out quickly. Enjoy your time in Italy and the exciting Winter Olympic Games!

Sources: ⓘ olympics.com ⓘ olympics.com ⓘ milanocortina2026.olympics.com ⓘ milanwintertravel.com

Search web ⓘ

Ask anything ⌂

Figure 6: Left: **phi-3-mini**'s completion without search. Right: **phi-3-mini**'s completion with search, using the default HuggingFace Chat-UI search ability. For reference, the 2026 Winter Olympic Games are scheduled to be held in Milano and Cortina in Italy, while the 2022 and 2018 Winter Olympic Games were held in Beijing, China and PyeongChang, Korea, respectively. Without the search results, the response is incorrect, while with the web search, not only does the response become accurate, but also gets more specific with suggestions.

	SM	MED	LG	CUP	POT
Espresso	1.30	1.60	2.00		
Americana	1.40	1.75	2.10		
Cappuccino	2.25	2.40	3.25		
Latte	2.35	3.00	3.25		
Mocha	2.35	3.00	3.50		
Chai	2.25	2.40	3.25		
Chai Bomb	2.75	3.25	3.75		
House Café	1.50	1.75	2.00		
Café au Lait	1.75	2.50	3.25		
Teas				1.50	4.00
GREEN BLACK WHITE OOLONG HERBAL					
*ADD .50 FOR HONEY **ADD \$1.00 FOR ICE**					

extract all displayed texts and keep original displayed format. For the left section, please use markdown table with column "coffee", "small", "median" and "large". Instead, for the right section, please use JSON format.

- Certainly! Below is the markdown table converted from the image:

Coffee	Small	Medium	Large
Espresso	1.30	1.60	2.00
Americana	1.40	1.75	2.10
Cappuccino	2.25	2.40	3.25
Latte	2.35	3.00	3.25
Mocha	2.35	3.00	3.50
Chai	2.25	2.40	3.25
Chai Bomb	2.75	3.25	3.75
House Café	1.50	1.75	2.00
Café au Lait	1.75	2.50	3.25

And here is the JSON representation of the teas offered:

```
{
  "Teas": {
    "CUP": {
      "GREEN": "$1.50",
      "BLACK": "$1.50",
      "WHITE": "$1.50",
      "OOLONG": "$1.50",
      "HERBAL": "$1.50"
    },
    "POT": {
      "GREEN": "$4.00",
      "BLACK": "$4.00",
      "WHITE": "$4.00",
      "OOLONG": "$4.00",
      "HERBAL": "$4.00"
    }
  },
  "*ADD .50 FOR HONEY*",
  "**ADD $1.00 FOR ICE**"
}
```

What is the price of a Cappuccino with large size?

- The price for a Cappuccino in large size is \$3.25.

How much does it cost to add ice to a tea?

- It costs an additional \$1.00 to add ice to any tea.

If someone wants to buy a pot of tea, how much would it cost?

- A pot of tea would cost \$4.00.

Figure 7: The demo case shows Phi-3.5-Vision’s capability in natural image understanding and reasoning.

ourselves, covering diverse domains and tasks such as general natural image understanding, chart/table-diagram understanding/reasoning, PowerPoint understanding, multi-image comparison, video summarization and model safety. The multimodal SFT data has about a total of 33B tokens. For DPO we mainly use a text DPO dataset and a relatively smaller-scale multimodal DPO dataset. For these two stages, we jointly train multimodal tasks and text-only tasks so that the model can achieve multi-modal reasoning while maintaining language capabilities as much as possible.

7.2 Academic benchmarks

7.2.1 Single-image Benchmarks

We report in Table 5 the evaluation results of Phi-3.5-Vision on nine open-source academic benchmarks. These benchmarks evaluate reasoning and perceptual capabilities on visual and text inputs and can be grouped in three categories: Science, Charts, and Generic knowledge. We compare Phi-3.5-Vision with the following baselines: MM1-3B-Chat [MGF⁺²⁴], MM1-7B-Chat [MGF⁺²⁴], Llava-1.6 Vicuna 7B [LLL23], Llava-1.6 Llama3-8B [LLL⁺²⁴], Qwen-VL-Chat [BBY⁺²³], Claude 3 Haiku [Ant24], Gemini 1.0 Pro V [TAB⁺²³], and GPT-4O. Our performance quality assessment setup used the same evaluation pipeline for all the baselines to ensure a fair comparison, with the exception of MM1-3B-Chat. We just copied and pasted their published numbers since the model is not publicly available.

Our evaluation setup aimed to mimic scenarios where regular users interact with a multi-modal model, i.e., users who are not experts in prompt engineering or know special techniques that can improve performance. For this reason, we adopted the evaluation setting used in Llava-1.5 [LLL23]. In this

setup, the prompts include instructions to select a single letter corresponding to an answer from a list of given options, or answer with a single word or phrase. In our prompts, we did not use specific tokens for multiple-choice questions. Moreover, we did not scale or pre-process any image in our benchmarking system. We placed the images as the first item in the prompts, except on the MMMU dataset where the prompts interleave the images anywhere in the question or the answers. Lastly, our evaluation setup only considered a 0-shot format. Because of these evaluation parameters, our reported numbers can differ from the published numbers of the considered baselines. As we can see, our Phi-3.5-Vision achieves super competitive results on all benchmarks and outperform other competitor models on most benchmarks while being smaller.

7.2.2 Multi-image Benchmarks

We report in Table 6 the evaluation results of Phi-3.5-Vision on one latest academic multi-image benchmark and one video benchmark. These benchmarks evaluate perceptual capabilities on multiple image/frames and text covering a wide range of general scenarios (e.g., Art and Style recognition, Forensic detection, and video understanding). We compare Phi-3.5-Vision with the following baseline methods: Llava Interleave-Qwen 7B [LZZ⁺²⁴], InternVL2 4B and 8B [CWT⁺²⁴], Gemini 1.5 Flash [TAB⁺²³], GPT-4o-mini, Claude 3.5 Sonnet [Ant24], Gemini 1.5 Pro [TAB⁺²³], and GPT-4O. Line in the single-frame evaluation case, our performance quality assessment setup used the same evaluation pipeline for all the baselines to ensure a fair comparison.

Our evaluation setup for multi-image also followed the Llava setup where prompts include instructions to select a single letter corresponding to an answer from a list of given options, or answer with a single word or phrase. Moreover, we did not use specific tokens for multiple-choice questions and we did not scale or pre-process any image in our benchmarking system. For most of the benchmarks, we placed the images as the first item in the prompts.

The evaluation pipelines for BLINK and VideoMME benchmarks differ from those published. In the case of BLINK, we do not use ChatGPT as the final answer selection mechanism. Instead, we instruct the evaluated model to select one answer directly from the given choices. The reason is that in this manner we ensure that the mistakes or successes come solely by the evaluated model. For the VideoMME benchmark, we extracted 16 frames from the video by sampling frames at a given rate that ensures a uniform time coverage of the entire video. We used 16 frames since this is the maximum number of images a prompt can contain for Azure OpenAI models. Unlike the proposed evaluation in VideoMME that uses the maximum number of frames a model can accept, we always pass the same amount of frames across all the considered model baselines. In this way we ensure the evaluations are fair since all the models receive the exact same input information (i.e., the prompt and set of images). As shown in Table 6, our Phi-3.5-Vision performs very competitively or outperforms baseline models under the similar model size in multi-image understanding scenarios as well.

7.3 Safety

To ensure the integration of **Phi-3.5-Vision** aligns with Microsoft’s Responsible AI (RAI) principles, we involved safety post-training in both Supervised Fine-Tuning (SFT) stage and Direct Preference Optimization (DPO) stage. In creating the safety training datasets, we utilized not only the text-only RAI datasets, but also a variety of in-house Multi-Modal (MM) RAI datasets that cover various harm categories identified in both public and internal MM RAI benchmarks. For the purpose of RAI evaluation, we performed a rigorous quantitative assessment on both public and internal benchmarks, this was done in conjunction with a human evaluation conducted by Microsoft’s internal red team.

	Phi-3.5-Vision 4.2b	MM1-3B-Chat 3.6b [MGF ⁺²⁴]	MM1-7B-Chat 7.6b [MGF ⁺²⁴]	LLaVA-1.6 Vicuna-7b [LLL ⁺²³]	LLama-Next LLama-3-8b [LLL ⁺²⁴]	Qwen-VL-Chat 9.6b [BBY ⁺²³]	Claude 3 haiku Ant[24]	Gemini 1.0 Pro V TAB ⁺²³	GPT-4O 2024-05-13
MMMU (val) [YNZ ⁺²³]	43.0	33.9	37.0	34.2	36.4	39.0	40.7	42.0	61.8
ScienceQA (test) [LMX ⁺²²]	91.3	69.4	72.6	70.6	73.7	67.2	72.0	79.7	88.5
MathVista (testmini) [LBX ⁺²⁴]	43.9	32.0	35.9	31.5	34.8	29.4	33.2	35.0	54.4
Inter-GPS (test) [LGJ ⁺²¹]	36.3	-	-	20.5	24.6	22.3	32.1	28.6	46.9
MMBench (dev-en) [LDZ ⁺²⁴]	81.9	75.9	79.0	76.3	79.4	75.8	62.4	80.0	88.4
POPE (test) [LDZ ⁺²³]	86.1	87.4	86.6	87.2	87.0	82.6	74.4	84.2	87.0
AI2D (test) [KSK ⁺¹⁶]	78.1	-	-	63.1	66.9	59.8	60.3	62.8	82.8
ChartQA (test) [MLT ⁺²²]	81.8	-	-	55.0	65.8	50.9	59.3	58.0	64.0
TextVQA (test) [SNS ⁺¹⁹]	72.0	71.9	72.8	64.6	55.7	59.4	62.7	64.7	75.6

Table 5: Comparison results on public MLLM benchmarks. All the reported numbers are produced with the exact same pipeline to ensure that the numbers are comparable except for MM1-3B-Chat [MGF⁺²⁴] and MM1-7B-Chat [MGF⁺²⁴], which are not publicly available. We adopted the evaluation setting used in Llava-1.5 [LLL⁺²³], without any specific prompt or pre-processing image for all results. These numbers might differ from other published numbers due to slightly different prompts.

In Table 7, we present the evaluation outcomes of Phi-3.5-Vision on three MM RAI benchmarks: one internal and two public benchmarks (specifically, RTVLM [LLY⁺²⁴] and VLGuard [ZBY⁺²⁴]). We juxtapose these results with those of other open-source models such as Llava-1.5 [LLL⁺²³], Llava-1.6 [LLL⁺²⁴], Qwen-VL-Chat [BBY⁺²³], and GPT4-V [Ope23]. The results clearly indicate that safety post-training notably enhances the RAI performance of Phi-3.5-Vision across all RAI benchmarks. In Figure 8, we further breakdown the performance across different RAI categories of the VLGuard and Internal benchmarks, demonstrating that safety post-training can aid Phi-3.5-Vision in improving RAI performance in nearly all categories.

7.4 Weakness

Regarding the multi-modal LLM capabilities of our Phi-3.5-Vision, it performs admirably across various fields. However, we have identified certain limitations, particularly with questions necessitating high-level reasoning abilities. Additionally, the model has been observed to occasionally generate ungrounded outputs, making it potentially unreliable in sensitive areas, such as finance. To mitigate these issues, we will incorporate more reasoning-focused and hallucination-related DPO data into post-training in the

	Phi-3.5-Vision 4.2b	Llava-interleave Qwen 7b [ZZZ ⁺²⁴]	InternVL2 4b [CWT ⁺²⁴]	InternVL2 8b [CWT ⁺²⁴]	Gemini 1.5 Flash [TAB ⁺²³]	GPT4O mini 2024-07-18	Claude 3.5 Sonnet [Ant ²⁴]	Gemini 1.5 Pro [TAB ⁺²³]	GPT-4O 2024-05-13
BLINK (val) [FHL ⁺²⁴]	57.0	53.1	45.9	45.4	45.8	51.9	56.5	61.0	63.2
VideoMME (test) [FDL ⁺²⁴]	50.8	50.2	49.9	52.6	62.3	61.2	55.9	62.6	68.4

Table 6: Comparison results on public multi-image/video MLLM benchmarks. All the reported numbers are produced with the exact same pipeline to ensure that the numbers are comparable.

	Phi-3.5-Vision 3.8b+0.3b	Phi-3.5-Vision w/o safety 3.8b+0.3b	Llava-1.6 Vicuna 7b+0.3b	Qwen-VL-Chat 7.7b+1.9b	GPT4-V N/A
Internal (private)	8.16	7.06	5.44	7.27	8.55
RTVLM (public)	5.44	3.56	3.86	4.78	6.81
VLGUARD (public)	9.10	4.66	5.62	8.33	8.90

Table 7: Comparison results on public and private multi-modal RAI benchmarks. Note that all metrics in the table are [0,10] and a higher value indicates a better performance.

future.

From a responsible AI standpoint, whilst safety post-training has made significant strides, our Phi-3.5-Vision occasionally fails to refrain from answering harmful or sensitive inquiries. Examples of such occasions include deciphering particular types of captcha and describing scam images containing disinformation or hallucination. We find that this issue partly arises from the capabilities, such as OCR, acquired during the training process with normal instruct tuning datasets, which can be regarded as the trade-off between helpfulness and harmlessness. Moving forward, we need to further explore this area to achieve a better balance.

References

- [AI23] Meta AI. Introducing meta llama 3: The most capable openly available llm to date, 2023.
- [Ant24] AI Anthropic. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 2024.
- [AON⁺²¹] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.
- [BBY⁺²³] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.



Figure 8: Comparison of categorized RAI performance of Phi-3.5-Vision with and without the safety post-training on the VLGGuard (left) and Internal (right) benchmark, respectively. It clearly indicates that safety post-training can enhance the RAI performance across nearly all the RAI categories.

- [BJN⁺22] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.
- [BSA⁺24] Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions, 2024.
- [BZGC19] Yonatan Bisk, Rowan Zellers, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. *arXiv preprint arXiv:1911.11641*, 2019.
- [CCE⁺18] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018.
- [CKB⁺21] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- [CLC⁺19] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 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, 2019.
- [CTJ⁺21] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray,

Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021.

- [CWT⁺24] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.
- [DFE⁺22] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.
- [DZZ⁺24a] Yiran Ding, Li Lyra Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens, 2024.
- [DZZ⁺24b] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, et al. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *arXiv preprint arXiv:2404.06512*, 2024.
- [FDL⁺24] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.
- [FHL⁺24] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024.
- [GZA⁺23] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Gustavo de Rosa Piero Kauffmann, Olli Saarikivia, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.
- [HBK⁺21a] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset, 2021.
- [HBK⁺21b] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021.

- [HBM⁺22] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Eliza Rutherford Trevor Cai, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.
- [HSK⁺24] Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What’s the real context size of your long-context language models?, 2024.
- [JBA⁺23] Mojtaba Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, Suriya Gunasekar, Piero Kauffmann, Yin Tat Lee, Yuanzhi Li, Anh Nguyen, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Michael Santacroce, Harkirat Singh Behl, Adam Tumann Kalai, Xin Wang, Rachel Ward, Philipp Witte, Cyril Zhang, and Yi Zhang. Phi-2: The surprising power of small language models. *Microsoft Research Blog*, 2023.
- [JCWZ17] Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension, 2017.
- [JLD⁺23] Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Chi Zhang, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset, 2023.
- [JPO⁺20] Di Jin, Eileen Pan, Nassim Oufattolle, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams, 2020.
- [JSM⁺23] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.
- [JSR⁺24] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024.
- [KLZ⁺23] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.
- [KMH⁺20] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

- [KSK⁺16] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images, 2016.
- [LBE⁺23] Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*, 2023.
- [LBX⁺24] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts, 2024.
- [LDL⁺23] Liyuan Liu, Chengyu Dong, Xiaodong Liu, Bin Yu, and Jianfeng Gao. Bridging discrete and backpropagation: Straight-through and beyond. *arXiv:2304.08612*, 2023.
- [LDZ⁺23] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models, 2023.
- [LDZ⁺24] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024.
- [LGC23] Liyuan Liu, Jianfeng Gao, and Weizhu Chen. Sparse backpropagation for moe training. *arXiv:2310.00811*, 2023.
- [LGJ⁺21] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning, 2021.
- [LHE22] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022.
- [LLL⁺24] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.
- [LLLL23] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.
- [LLY⁺24] Mukai Li, Lei Li, Yuwei Yin, Masood Ahmed, Zhenguang Liu, and Qi Liu. Red teaming visual language models. *arXiv preprint arXiv:2401.12915*, 2024.
- [LMX⁺22] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.
- [LST⁺24] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36, 2024.
- [LTD⁺24] Jiawei Liu, Jia Le Tian, Vijay Daita, Yuxiang Wei, Yifeng Ding, Yuhan Katherine Wang, Jun Yang, and Lingming Zhang. Repoqa: Evaluating long context code understanding, 2024.

- [LZZ⁺24] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.
- [MCKS18] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering, 2018.
- [MGF⁺24] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Ankur Jain, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Guoli Yin, Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch, Alexander Toshev, and Yafei Yang. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*, 2024.
- [MHJ⁺23] Ahmed Magooda, Alec Helyar, Kyle Jackson, David Sullivan, Chad Atalla, Emily Sheng, Dan Vann, Richard Edgar, Hamid Palangi, Roman Lutz, Hongliang Kong, Vincent Yun, Eslam Kamal, Federico Zarfati, Hanna Wallach, Sarah Bird, and Mei Chen. A framework for automated measurement of responsible ai harms in generative ai applications, 2023.
- [MLT⁺22] Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics.
- [MRB⁺23] Niklas Muennighoff, Alexander M Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. *arXiv preprint arXiv:2305.16264*, 2023.
- [NWD⁺20] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding, 2020.
- [Ope23] OpenAI. Gpt-4v(ision) system card, 2023. https://cdn.openai.com/papers/GPTV_System_Card.pdf.
- [RHS⁺23] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark, 2023.
- [RKH⁺21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [RWC⁺19] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [SLBBC19] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *arXiv preprint arXiv:1907.10641*, 2019.
- [SNS⁺19] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read, 2019.

- [SRR⁺22] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.
- [SSS⁺22] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them, 2022.
- [TAB⁺23] Gemini Team, Rohan Anil, Sébastien Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [THLB19] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge, 2019.
- [TLI⁺23] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [TMH⁺24] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology, 2024.
- [VSP⁺17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
- [WXW⁺24] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. 2024.
- [YHB⁺22] Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer. 2022.
- [YNZ⁺23] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2023.
- [ZBY⁺24] Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Timothy Hospedales. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. *arXiv preprint arXiv:2402.02207*, 2024.
- [ZCG⁺23] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models, 2023.

[ZCS⁺23] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zuoquan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

[ZHB⁺19] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, 2019.

A Example prompt for benchmarks

Question:
Solve for x : $(-\frac{1}{3})(-4 - 3x) = \frac{1}{2}$
Options:
A. $-\frac{5}{6}$
B. $\frac{7}{6}$
C. $\frac{5}{3}$
D. $\frac{1}{6}$
Answer: A
Question:
Which of the following is the body cavity that contains the pituitary gland?
Options:
A. Abdominal
B. Cranial
C. Pleural
D. Spinal
Answer: B
Question:
Where was the most famous site of the mystery cults in Greece?
Options:
A. Ephesus
B. Corinth
C. Athens
D. Eleusis
Answer:

B Authors (alphabetical)

Marah Abdin	Xin Jin	Adil Salim
Jyoti Aneja	Nikos Karampatziakis	Michael Santacroce
Hany Awadalla	Piero Kauffmann	Shital Shah
Ahmed Awadallah	Mahoud Khademi	Ning Shang
Ammar Ahmad Awan	Dongwoo Kim	Hiteshi Sharma
Nguyen Bach	Young Jin Kim	Yelong Shen
Amit Bahree	Lev Kurilenko	Swadheen Shukla
Arash Bakhtiari	James R. Lee	Xia Song
Jianmin Bao	Yin Tat Lee	Masahiro Tanaka
Harkirat Behl	Yuanzhi Li	Andrea Tupini
Alon Benhaim	Yunsheng Li	Praneetha Vaddamanu
Misha Bilenko	Chen Liang	Chunyu Wang
Johan Bjorck	Lars Liden	Guanhua Wang
Sébastien Bubeck	Xihui Lin	Lijuan Wang
Martin Cai	Zeqi Lin	Shuohang Wang
Qin Cai	Ce Liu	Xin Wang
Vishrav Chaudhary	Liyuan Liu	Yu Wang
Dong Chen	Mengchen Liu	Rachel Ward
Dongdong Chen	Weishung Liu	Wen Wen
Weizhu Chen	Xiaodong Liu	Philipp Witte
Yen-Chun Chen	Chong Luo	Haiping Wu
Yi-Ling Chen	Piyush Madan	Xiaoxia Wu
Hao Cheng	Ali Mahmoudzadeh	Michael Wyatt
Parul Chopra	David Majercak	Bin Xiao
Xiyang Dai	Matt Mazzola	Can Xu
Matthew Dixon	Caio César Teodoro Mendes	Jiahang Xu
Ronen Eldan	Arindam Mitra	Weijian Xu
Victor Fragoso	Hardik Modi	Jilong Xue
Jianfeng Gao	Anh Nguyen	Sonali Yadav
Mei Gao	Brandon Norick	Fan Yang
Min Gao	Barun Patra	Jianwei Yang
Amit Garg	Daniel Perez-Becker	Yifan Yang
Allie Del Giorno	Thomas Portet	Ziyi Yang
Abhishek Goswami	Reid Pryzant	Donghan Yu
Suriya Gunasekar	Heyang Qin	Lu Yuan
Emman Haider	Marko Radmilac	Chenruidong Zhang
Junheng Hao	Liliang Ren	Cyril Zhang
Russell J. Hewett	Gustavo de Rosa	Jianwen Zhang
Wenxiang Hu	Corby Rosset	Li Lyna Zhang
Jamie Huynh	Sambudha Roy	Yi Zhang
Dan Iter	Olatunji Ruwase	Yue Zhang
Sam Ade Jacobs	Olli Saarikivi	Yunan Zhang
Mojan Javaheripi	Amin Saied	Xiren Zhou

C Acknowledgements

We would like to thank Zhuohan Li, Simon Mo from UC Berkeley and Kaichao You from Tsinghua University for sharing their insights on the vLLM kernel.