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. The innovation lies entirely in our dataset for training, a scaled-up version of the one used for **phi-2**, composed of heavily filtered web data and synthetic data. The model is also further aligned for robustness, safety, and chat format. We also provide some initial parameter-scaling results with a 7B and 14B models trained for 4.8T tokens, called **phi-3-small** and **phi-3-medium**, both significantly more capable than **phi-3-mini** (e.g., respectively 75% and 78% on MMLU, and 8.7 and 8.9 on MT-bench).

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 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 inferenced locally on a modern phone (see Figure 1), 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 [VSP+17], with default context length $4K$. We also introduce a long context version via LongRope [DZZ+24] 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 32064[1]. 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 100352 and has default context length $8K$. It follows the standard decoder architecture of a 7B model class, having 32 layers and a hidden size of 4096. To minimize KV cache footprint, the model also leverages a grouped-query attention, with 4 queries sharing 1 key. Moreover **phi-3-small** uses alternative layers of dense attention and a novel blocksparse attention to further optimize on KV cache savings while maintaining long context retrieval performance. An additional 10% multilingual data was also used for this model.

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 ≈ 1.8GB 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 of consists of heavily filtered 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*.[2] We try to calibrate the training data to be closer to the "data optimal" regime for small models. In particular, we filter the 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 2.

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

[2] Just 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.

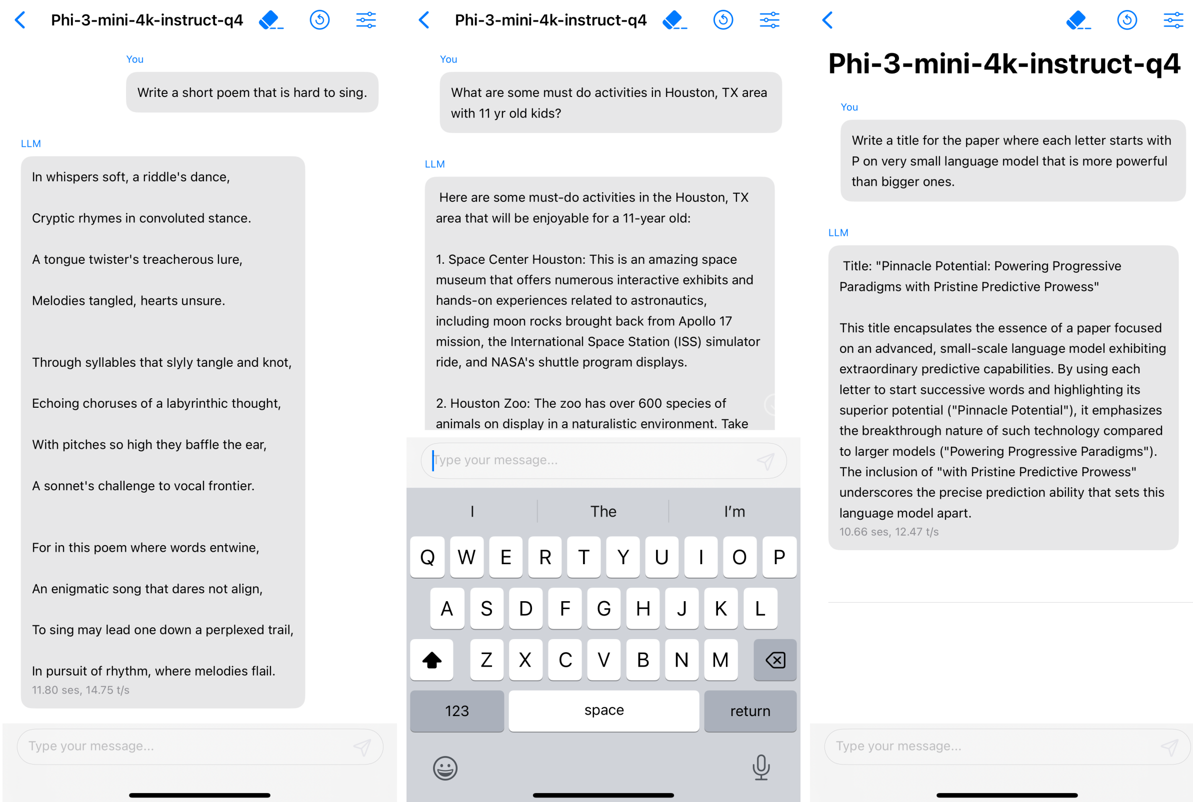

Figure 1: 4-bit quantized **phi-3-mini** running natively on an iPhone with A16 Bionic chip, generating over 12 tokens per second.

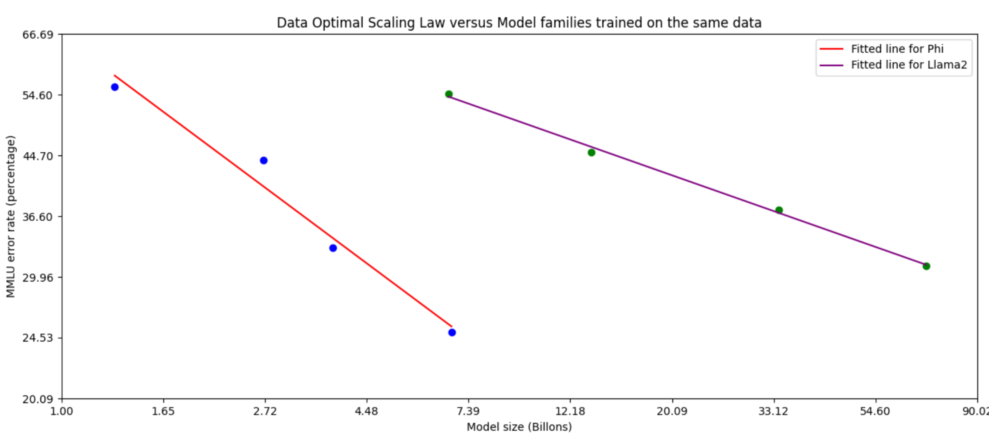

Figure 2: Scaling law close to the "Data Optimal Regime" (from left to right: phi-1.5, phi-2, phi-3-mini, phi-3-small) versus Llama-2 family of models (7B, 13B, 34B, 70B) that were trained on the same fixed data. We plot the log of MMLU error versus the log of model size.

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. We are still actively investigating some of those benchmarks (including a regression on HumanEval), hence the numbers for **phi-3-medium** should be considered as a "preview".

Post-training. Our models went through post-training with both supervised instruction fine-tuning, and preference tuning with DPO. We have worked on generating and curating various instruction and preference data. This has improved the model chat capabilities, robustness, as well as its safety.

3 Academic benchmarks

On the next page we report the results for **phi-3-mini** 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** models.[3] The number of k–shot examples is listed per-benchmark. An example of a 2-shot prompt is described in Appendix A.

[3]For 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 (preview)	Phi-2 2.7b	Mistral 7b	Gemma 7b	Llama-3-In 8b	Mixtral 8x7b	GPT-3.5 version 1106
MMLU (5-Shot) [HBK+21]	68.8	75.3	78.2	56.3	61.7	63.6	66.0	68.4	71.4
HellaSwag (5-Shot) [ZHB+19]	76.7	78.7	83.0	53.6	58.5	49.8	69.5	70.4	78.8
ANLI (7-Shot) [NWD+20]	52.8	55.0	58.7	42.5	47.1	48.7	54.8	55.2	58.1
GSM-8K (0-Shot; CoT) [CKB+21]	82.5	88.9	90.3	61.1	46.4	59.8	77.4	64.7	78.1
MedQA (2-Shot) [JPO+20]	53.8	58.2	69.4	40.9	49.6	50.0	58.9	62.2	63.4
AGIEval (0-Shot) [ZCG+23]	37.5	45.0	48.4	29.8	35.1	42.1	42.0	45.2	48.4
TriviaQA (5-Shot) [JCWZ17]	64.0	59.1	75.6	45.2	72.3	75.2	73.6	82.2	85.8
Arc-C (10-Shot) [CCE+18]	84.9	90.7	91.0	75.9	78.6	78.3	80.5	87.3	87.4
Arc-E (10-Shot) [CCE+18]	94.6	97.1	97.8	88.5	90.6	91.4	92.3	95.6	96.3
PIQA (5-Shot) [BZGC19]	84.2	87.8	87.7	60.2	77.7	78.1	77.1	86.0	86.6
SociQA (5-Shot) [BZGC19]	76.6	79.0	80.2	68.3	74.6	65.5	73.2	75.9	68.3
BigBench-Hard (0-Shot) [SRR+22, SSS+22]	71.7	75.0	81.3	59.4	57.3	59.6	68.9	69.7	68.32
WinoGrande (5-Shot) [SLBBC19]	70.8	82.5	81.4	54.7	54.2	55.6	58.0	62.0	68.8
OpenBookQA (10-Shot) [MCKS18]	83.2	88.4	87.2	73.6	79.8	78.6	81.6	85.8	86.0
BoolQ (0-Shot) [CLC+19]	77.2	82.9	86.6	–	72.2	66.0	78.3	77.6	79.1
CommonSenseQA (10-Shot) [THLB19]	80.2	80.3	82.6	69.3	72.6	76.2	73.6	78.1	79.6
TruthfulQA (10-Shot) [LHE22]	65.0	68.7	75.7	–	52.1	53.0	62.0	60.1	85.8
HumanEval (0-Shot) [CTJ+21]	58.5	59.1	55.5	59.0	28.0	34.1	38.4	37.8	62.2
MBPP (3-Shot) [AON+21]	70.0	71.4	74.5	60.6	50.8	51.5	65.3	60.2	77.8
Average	71.2	74.9	78.2	–	61.0	62.0	68.0	69.9	75.3
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 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,

	Phi-3-Mini-4k 3.8b	Phi-3-Mini-128k 3.8b	Phi-2 2.7b	Mistral 7b	Gemma 7b	Llama-3-In 8b
Ungroundedness	0.603	0.637	1.481	0.935	0.679	0.328
Intellectual Property (DR-1)	23.95%	21.50%	24.00%	56.20%	38.33%	37.30%
Harmful Content Continuation (DR-3)	0.75%	1.08%	2.93%	2.58%	1.28%	1.30%
Harmful Content Summarization (DR-3)	10.00%	10.20%	14.35%	22.33%	10.33%	8.20%
Jailbreak (DR-1)	12.29%	12.57%	15.00%	15.57%	11.43%	13.00%

Table 1: Comparison of Microsoft internal multi-turn conversation RAI benchmark results of **phi-3-mini** and other models. Note that a lower value indicates a better performance for all metrics in the table.

as shown in Figure 3.

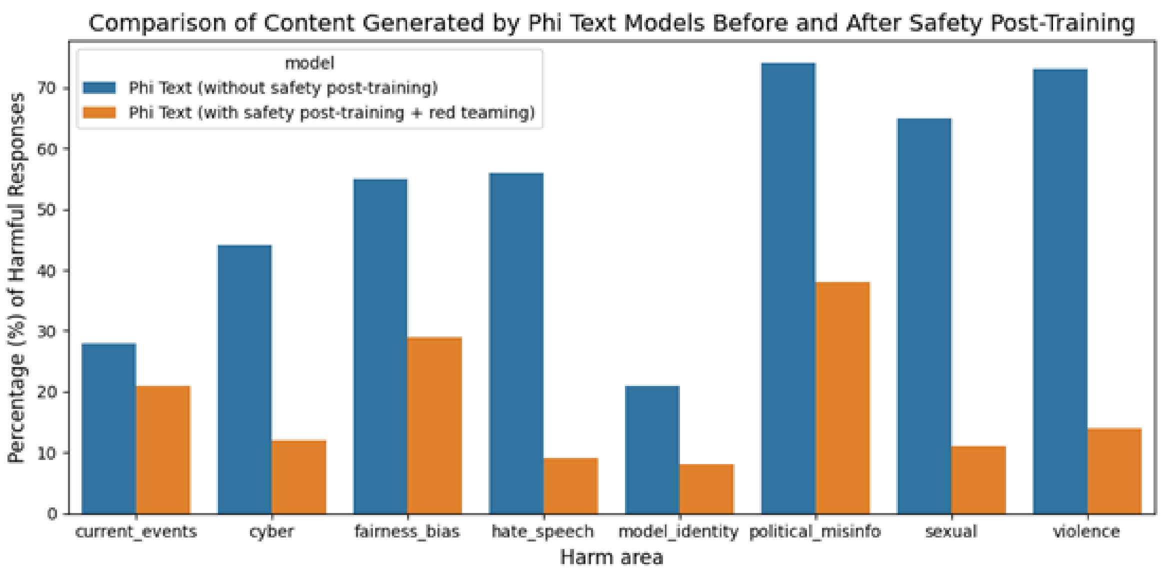

Figure 3: Comparison of harmful response percentages by Microsoft AI Red Team between **phi-3-mini** before and after the safety alignment. Note that the harmful response percentages in this chart are inflated numbers as the red team tried to induce **phi-3-mini** in an adversarial way to generate harmful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

Table 1 shows the results of in-house RAI benchmarks for **phi-3-mini-4k** and **phi-3-mini-128k** compared to phi-2 [JBA+23], Mistral-7b-v0.1 [JSM+23], Gemma 7b [TMH+24]. This benchmark utilized GPT-4 to simulate multi-turn conversations in five different categories 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.

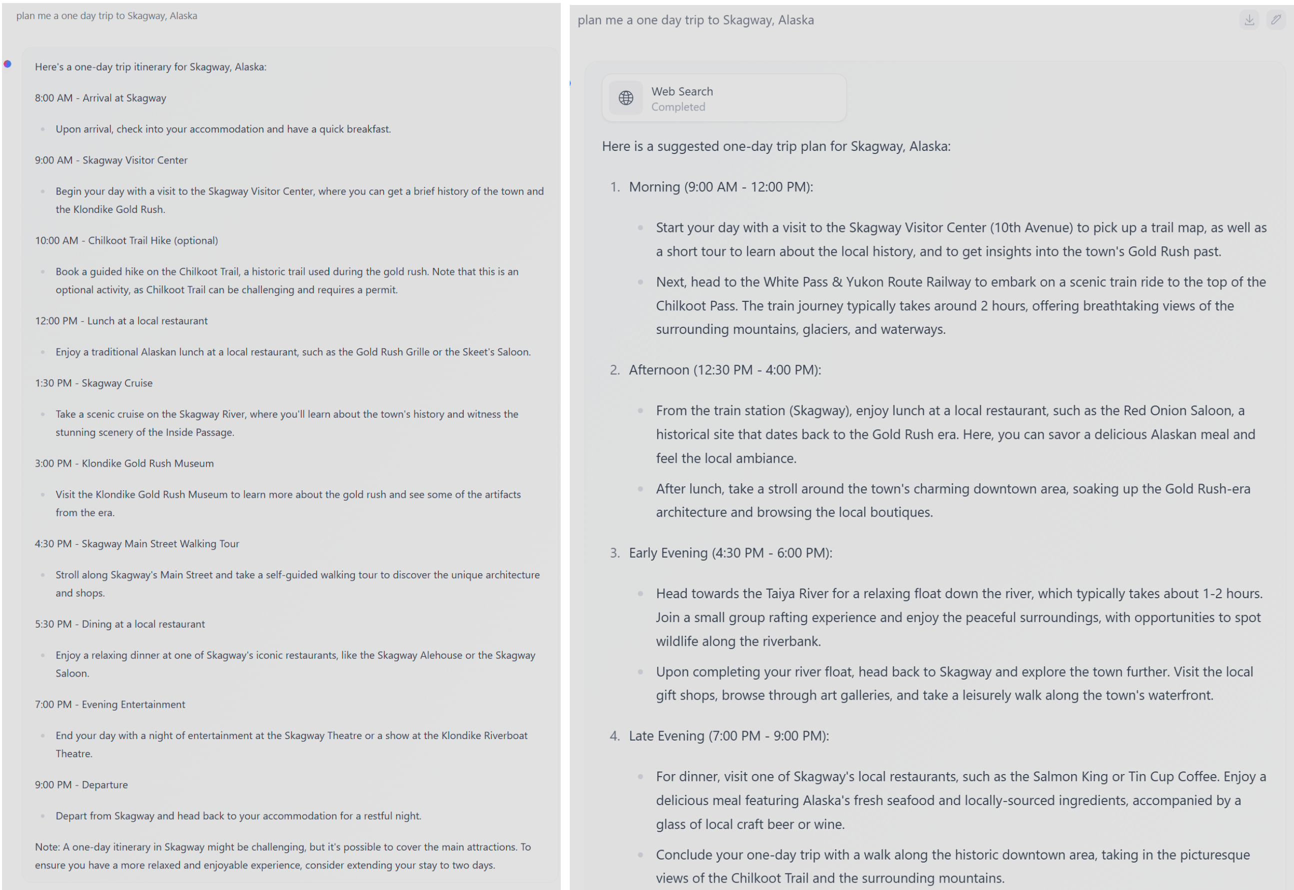

Figure 4: Left: **phi-3-mini**'s completion without search. Right: **phi-3-mini**'s completion with search, using the default HuggingFace Chat-UI search ability.

5 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** in Figure 4. 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.

References

[AI23] Meta AI. Introducing meta llama 3: The most capable openly available llm to date, 2023.

[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.

[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.

[DZZ+24] Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens, 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+21] 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.

[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.

[JBA+23] Mojan 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 Taumann 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 Oufattole, 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.

[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.

[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.

[LHE22] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022.

[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.

[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.

[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.

[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.

[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.

[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, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[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, Zhuohan 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

Marah Abdin
Sam Ade Jacobs
Ammar Ahmad Awan
Jyoti Aneja
Ahmed Awadallah
Hany Awadalla
Nguyen Bach
Amit Bahree
Arash Bakhtiari
Harkirat Behl
Alon Benhaim
Misha Bilenko
Johan Bjorck
Sébastien Bubeck
Martin Cai
Caio César Teodoro Mendes
Weizhu Chen
Vishrav Chaudhary
Parul Chopra
Allie Del Giorno
Gustavo de Rosa
Matthew Dixon
Ronen Eldan
Dan Iter
Abhishek Goswami
Suriya Gunasekar
Emman Haider
Junheng Hao
Russell J. Hewett
Jamie Huynh
Mojan Javaheripi
Xin Jin
Piero Kauffmann
Nikos Karampatziakis
Dongwoo Kim
Mahoud Khademi
Lev Kurilenko
James R. Lee
Yin Tat Lee
Yuanzhi Li
Chen Liang
Weishung Liu
Eric Lin
Zeqi Lin
Piyush Madan
Arindam Mitra
Hardik Modi
Brandon Norick
Anh Nguyen
Barun Patra
Daniel Perez-Becker
Heyang Qin
Thomas Portet
Reid Pryzant
Sambudha Roy
Marko Radmilac
Corby Rosset
Olatunji Ruwase
Olli Saarikivi
Amin Saied
Adil Salim
Michael Santacroce
Shital Shah
Ning Shang
Hiteshi Sharma
Xia Song
Xin Wang
Rachel Ward
Guanhua Wang
Philipp Witte
Michael Wyatt
Jiahang Xu
Can Xu
Sonali Yadav
Fan Yang
Ziyi Yang
Donghan Yu
Chengruidong Zhang
Cyril Zhang
Jianwen Zhang
Li Lyna Zhang
Yi Zhang
Yunan Zhang
Xiren Zhou