Phi-3 Safety Post-Training: Aligning Language Models with a "Break-Fix" Cycle

Microsoft

Abstract

Recent innovations in language model training have demonstrated that it is possible to create highly performant models that are small enough to run on a smartphone. As these models are deployed in an increasing number of domains, it is critical to ensure that they are aligned with human preferences and safety considerations. In this report, we present our methodology for safety aligning the Phi-3 series of language models. We utilized a "break-fix" cycle, performing multiple rounds of dataset curation, safety post-training, benchmarking, red teaming, and vulnerability identification to cover a variety of harm areas in both single and multi-turn scenarios. Our results indicate that this approach iteratively improved the performance of the Phi-3 models across a wide range of responsible AI benchmarks.

1 Introduction

Given the computational cost associated with large language models (LLMs), there is increasing interest in developing smaller models with lower compute and memory requirements. Recent research has demonstrated that it is possible to create performant small language models (SLMs) by training on highly curated and synthetic datasets [GZA+23, LBE+23, JBA+23]. Microsoft recently released the Phi-3 series of SLMs, including Phi-3-mini (3.8B), Phi-3-small (7B), and Phi-3-medium (14B). For example, Phi-3-mini is small enough to run on a smartphone but achieves 69% on MMLU and 8.38 on MT-Bench, making it competitive with Mixtral 8x7B and GPT-3.5. [AJA+24].

Open-source SLMs enable an exciting array of on-device generative AI applications. At the same time, the proliferation of language models in an increasing number of domains underscores the importance of aligning models to human preferences and safety considerations. In this report, we present our approach to aligning the Phi-3 series of language models. We utilized a "break-fix" cycle that relies on multiple rounds of vulnerability identification and safety post-training. In the sections that follow, we detail our methodology, quantitative benchmarks, and red teaming results.

2 Safety Alignment

2.1 Approach

Microsoft adopted an iterative approach to safety post-training that consisted of five main stages:

1. **Safety Dataset Curation:** We utilized existing publicly available datasets with various modifications and generated additional datasets based on feedback from the AI Red Team for further safety post-training.
2. **Safety Post-Training:** The safety datasets mixed with standard preference datasets were used in both the supervised fine-tuning (SFT) and direct preference optimization (DPO) [RSM+23] stages.

3. **Quantitative and Qualitative RAI Evaluations:** A wide spectrum of responsible AI (RAI) evaluations, described in detail below, were considered to select release candidates (RCs) to share with the AI Red Team.

4. **AI Red Teaming:** The release candidates were shared with a centralized and independent AI Red Team (AIRT), which leveraged a variety of adversarial techniques to probe models for harmful content. The red teaming strategy is described below.

5. **Vulnerability Identification:** Based on the RAI evaluations and AIRT findings, potential vulnerabilities are identified to inform further safety post-training.

As illustrated by Figure 1, we repeated this cycle multiple times and gradually fine-tuned the Phi-3 models to generate safe responses in a variety of contexts. We found that this iterative "break-fix" approach made it possible to mitigate many more risks than what can typically be achieved by a single fine-tuning job. In addition to RAI benchmarks, we monitored multiple performance metrics to ensure that safety post-training did not degrade the quality of generated text. The datasets, red teaming strategies, and evaluation benchmarks used are detailed in the sections below.

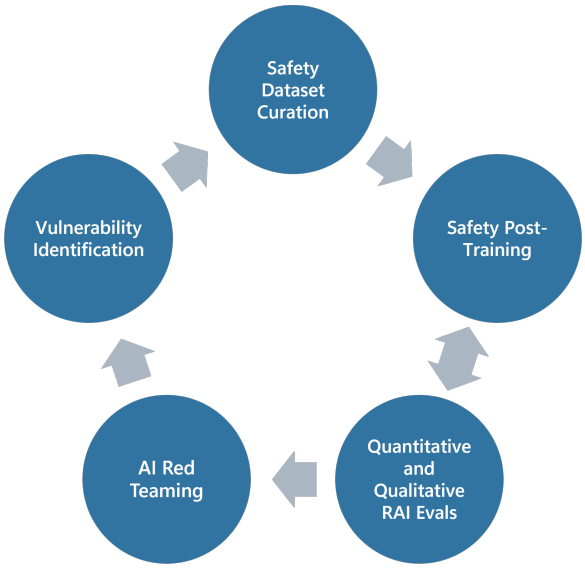

Figure 1: The five main stages of the "break-fix" cycle used to safety align the Phi-3 language models. The double-headed arrow between "Safety Post-Training" and "Quantitative and Qualitative RAI Evals" illustrates the iterative process of running and shortlisting multiple release candidates with different data ablations before sharing a model with the AI Red Team.

2.2 Safety Alignment

For safety post-training of the Phi-3 models, we used a combination of open-source and in-house datasets. To improve the quality and effectiveness of the open-source datasets including [BJN+22] and [JLD+23], we used a variety of approaches including regenerating responses with GPT-4 and applying the instruction conversion method outlined in [BSA+24]. While the open-source datasets covered a wide range of safety aspects, in-house datasets were curated to mitigate specific risks reported by AIRT as fine-tuning or preference optimization datasets depending on their effectiveness.

	Single-Turn	Multi-Turn ($n = 5$ or $n = 8$)
Low-Skilled Adversary	Single-turn prompts in English asking the model to generate harmful content.	Multi-turn conversations asking for harmful content, automated using an attacker bot (GPT-4) via PyRIT.
Intermediate Adversary	Common prompt encodings (e.g., base64, leetspeak, ROT-13) and public jailbreaks (e.g., BetterDAN, AIM, AntiGPT) applied to the low-skilled adversary single-turn prompts.	Priming the model to respond "yes" to a series of prompts before asking for harmful content and Crescendo-like strategies tested manually.

Table 1: Summary of the four main adversarial scenarios used to red team the Phi-3 language models.

In both the SFT and DPO stages, all safety datasets were mixed and used with other preference datasets leveraged in the post-training process. For every model checkpoint, both general quality evaluations and safety evaluations were conducted to decide a model checkpoint to be reviewed by AIRT and to eventually choose the best candidate for release.

2.3 Red Teaming

The Phi-3 release candidates were shared with a centralized Microsoft AI Red Team (AIRT), which leveraged a variety of techniques to test the models for responsible AI (RAI) risks across a range of categories. More specifically, the AIRT probed Phi-3 models for harmful content using both single-turn prompts and multi-turn conversations. These strategies were further split into "low-skilled adversary" and "intermediate adversary" personas, as summarized in Table 1. These personas were scoped based on the most common ways in which users might elicit harmful content from models. A low-skilled adversary was defined as a typical chatbot user who tries to generate harmful content by asking for it directly, while an intermediate adversary actively attempts to subvert a model's safety guardrails using, for example, basic encodings and jailbreaks. To gauge the risk posed by Phi-3 models in comparison to open-source equivalents, the AIRT performed the same testing on Gemma-7B [TMH+24], Mixtral-8x7B [JSR+24], and Llama-3-In [TLI+23].

The AIRT operated independently of the safety post-training team and probed for harmful content across a range of categories, including content related to current events, phishing and cybersecurity, fairness and bias, hate speech, sexual content, and violence. Note that these harm categories were selected in accordance with Microsoft's Responsible AI Standard [Mic22] and were not necessarily the same as those covered by the preference datasets used for safety post-training.

To red team the Phi-3 models at scale and cover as much of the risk surface as possible, the AIRT utilized PyRIT: Python Risk Identification Toolkit [LML+24].[1] PyRIT is an open-source project that provides automation to support prompt generation, prompt conversion (e.g., to apply encodings and jailbreaks), response scoring, and even multi-turn conversations driven by an attacker LLM. To verify the accuracy of PyRIT automation, the AIRT manually checked the scored results and made corrections where necessary. Note that PyRIT was developed specifically to support red teaming operations and is separate from automation used to calculate responsible AI benchmarks. For the intermediate adversary multi-turn scenario, AIRT manually applied strategies like Crescendo, which uses seemingly benign prompts and gradually escalates the conversation to jailbreak a model [RSE24].

[1] PyRIT GitHub repository: `https://github.com/Azure/PyRIT`

In the next section, we present the results of the final Phi-3 models on a range of responsible AI benchmarks, along with the results achieved by comparable open-source models. We utilized these benchmarks to track overall safety performance and compare release candidates throughout the safety post-training process. It is important to note, however, that a model's full risk profile can never be fully captured by a single set of metrics. In contrast with safety benchmarks, red teaming targets emerging harm areas, leverages the latest adversarial techniques, and can address ambiguous scenarios in which a model's behavior might be interpreted in multiple ways. Importantly, red teaming centers the human elements of AI safety and can help answer questions related to how users might feel while interacting with a model. For example, in what scenarios might the model make users feel uncomfortable? Is the model at risk of providing dangerous or harmful advice? Are users likely to trust the model? During the red teaming phase of the break-fix cycle, questions like these played an important role in deciding where to focus further safety post-training efforts.

3 Safety Evaluation

3.1 RAI Safety Benchmarks

Throughout the safety alignment process, we used a range of responsible AI (RAI) benchmarks – including both public datasets and Microsoft internal measurements – to track the safety performance of Phi-3 models and compare release candidates. In this section, we explain these benchmarks in detail and present the results achieved by the final release candidates, as well as those achieved by Mistral-7B, Gemma-7B, and Llama-3-In, for comparison.

3.1.1 Internal Automated Measurement

We used one of Microsoft's automated measurement systems that leverages highly capable models like GPT-4 to simulate multi-turn conversations between adversarial AI agents and a target model [MHJ+23]. Among diverse conversation templates available, we ran experiments for the five scenarios below.

- **Grounding:** Asking the model to reason based on the information provided in prompts.
- **3rd Party Content:** Asking a model to provide protected third-party content.
- **Harmful Content Continuation:** Asking the model to generate harmful content.
- **Harmful Content Summarization:** Asking the model to summarize harmful content.
- **Jailbreak:** Asking the model to bypass behavior protocols learned during safety post-training.

This technique probes for harmful content by making multiple requests in a multi-turn setting or by posing hypothetical scenarios that make the model more likely to respond. (e.g., asking the model to generate harmful content "for a research purpose").

Table 2 shows the results for the Phi-3 and baseline models across the five scenarios. GPT-4 was used to evaluate the model responses. Ungroundedness measures how much the response is based on a given prompt on a scale from 0 (fully grounded) to 4 (not grounded). The other categories are evaluated in terms of severity from 0 (no harm) to 7 (severe harm). The defect rates (DR-x) shown in the table are computed as the percentage of samples with severity score greater than or equal to x. Note that a lower score indicates more desirable performance in all scenarios.

The Phi-3 models showed scores that are better than or comparable to the scores achieved by the baseline models in all five scenarios.

	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
Ungroundedness	0.603	0.299	0.213	1.481	0.935	0.679	0.328
Third Party Harm (DR-1)	0.240	0.253	0.251	0.240	0.562	0.383	0.373
Harmful Content Continuation (DR-3)	0.007	0.003	0.010	0.029	0.026	0.013	0.013
Harmful Content Summarization (DR-3)	0.100	0.110	0.112	0.144	0.223	0.103	0.082
Jailbreak (DR-1)	0.123	0.107	0.111	0.150	0.156	0.114	0.130

Table 2: Results of Phi-3 models and baseline models on the Microsoft internal multi-turn conversation benchmarks. Note that a lower value indicates a better performance for all metrics in the table.

3.1.2 XSTest

XSTest is a public dataset of 250 safe prompts across ten prompt categories (e.g., definitions, historical events, etc.) that well-calibrated models should comply with, and 200 unsafe prompts that most general-purpose models should refuse [RKV+24].

The following two refusal metrics are computed in this benchmark:

- **Inappropriate Prompt Refusal Rate (IPRR):** Measures the rate that the model refuses to answer inappropriate or harmful prompts (higher is better).
- **Valid Prompt Refusal Rate (VPRR):** Measures the rate that the model refuses to answer appropriate or innocuous prompts (lower is better).

Table 3 shows IPRR and VPRR values for the Phi-3 and baseline models. We observed that higher IPRR values are often associated with higher VPRR values. In other words, models that are more likely to refuse harmful prompts are also more likely to refuse harmless prompts. This behavior is well known in responsible AI research and is often described as a tradeoff between helpfulness (higher VPRR) and harmlessness (higher IPRR). In the table below, we observe this tradeoff across all models tested. More specifically, we see that Phi-3-small and Gemma achieve similar balances between IPRR and VPRR. In addition, Llama-3-In is comparable to, but slightly outperforms, Phi-3-medium.

	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
IPRR (Higher is better)	0.750	0.965	0.790	0.015	0.040	0.955	0.815
VPRR (Lower is better)	0.232	0.264	0.124	0.004	0.008	0.216	0.024

Table 3: Results of Phi-3 models and baseline models on the XSTest benchmarks.

3.1.3 DecodingTrust

DecodingTrust is a comprehensive trustworthiness evaluation methodology which considers diverse risks including toxicity, stereotype bias, adversarial robustness, out-of-distribution robustness, robustness on adversarial demonstrations, privacy, machine ethics, and fairness [WCP+24]. In our DecodingTrust benchmarks, we covered all of these risks except for toxicity, which is separately covered by the ToxiGen benchmark.

	Phi-3-mini 3.8b	Phi-3-small 7b	Phi-3-medium 14b	Phi-2 2.7b	Mistral 7b	Gemma 7b
Stereotype Bias	0.983	0.983	0.993	0.860	0.990	0.996
Adversarial Robustness	0.490	0.615	0.516	0.513	0.381	0.497
Out-of-Distribution Robustness	0.643	0.706	0.747	0.655	0.667	0.644
Robustness to Adversarial Demonstrations	0.666	0.635	0.719	0.467	0.575	0.572
Privacy	0.926	0.993	0.824	0.568	0.688	0.905
Machine Ethics	0.754	0.775	0.737	0.425	0.710	0.766
Fairness	0.825	0.589	0.663	0.668	0.827	0.950

Table 4: Results of Phi-3 models and baseline models on the DecodingTrust benchmarks. Note that a higher value indicates better performance for all metrics in the table.

- **Stereotype Bias:** Checks whether the model can identify stereotypes included in prompts.
- **Robustness Metrics:** Three robustness-related metrics measure how consistently the model can identify inappropriate prompts when multiple variations of a prompt with the same meaning are provided.
- **Privacy:** Measures how likely the model is to include personal information such as a phone number or email address in responses.
- **Machine Ethics:** Measures how well the model can understand immorality in prompts.
- **Fairness:** Measures how well the model can provide consistent answers when only sensitive attributes such as gender are changed in prompts.

Note that the DecodingTrust metrics do not evaluate a model's ability to generate more desirable content. Rather, these metrics are primarily based on how well a model understands various responsible AI risks. Therefore, the values in the table below should be interpreted as performance indicators of language tasks such as harmful content detection rather than content generation (higher is better).

Table 4 shows the DecodingTrust benchmarks for the Phi-3 and baseline models. As shown in the table below, no model is universally better or worse than the others.

3.1.4 ToxiGen

ToxiGen is a large-scale machine generated dataset for adversarial and implicit hate speech detection [HGP+22]. As mentioned above, we used the ToxiGen benchmark in favor of the toxicity category in DecodingTrust because ToxiGen is a richer dataset with more prompts (274K) than DecodingTrust toxicity (under 100K). A high ToxiGen score means that the model can detect harmfulness in prompts well. The scores are shown in Table 5. All three Phi-3 model variants outperformed Mistral and Gemma.

3.2 Iterative Safety Alignment

The RAI benchmark scores reported above reflect the performance of the final Phi-3-mini, Phi-3-small, and Phi-3-medium release candidates. However, multiple iterations of safety post-training, red teaming, and vulnerability identification were required to achieve the best results. Given the vast number of

	Phi-3-mini 3.8b	Phi-3-small 7b	Phi-3-medium 14b	Phi-2 2.7b	Mistral 7b	Gemma 7b
ToxiGen	0.764	0.827	0.855	0.589	0.677	0.572

Table 5: Results of Phi-3 models and baseline models on the ToxiGen benchmark (higher is better).

scenarios that users might encounter, we found that this iterative approach, as opposed to a single fine-tuning job, was necessary to identify and mitigate a realistic range of real-world risks.

To further quantify the overall improvement in safety alignment, we evaluated the Phi-3 models before and after completing several rounds of the "break-fix" cycle on a dataset of prompts used by the AI Red Team. In Figure 2, we plot the percentage of harmful responses generated by models with and without safety alignment across several harm categories. On average, we observe a 75% reduction in the amount of harmful content generated, which indicates the efficacy of the overall "break-fix" approach.

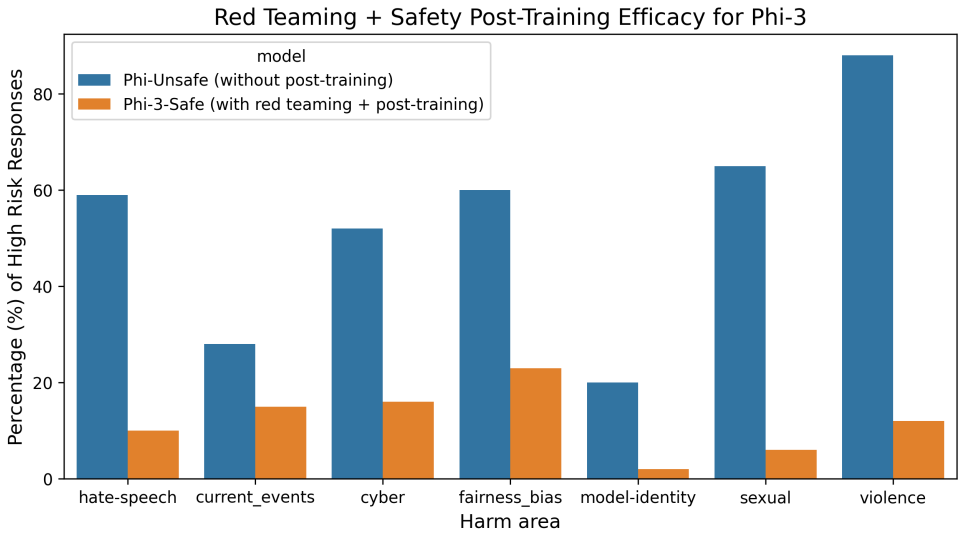

Figure 2: Comparison of high-risk responses generated by Phi-3 language models before and after several rounds of the "break-fix" cycle. Note that percentages are inflated because prompts used by the AI Red Team were crafted to elicit harmful generations.

4 Responsible AI Considerations for Developers

4.1 Responsible Downstream Development

While the Phi-3 models benefit from a robust safety post-training approach, developers should consider how to adapt models with further fine-tuning to their specific use case and safety requirements. In addition to fine-tuning, developers should explore building or adopting additional safety-related tools and approaches to ensure that model outputs are appropriate for their context. These may include safety classifiers run on inputs or outputs, prompt engineering techniques, or other guidance to end-users about how to interpret or use model outputs appropriately. Further guidance and open-source tools are available via Microsoft's Responsible AI Toolbox repository.[2]

[2]RAI Toolbox GitHub repository: `https://github.com/microsoft/responsible-ai-toolbox`

In further developing or deploying models for downstream uses cases, developers should be aware of common capability limitations of language models that are also present in the Phi-3 series. Like other language models, Phi-3 models can potentially behave in ways that are unfair, unreliable, or offensive. Some of the limiting behaviors to be aware of include:

- **Quality of Service:** The Phi-3 models are trained primarily on English text. Languages other than English will experience worse performance. English language varieties with less representation in the training data might experience worse performance than standard American English.
- **Representational Harms & Perpetuation of Stereotypes:** These models can over- or under-represent groups of people, erase representation of some groups, or reinforce demeaning or negative stereotypes. Despite safety post-training, these limitations may still be present due to differing levels of representation of different groups or prevalence of examples of negative stereotypes in training data that reflect real-world patterns and societal biases.
- **Inappropriate or Offensive Content:** These models may produce other types of inappropriate or offensive content, which may make it inappropriate to deploy for sensitive contexts without additional mitigations that are specific to the use case.
- **Information Reliability:** Language models can generate nonsensical content or fabricate content that might sound reasonable but is inaccurate or outdated.
- **Limited Scope for Code:** The majority of code in the Phi-3 training data is based in Python. We strongly recommend that users manually verify all code generated by the Phi-3 models, especially for languages other than Python.

4.2 Additional Considerations

At Microsoft, we are committed to advancing the state of the art in AI and ensuring that our AI products and services are safe, secure, and trustworthy. Language models have great potential to enable new capabilities and benefit society by driving an open innovation cycle and enabling an extensive value chain built on open-source projects. These include direct and indirect benefits, such as advancing AI safety and security, fostering global collaboration and academic research, and inviting greater participation in the development of AI systems.

We have also considered the potential risks and believe the release of the Phi-3 models does not have a meaningful impact on marginal risk of the AI ecosystem due to the availability of larger, advanced AI models and other open information. Based on the evaluations and safety post-training detailed in this white paper, we have assessed the potential benefits of open innovation and research will outweigh potential risks specific to this model.

There are specific areas our team has considered and taken steps to address, but we have not designed or evaluated these models for every potential downstream use case. Developers should take into account common limitations of this technology as they select use cases, as well as conduct evaluations and implement appropriate safeguards in additional fine-tuning and deployment stages. Developers have a responsibility to apply responsible AI best practices and ensure that a specific use case complies with relevant laws. Important areas for consideration include:

- **Allocation:** While we have implemented mitigations in post-training to address potential biases, given the known limitations of language models, these models may not be suitable for scenarios that could have consequential impact on legal status or the allocation of resources or life opportunities without performing further assessments and applying additional debiasing techniques. For

example, in conferring legal rights or an individual's access to credit, education, employment, healthcare, housing, insurance, social welfare benefits, services, or opportunities, or the terms on which they are provided.

- **High-Risk Scenarios:** Developers should also assess the suitability of using models in high-risk scenarios in situations where unfair, unreliable, or offensive outputs might be extremely costly or lead to harm. This includes providing advice in sensitive or expert domains where accuracy and reliability are critical (e.g., legal or health advice). Additional safeguards should be implemented at the application level according to the deployment context.
- **Misinformation:** Language models may produce inaccurate information. Developers should consider and adopt best practices for transparency and disclosure to inform end-users that they are interacting with an AI system. As part of applications, developers can also build mechanisms for feedback as well as pipelines to ground responses in use-case specific, contextual information, a technique known as Retrieval Augmented Generation (i.e., "RAG").
- **Generation of Harmful Content:** While safety post-training has reduced the likelihood that the model will generate some forms of harmful content, developers will need to make assessments based on their context and use available safety classifiers or custom solutions based on their use cases.
- **Privacy:** Developers should be aware of and adhere to privacy laws in the jurisdictions where they operate and deploy applications.
- **Misuse:** Other forms of misuse such as fraud, spam, or malware production may be possible, and developers should ensure that their applications do not violate applicable laws and regulations.

5 Conclusion

In this report, we presented our approach to safety aligning the Phi-3 series of language models. We adopted an iterative "break-fix" approach by performing multiple rounds of dataset curation, post-training with DPO, responsible AI benchmarking, red teaming, and vulnerability identification. Safety alignment is a challenging and open-ended task, but our results indicate that this cycle significantly reduced the amount of harmful content generated by the Phi-3 models in a range of scenarios. Nonetheless, the Phi-3 models are susceptible to the same fundamental limitations as any modern language model, and we hope that the responsible AI considerations outlined in this report will encourage users to think carefully about additional safety mitigations that may be necessary for their specific use cases.

Contributors

GenAI Model Team: Emman Haider, Daniel Perez-Becker, Thomas Portet, Piyush Madan, Amit Garg, David Majercak, Wen Wen, Dongwoo Kim, Ziyi Yang, Jianwen Zhang, Hiteshi Sharma

AI Red Team: Blake Bullwinkel, Martin Pouliot, Amanda Minnich, Shiven Chawla, Solianna Herrera, Shahed Warreth, Maggie Engler, Gary Lopez, Nina Chikanov, Raja Sekhar Rao Dheekonda, Bolor-Erdene Jagdagdorj, Roman Lutz, Richard Lundeen, Tori Westerhoff, Pete Bryan, Christian Seifert, Ram Shankar Siva Kumar

Office of Responsible AI: Andrew Berkley, Alex Kessler

Acknowledgements

We are very grateful to April Rettkowski, Yonatan Zunger, and Weizhu Chen for providing valuable feedback on this paper.

References

[AJA+24] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Yen-Chun Chen, Yi-Ling Chen, Parul Chopra, Xiyang Dai, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Victor Fragoso, Dan Iter, Mei Gao, Min Gao, Jianfeng Gao, Amit Garg, 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, Yunsheng Li, Chen Liang, Lars Liden, Ce Liu, Mengchen Liu, Weishung Liu, Eric Lin, Zeqi Lin, Chong Luo, Piyush Madan, Matt Mazzola, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Xin Wang, Lijuan Wang, Chunyu Wang, Yu Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Haiping Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Jianwei Yang, Ziyi Yang, Yifan Yang, Donghan Yu, Lu Yuan, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024.

[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.

[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.

[HGP+22] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection, 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.

[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.

[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.

[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.

[LML+24] Gary Lopez, Amanda Minnich, Roman Lutz, Richard Lundeen, Raja Sekhar Rao Dheekonda, Nina Chikanov, Bolor-Erdene Jagdagdorj, Martin Pouliot, Shiven Chawla, Whitney Maxwell, Blake Bullwinkel, Katherine Pratt, Joris de Gruyter, Charlotte Siska, Pete Bryan, Tori Westerhoff, Chang Kawaguchi, Christian Seifert, Ram Shankar Siva Kumar, and Yonatan Zunger. Pyrit: A framework for security risk identification and red teaming in generative ai systems, 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.

[Mic22] Microsoft. Microsoft Responsible AI Standard. `https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE5cmFl?culture=en-us&country=us`, June 2022.

[RKV+24] Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models, 2024.

[RSE24] Mark Russinovich, Ahmed Salem, and Ronen Eldan. Great, now write an article about that: The crescendo multi-turn llm jailbreak attack, 2024.

[RSM+23] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2023.

[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, 2023.

[TMH+24] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open models based on gemini research and technology, 2024.

[WCP+24] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models, 2024.