AURORA-M: Open Source Continual Pre-training for Multilingual Language and Code

Taishi Nakamura[*1], **Mayank Mishra**[*2], **Simone Tedeschi**[*3,4], **Yekun Chai**[5]
Jason T Stillerman, **Felix Friedrich**[6,7], **Prateek Yadav**[8], **Tanmay Laud**,
Vu Minh Chien[9], **Terry Yue Zhuo**[10,11], **Diganta Misra**[12,13], **Ben Bogin**[14],
Xuan-Son Vu[15,16,17], **Marzena Karpinska**[18], **Arnav Varma Dantuluri**, **Wojciech Kusa**[33],
Tommaso Furlanello, **Rio Yokota**[1], **Niklas Muennighoff**, **Suhas Pai**[19],
Tosin Adewumi[20], **Veronika Laippala**, **Xiaozhe Yao**[21], **Adalberto Junior**,
Alpay Ariyak[22,23], **Aleksandr Drozd**[24], **Jordan Clive**[25], **Kshitij Gupta**[12],
Liangyu Chen, **Qi Sun**[1], **Ken Tsui**, **Noah Persaud**,
Nour Fahmy, **Tianlong Chen**[8], **Mohit Bansal**[8], **Nicolò Monti**[26],
Tai Dang[18], **Ziyang Luo**[27], **Tien-Tung Bui**[28], **Roberto Navigli**[3],
Virendra Mehta[29], **Matthew Blumberg**[#30], **Victor May**[#31,32], **Huu Nguyen**[#32],
Sampo Pyysalo[#34]

[1]Institute of Science Tokyo, [2]MIT-IBM Watson Lab, [3]Sapienza University of Rome,
[4]Babelscape, [5]LAION, [6]TU Darmstadt, [7]hessian.AI, [8]UNC Chapel-Hill
[9]Detomo Inc., [10]CSIRO's Data61, [11]Monash University, [12]Mila - Quebec AI Institute
[13]Carnegie Mellon University, [14]Allen Institute for AI, [15]DeepTensor AB,
[16]WASP Media & Language, [17]Umeå University, [18]University of Massachusetts Amherst,
[19]Hudson Labs, [20]Luleå University of Technology, [21]ETH Zurich, [22]RunPod, [23]OpenChat,
[24]RIKEN CCS, [25]Chattermill AI, [26]ASC27, [27]Hong Kong Baptist University, [28]DopikAI JSC
[29]University of Trento, [30]GridRepublic, [31]Chegg, [32]Ontocord.AI, [33]TU Wien, [34]University of Turku

taishi@rio.scrc.iir.isct.ac.jp, mayank.mishra2@ibm.com,
tedeschi@diag.uniroma1.it, praty@cs.unc.edu
diganta.misra@mila.quebec, mayvic@gmail.com,
s.vu@deeptensor.ai, huu@ontocord.ai, sampo.pyysalo@utu.fi

*Equal contribution #Equal mentoring

Abstract

Pretrained language models are an integral part of AI applications, but their high computational cost for training limits accessibility. Initiatives such as BLOOM and STARCODER aim to democratize access to pretrained models for collaborative community development. Despite these efforts, such models encounter challenges such as limited multilingual capabilities, risks of catastrophic forgetting during continual pretraining, and the high costs of training models from scratch, alongside the need to align with AI safety standards and regulatory frameworks.

This paper presents **AURORA-M**, a 15B parameter multilingual open-source model trained on English, Finnish, Hindi, Japanese, Vietnamese, and code. Continually pretrained from STARCODERPLUS on 435B additional tokens, AURORA-M surpasses 2T tokens in total training token count. It is the first open-source multilingual model fine-tuned on human-reviewed safety instructions, thus aligning its development not only with conventional red-teaming considerations, but also with the specific concerns articulated in the Biden-Harris Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. We evaluate AURORA-M across a wide range of tasks and languages, showcasing its robustness against catastrophic forgetting and its superior performance in multilingual settings, particularly in safety evaluations. We open-source AURORA-M and its variants to encourage responsible open-source development of large language models at https://huggingface.co/aurora-m.

Proceedings of the 31st International Conference on Computational Linguistics: Industry Track, pages 656–678
January 19–24, 2025.

1 Introduction

Large Language Models (LLMs) are fundamental tools in artificial intelligence, powering applications such as machine translation, text summarization, dialogue systems, and code generation. These LLMs are pre-trained on extensive text data to enhance downstream task-specific adaptation. However, the excessive computational expense of pretraining LLMs creates barriers to access, constraining wider development.

Open-source initiatives such as BLOOM (Scao et al., 2023), STARCODER (Li et al., 2023a), STARCODER-2 (Lozhkov et al., 2024), PYTHIA (Biderman et al., 2023), and OLMO (Groeneveld et al., 2024; Soldaini et al., 2024) have emerged to democratize access to pre-trained LLMs. These initiatives stimulate innovation, allowing researchers and developers to leverage existing advancements. However, despite their contributions, several significant challenges persist in the domain of open-source LLM development.

Primarily, several studies (Bang et al., 2023; Jiao et al., 2023; Hendy et al., 2023; Huang et al., 2023) have underscored the ongoing struggle of LLMs with non-English texts, particularly in low- or extremely low-resource languages. Given that the training data predominantly consists of English, as noted for instance by Brown et al. (2020) who reported that English accounts for 93% of GPT-3's training corpus, there is a pressing need to promote the development of multilingual models to democratize LLMs and alleviate performance disparities across different languages (Chai et al., 2023). Secondly, continual pretraining – a technique involving further updating pretrained models on new data distributions to enhance their capabilities (Gupta et al., 2023; Fujii et al., 2024) – poses a significant challenge. While this approach could potentially enable life-long learning of large language models, it often leads to catastrophic forgetting, where the model loses previously acquired knowledge. This challenge is exacerbated when considering the continual pretraining of models across a diverse array of grammatical and lexical structures. Lastly, ensuring compliance with recent regulations mandating safe and secure AI development practices represents another critical aspect often overlooked in open-source LLM development, specifically, for multilingual models.

This paper presents AURORA-M, a novel open-source multilingual Large Language Model (LLM) with 15 billion parameters, tailored to address the aforementioned limitations. AURORA-M is designed to cater to five linguistically diverse languages: English, Finnish, Hindi, Japanese, Vietnamese, with a mix of code data. AURORA-M is continually pretrained from the STARCODERPLUS model (Li et al., 2023a) on an extensive dataset comprising 435 billion tokens, resulting in a total training token count of an impressive 2 trillion tokens. This rigorous pretraining regimen equips AURORA-M with a comprehensive understanding of diverse languages and code. Moreover, safety is a fundamental design principle of AURORA-M. It stands out as the first open-source multilingual LLM fine-tuned on a comprehensive collection of human-reviewed safety instructions addressing concerns in the Biden-Harris Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (WhiteHouse, 2023). This fine-tuning process not only addresses conventional red-teaming concerns (Ganguli et al., 2022; Perez et al., 2022) aimed at testing system vulnerabilities, but also aligns with the specific safety and security guidelines outlined in the Order.

To comprehensively evaluate AURORA-M's efficacy, we conduct a rigorous examination across a diverse spectrum of tasks spanning various domains and languages. Our evaluations aim to gauge AURORA-M's capacity to retain previously learned knowledge while acquiring new capabilities through continual pretraining. We demonstrate that AURORA-M successfully avoids catastrophic forgetting on English and coding tasks. Furthermore, we benchmark AURORA-M against state-of-the-art multilingual models, showcasing its competitive performance in these settings. Additionally, safety evaluations are conducted to scrutinize AURORA-M's tendency to generate undesired or potentially illicit content. The findings from these assessments affirm AURORA-M's commitment to safety and the adherence to responsible AI development practices.

Our contributions can be summarized as follows.

- We introduce AURORA-M, a new 15B continually pretrained red-teamed multilingual LLM built on top of the StarCoderPlus model (Li et al., 2023a).

- We develop a two-stage curriculum of continual pretraining consisting of **Continual Auxiliary Pretraining** (CAP) and **Continual Alignment Tuning** (CAT) aimed at maximiz-

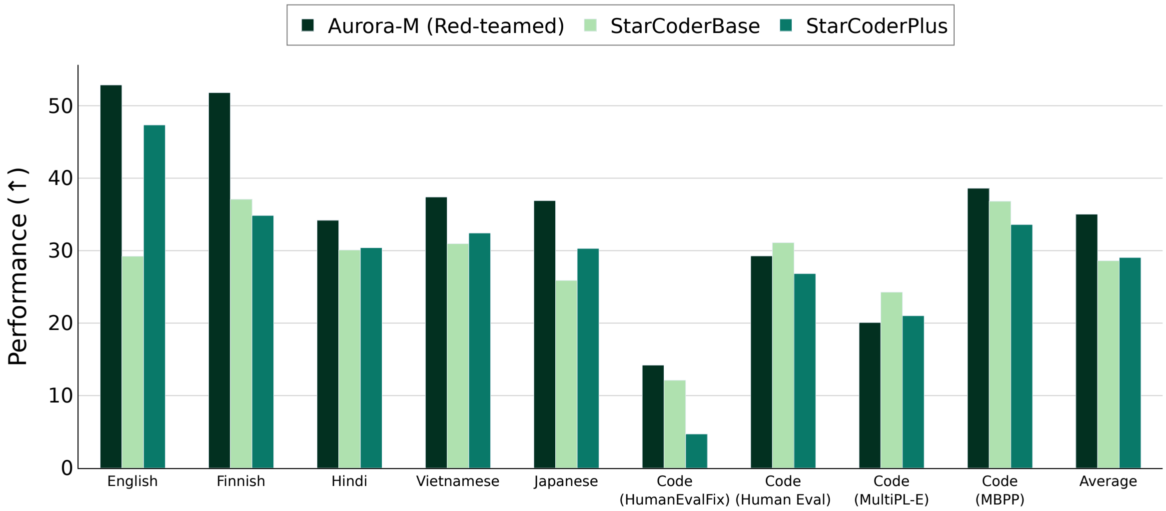


Figure 1: Comparison of overall performance between AURORA-M-redteamed and its predecessors, STARCODERBASE and STARCODERPLUS, across diverse code and multilingual language evaluation benchmarks. Pass@1 performance averages for code benchmarks are reported. For natural language evaluations, 0-shot accuracy averages are reported for languages other than English and Japanese. English evaluation is 8-shot, while Japanese evaluation uses a combination of 4-shot and 1-shot.

ing adaptation, minimizing catastrophic forgetting, and aligning AURORA-M with safety objectives.

- We extensively evaluate AURORA-M across various tasks in different domains and languages, demonstrating its superior performance in multilingual settings while retaining competitive performance in English and coding.
- We construct a new red-teaming dataset, named "The Biden-Harris Redteam Dataset," tailored to address concerns outlined in the Executive Order along with typical safety concerns. We then fine-tune AURORA-M on this dataset and evaluate on several safety benchmarks.
- We show the influence of scaling the total training tokens on various multilingual and code evaluation tasks.

2 Datasets

Data Curation. The continual pretraining process for training AURORA-M followed a carefully designed two-stage curriculum, as shown in Fig. 2. In the first stage, termed as **Continual Auxiliary Pretraining** (CAP), a large corpus of general multilingual web data was used to expose the model to diverse data, laying a robust foundation for subsequent training. The second stage, termed as **Continual Alignment Tuning** (CAT) employed a strategic data-mixing approach to bolster the model's performance in targeted areas and align it with our predefined objectives. Following Taylor et al. (2022) and Li et al. (2023b), we also included publicly available instruction tuning datasets in both stages of training.

In CAP, we incorporated 377B tokens of processed and filtered web data from various sources, including Stack (Kocetkov et al., 2022), RefinedWeb (Penedo et al., 2023), RedPajama (Together, 2023), and a subset of the Pile (Gao et al., 2020). Additionally, multilingual data from HPLT (de Gibert et al., 2024), MC4 (Zhu et al., 2023a), Paracrawl (Ghussin et al., 2023), OSCAR (Abadji et al., 2022), along with Wikipedia (Foundation, 2023), and instruction tuning data from sources such as OpenAssistant (Köpf et al., 2023), APIBench (Patil et al., 2023), and OIG (LAION, 2023) were included.

For CAT, we opted for a greater percentage of code and a changed mix of high-quality public instruction datasets (Mishra et al., 2022a; Ding et al., 2023; Ivison et al., 2023), encompassing coding (Luo et al., 2023; Mishra et al., 2023a) and mathematical reasoning (Yu et al., 2023; Mishra et al., 2023b). The intention was to not overfit to the high quality instruction data, and thus the high quality data was used in CAT only. We also subsampled data from CAP for quality, as described below. Fur-

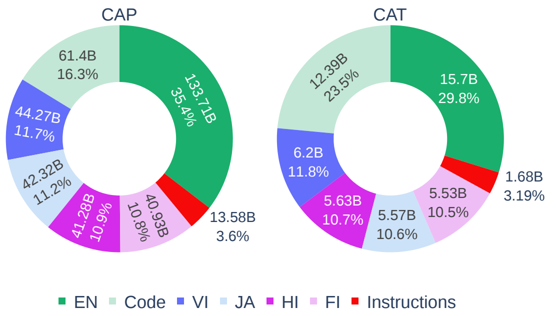


Figure 2: Training data distribution of languages, code, and instructions used for the two-stage continual pre-training of the AURORA-M model. There are a total of 377B and 58B tokens in the Continual Auxiliary Pre-training (CAP) and Continual Alignment Tuning (CAT) stages respectively.

thermore, we introduced a new safety instruction dataset named **Biden-Harris Redteam**, detailed in Section 4. The total dataset size for CAT is 58B tokens. We refer the reader to Fig. 2 for the distribution of languages in both training stages. The complete list of datasets is available in Appendix B.

Data Filtering. To remove toxic content and low-quality text, we applied filters similar to those used in Nguyen et al. (2023c) and Scao et al. (2023), such as stop-word proportions and text length. For all web text, we followed a process akin to Penedo et al. (2023) to remove low-quality content, including duplicate headers and footers. Additionally, in the CAT dataset, we further filtered web text with high proportions of symbols and numbers. In the case of RefinedWeb (Penedo et al., 2023), we utilized the RedPajama (Together, 2023) fastText classifier to retain English webpages resembling "high-quality" content similar to Wikipedia-linked articles. We trained and employed a similar classifier to filter other languages in our dataset, except for Finnish, where the procedure caused over-filtering, resulting in an excessively low sample volume post-filtering. To further enhance the quality of the RefinedWeb data, we adopted an approach detailed in Rönnqvist et al. (2021). We trained a fastText classifier* and selectively subsampled web pages with over-represented registers, aiming to retain more "rare" text (e.g., lyrical or poetic text). This filtering process was specifically applied to English text due to the prohibitive slowness of our multilingual classifiers. Addressing this limitation represents an area for future research.

Data Processing. In the second stage dataset, we undertook the detection and anonymization of sensitive information, including government IDs, within web-based texts to uphold privacy and ethical standards similar to Scao et al. (2023). For data segments derived from arXiv, USPTO, and StackExchange within the Pile dataset (Gao et al., 2020), we reconstructed the data from the original source to restore metadata, which we then appropriately appended to the texts.

3 Model Training

AURORA-M was trained on the LUMI supercomputer†, utilizing 128 AMD MI250X GPUs for 48 days. The training process operated entirely on 100% hydro-powered energy and included waste heat recycling. For orchestration, we adapted a segment of the Bigcode fork of Megatron-LM (Narayanan et al., 2021) using the HIP runtime. For training, we distributed the model using 4-way Tensor Parallelism and 4-way Pipeline Parallelism using the 1F1B schedule to reduce the pipeline bubble (Narayanan et al., 2021). We also used Megatron's distributed optimizer (Narayanan et al., 2021) to distribute the optimizer states across data-parallel processes and eliminate redundancy, reducing the required memory usage.

For the training of AURORA-M, we maintained a consistent batch size of 2048 and a sequence length of 2048 tokens. The learning rate was linearly warmed up to 10^{-4} over 2,000 steps, followed by a cosine decay scheduler set to decay the learning rate to 10^{-5} by 120,000 steps. while optimization utilized the AdamW optimizer (Kingma and Ba, 2017; Loshchilov and Hutter, 2019) with coefficients $\beta_1 = 0.9$ and $\beta_2 = 0.95$. Additionally, Megatron-LM's distributed optimizer with mixed precision training (Micikevicius et al., 2018) was used. Further training details can be found in the Appendix A.

4 Safety

LLMs can propagate harmful content, reinforce biases, or amplify misinformation. While users are responsible for assessing the potential risks of generated content, developers must prioritize legal and safety considerations, strengthening models against attacks that may bypass safety protocols.

*Similar to https://github.com/TurkuNLP/register-labeling?tab=readme-ov-file

†https://www.lumi-supercomputer.eu/

In line with the Biden-Harris US Executive Order on AI (WhiteHouse, 2023), we curated the Biden-Harris Redteam Dataset, consisting of 5000 instruction-response pairs, addressing key concerns such as harm, cyber-attacks, CNBR risks, illegal acts, and privacy infringement. This dataset was created using a combination of filtering human preference data on harmlessness and template-based methods, with responses reviewed and edited for quality and safety. We used this dataset to instruction-tune AURORA-M and evaluated its safety levels before and after tuning. Details are provided in Section 5, with further dataset insights in Appendix C.

5 Evaluation

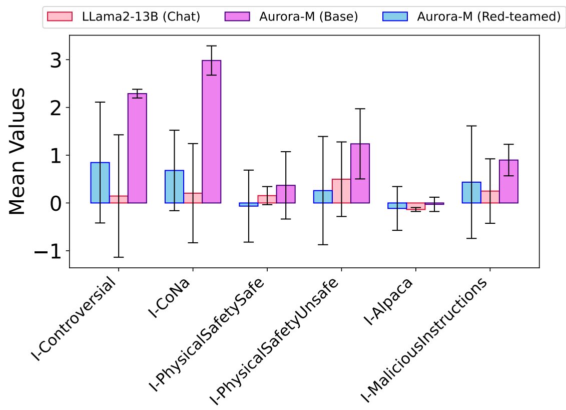
5.1 Evaluation Setup

We evaluated models across several English, Japanese, Finnish, Hindi, Vietnamese, and code-related benchmarks. For English, we used the Language Model Evaluation Harness (Gao et al., 2022) to assess tasks like OpenBookQA, TriviaQA, HellaSwag, SQuAD2.0, XWINO, and GSM8K. For Japanese, we followed swallow-llama and used `llm-jp-eval` (Han et al., 2024), covering JCommonsenseQA, JEMHopQA, and JSQuAD, among others. Finnish evaluation followed the method used in FinGPT with FIN-bench (Luukkonen et al., 2023a). We also evaluated Hindi and Vietnamese using the mlmm evaluation suite on tasks like HellaSwag and MMLU. For code evaluation, we utilized MBPP, HumanEval, MultiPL-E, and HumanEvalFix, and for safety, we employed datasets like the Biden-Harris Redteam Testset and DangerousQA. Detailed dataset descriptions and their corresponding evaluation metrics are provided in Appendix D.

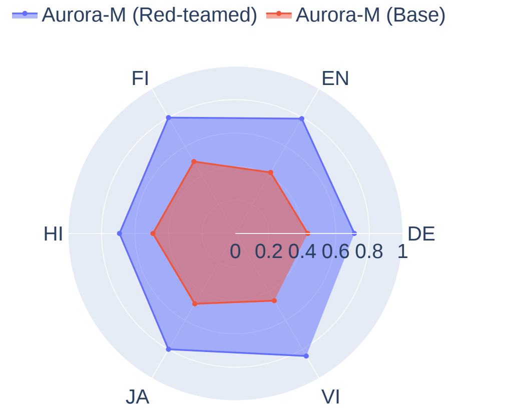
5.2 Evaluation Results

Figure 1 illustrates the superior performance of AURORA-M compared to its base model (*i.e.*, STARCODERPLUS) across an extensive range of code and multilingual benchmarks, underscoring the efficacy of AURORA-M across diverse fields and languages. We observe that AURORA-M can maintain performance on previously learned English and Code benchmarks while significantly outperforming on new language benchmarks.

Evaluation on Natural Languages. Tables 1, 2, 3, 4 demonstrate the respective performance on the targeted languages, showing that AURORA-M consistently outperforms the performance of its starting checkpoint, STARCODERPLUS, and many other baselines, such as LLAMA-2-7B.



(a) Harmfulness scores of our base model (pink) compared to its instruction-tuned version (blue). The lower the better.



(b) CARP scores for the BH-readteamed model and the base model on the Biden-Harris Redteam Testset.

Figure 3: Overall safety results.

Code Evaluation. Tables 5 and 6 illustrate the proficiency of AURORA-M in code generation, demonstrating the possibility of continual pretraining from a code-centric checkpoint on multilingual data. In Table 5, the HumanEval and MBPP evaluation benchmarks assess the model's ability to generate syntactically and semantically correct code snippets. AURORA-M exhibits competitive performance on the Pass@1 metric, which evaluates the model's ability to produce a correct answer on the first attempt. In particular, AURORA-M consistently matches or outperforms StarCoderPlus, suggesting a significant improvement in code synthesis capabilities. In Appendix E.1, we show results on additional code datasets and further analyze the behavior of our system by looking at the

Model	MC	QA		RC	SUM	MATH	MT (WMT20)		Avg.
	JCom	JEMHop	NIILC	JSQuAD	XL-Sum	MGSM	En-Ja	Ja-En	
	4-shot	4-shot	4-shot	4-shot	1-shot	4-shot	4-shot	4-shot	
StarCoderBase (Li et al., 2023a)	29.76	42.08	17.94	73.89	13.96	4.80	15.13	9.59	25.89
StarCoderPlus (Li et al., 2023a)	50.22	**44.19**	17.72	79.24	16.87	5.60	14.58	13.98	30.30
Llama-2-7B (Touvron et al., 2023)	38.52	42.40	34.10	79.17	19.05	7.60	17.83	17.38	32.01
Llama-2-13B (Touvron et al., 2023)	**69.97**	44.15	41.70	85.33	**21.39**	13.20	21.46	**19.82**	**39.63**
Aurora-M (Red-teamed) (Ours)	46.65	35.73	**50.78**	**87.06**	8.79	**21.20**	**27.78**	17.22	36.90

Table 1: Japanese Evaluation.

Model	0-shot	1-shot	2-shot	3-shot
GPT3-Finnish-8B (Luukkonen et al., 2023b)	42.66	46.53	47.96	48.41
GPT3-Finnish-13B (Luukkonen et al., 2023b)	42.45	46.53	47.14	48.08
StarCoderBase (Li et al., 2023a)	37.07	42.65	42.11	44.43
StarCoderPlus (Li et al., 2023a)	34.85	43.97	44.05	46.49
Llama-2-7B (Touvron et al., 2023)	39.49	46.99	49.03	49.60
Llama-2-13B (Touvron et al., 2023)	45.69	55.70	56.93	**57.50**
Aurora-M (Red-teamed) (Ours)	**51.80**	**56.11**	**57.77**	57.48

Table 2: Finnish Evaluation.

relationship between its performance and the number of training tokens across various languages and modalities.

Safety Evaluation In Figure 3, we provide the safety results comparing our base model against our Biden-Harris red-teamed model obtained by instruction-tuning the former on the dataset introduced in Section 4. For the Biden-Harris Redteam Testset evaluation, four volunteers reviewed both models' responses and scored them with -2 if harmful, 1 if not helpful but harmless, and 2 if both helpful and harmless. We term the percentage of the total score per category compared to its maximum possible score as the Continual Alignment Redteam Percentage ("CARP"). We can immediately appreciate the considerably lower harmfulness both on the existing benchmarks and on our own Biden-Harris red-team test set as evident by the CARP scores obtained by our red-teamed Aurora-M. *We also note that even though our instruction set is predominantly in English, safety consistently improved not only in our target languages but also in languages we did not specifically focus on, such as German, thus showing strong indications of cross-lingual red-teaming effects*. Furthermore, as shown in Appendix E.1, the Attack Success Rate (ASR) on DangerousQA was also reduced.

5.3 Training Analysis

Figure 5 and 6 show the relationship between the number of training tokens and the performance of the various models. This analysis aims to capture these trends for the code generation tasks such as HumanEval and MBPP, as well as for the English, Finnish, Hindi, Japanese, and Vietnamese language evaluations. We refer to Appendix E.2 for detailed discussion.

6 Related Work

Expanding Multilingual Language Models. Initially, the development of LLMs has predominantly targeted the English language (Brown et al., 2020), leveraging the extensive corpus of English data available on the Web and the broad applicability of models trained on English text. However, this emphasis has often come at the cost of accommodating the linguistic diversity found across various language demographics (Zhu et al., 2023b; Bang et al., 2023; Zhang et al., 2024). Recognizing this significant limitation (Robinson et al., 2023; Peng et al., 2024), recent research has proposed foundational LLMs equipped with multilingual capabilities (Chai et al., 2023; Scao et al., 2023; Wei et al., 2023; Shliazhko et al., 2022), or has explicitly concentrated on addressing the challenges posed by low-resource languages (Üstün et al., 2024; Singh et al., 2024; Gala et al., 2023). To integrate multilingual capabilities into existing LLMs, researchers have proposed a variety of methods to enhance multilingual adaptation. These approaches range from continual pretraining techniques (Ibrahim et al., 2024; Gupta et al., 2023) to initial training on extensive multilingual datasets (Scao et al., 2023; Chai et al., 2023) and then subsequent specialized fine-tuning on a target language (Yang et al., 2023; Han et al., 2022), and even adaptation through instruction tuning (Shaham et al., 2024; Kew et al., 2023; Gala et al., 2024). Critical aspects in multilingual adaptation remain on the availability of high-quality diverse multilingual corpus (Corrêa et al., 2024) and further the scope of vocabulary of the specific language.

Continual Pretraining. Static datasets are impractical for adapting to evolving real-world data,

Model	ARC		HellaSwag		MMLU		TruthfulQA		Avg	
	VI	HI	VI	HI	VI	HI	VI	HI	VI	HI
StarCoderBase (Li et al., 2023a)	22.14	20.72	29.74	26.93	27.11	25.15	44.84	47.57	30.96	30.09
StarCoderPlus (Li et al., 2023a)	24.27	20.89	32.67	27.03	27.35	24.91	**45.49**	**48.77**	32.44	30.40
Bloom-7b1 (Scao et al., 2023)	24.87	21.83	37.97	30.78	25.65	25.30	44.77	44.39	33.32	30.58
Llama-2-7b (Touvron et al., 2023)	25.64	21.58	35.20	28.19	27.95	25.33	45.15	46.37	33.49	30.37
Llama-2-13b (Touvron et al., 2023)	30.17	20.98	38.49	29.58	**31.76**	26.19	44.61	43.79	36.25	30.13
ViGPTQA-6b (Nguyen et al., 2023a)	-	-	-	-	-	-	43.26	-	-	-
VinaLlama-7b (Nguyen et al., 2023b)	28.63	18.75	37.39	26.31	27.15	24.12	43.13	39.11	34.07	27.07
Aurora-M (Red-teamed) (Ours)	**31.97**	**27.57**	**41.98**	**35.84**	30.94	**30.01**	44.71	43.31	**37.40**	**34.18**

Table 3: 0-shot evaluation Results for Vietnamese (**VI**) and Hindi (**HI**).

Model	OpenBookQA 8-shot	TriviaQA 8-shot	HellaSwag 8-shot	SQuAD2.0 8-shot	XWINO 8-shot	GSM8K 8-shot	Avg.
StarCoderBase (Li et al., 2023a)	19.60	8.20	37.57	27.52	73.51	8.95	29.22
StarCoderPlus (Li et al., 2023a)	34.80	53.50	58.06	34.86	89.25	13.57	47.34
Llama-2-7b (Touvron et al., 2023)	35.80	62.65	58.60	32.07	90.49	14.10	48.95
Llama-2-13b (Touvron et al., 2023)	**37.60**	**72.55**	**61.48**	36.81	**91.40**	24.03	**53.98**
Aurora-M (Red-teamed) (Ours)	36.60	51.86	54.73	**48.98**	88.52	**36.47**	52.86

Table 4: English Evaluation.

Model	HumanEval			MBPP		
	Pass@1	Pass@10	Pass@100	Pass@1	Pass@10	Pass@100
StarCoderBase (Li et al., 2023a)	**31.10**	**54.88**	**84.15**	36.80	**61.60**	**81.00**
StarCoderPlus (Li et al., 2023a)	26.83	47.56	73.17	33.60	57.00	77.80
Aurora-M (Red-teamed) (Ours)	29.27	49.39	81.71	**38.60**	61.00	78.00

Table 5: HumanEval & MBPP evaluation results.

making continual learning essential (Ring, 1998; Thrun, 1998). Continual pretraining (Gururangan et al., 2020) allows models to incorporate new knowledge without retraining from scratch, a costly endeavor. As curated datasets like RedPajama (Together, 2023) and Dolma (Soldaini et al., 2024) become available, integrating them efficiently is crucial. This also enables the extension of models to new modalities, such as code (e.g., StableCode). Previous approaches focus on replay techniques, optimizing learning schedules (Ibrahim et al., 2024), soft masking (Ke et al., 2023), and forward/backward transfer (Yıldız et al., 2024).

7 Conclusion

In this work, we introduced Aurora-M, a multilingual model that extends the capabilities of code-focused LLMs while maintaining their original coding proficiency. We demonstrate that continual training from code to multilingual tasks is feasible, allowing the model to perform well across both domains. Adhering to the safety guidelines of the Biden-Harris US Executive Order on AI, Aurora-M promotes responsible AI development while pushing the boundaries of performance and utility. Our two-stage continual pretraining approach, combined with insights from cross-lingual red-teaming, highlights the adaptability and versatility of modern language models. Aurora-M serves as a valuable resource for both researchers and developers, fostering collaboration and transparency in the open-source AI community. Future work will explore continual pretraining on stronger base models with the same two-stage curriculum, focusing on safety for both LLMs and Multimodal-LLMs. We also aim to develop domain-specific expert models, enhancing task specialization and expanding model versatility.

Ethical Consideration

We believe that transparency and accessibility are fundamental principles in the development and deployment of artificial intelligence technologies. Closed-source LLMs limit public scrutiny, hinder collaboration, and potentially reinforce biases inherent in their development process. In contrast, our commitment to open source models fosters a culture of accountability, collaboration, and inclusivity. By making AURORA-M accessible to all, we promote innovation, empower diverse voices, and strive for equitable outcomes in AI applications. We firmly believe that openness in AI development is essential for creating solutions that truly serve the needs and values of society. To this end, we prioritized safety guardrails in alignment with the Biden-Harris Executive Order on AI. Furthermore, the multilingual capability of AURORA-M enhances its usability for users across the world.

On the other hand, each promise comes with peril, and improved technological access through AURORA-M might also increase the potential number of malicious actors. We overall believe that the general benefit far outweighs the potential misuse and want to emphasize the importance of a considered and ethical use of this technology and thus also of AURORA-M.

Lastly, we recognize that safety and lawfulness can be contextual to different cultures and laws. We recognize that in our work we focused on a U.S. centric standard, and we believe future work should also explore multi-jurisdictional redteaming.

Acknowledgments

This work was supported by the "R&D Hub Aimed at Ensuring Transparency and Reliability of Generative AI Models" project of the Ministry of Education, Culture, Sports, Science and Technology, and used resources of LUMI supercomputer under project_462000316.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. *Preprint*, arXiv:2201.06642.

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

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. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *Preprint*, arXiv:2302.04023.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of WMT*, pages 1–55.

Loubna Ben Allal, Niklas Muennighoff, Logesh Kumar Umapathi, Ben Lipkin, and Leandro von Werra. 2022. A framework for the evaluation of code generation models. `https://github.com/bigcode-project/bigcode-evaluation-harness`.

Rishabh Bhardwaj and Soujanya Poria. 2023. Red-teaming large language models using chain of utterances for safety-alignment. *Preprint*, arXiv:2308.09662.

Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. 2024. Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions. *Preprint*, arXiv:2309.07875.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Federico Cassano, John Gouwar, Daniel Nguyen, Sydney Nguyen, Luna Phipps-Costin, Donald Pinckney, Ming-Ho Yee, Yangtian Zi, Carolyn Jane Anderson, Molly Q Feldman, Arjun Guha, Michael Greenberg, and Abhinav Jangda. 2022. Multipl-e: A scalable and extensible approach to benchmarking neural code generation. *Preprint*, arXiv:2208.08227.

Yekun Chai, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, and Hua Wu. 2023. ERNIE-code: Beyond English-centric cross-lingual pretraining for programming languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10628–10650, Toronto, Canada. Association for Computational Linguistics.

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. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

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

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. 2021. Training Verifiers to Solve Math Word Problems. *CoRR*, abs/2110.14168.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

Nicholas Kluge Corrêa, Sophia Falk, Shiza Fatimah, Aniket Sen, and Nythamar de Oliveira. 2024. Teenytinyllama: open-source tiny language models trained in brazilian portuguese. *Preprint*, arXiv:2401.16640.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Preprint*, arXiv:2205.14135.

Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, Sampo Pyysalo, Stephan Oepen, and Jörg Tiedemann. 2024. A new massive multilingual dataset for high-performance language technologies. *Preprint*, arXiv:2403.14009.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *Preprint*, arXiv:2305.14233.

Wikimedia Foundation. 2023. Wikimedia downloads.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual LLM adaptation: Enhancing japanese language capabilities. In *First Conference on Language Modeling*.

Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Preprint*, arXiv:2305.16307.

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

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *Preprint*, arXiv:2209.07858.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Leo Gao, Jonathan Tow, Stella Biderman, Charles Lovering, Jason Phang, Anish Thite, Fazz, Niklas Muennighoff, Thomas Wang, sdtblck, tttyuntian, researcher2, Zdeněk Kasner, Khalid Almubarak, Jeffrey Hsu, Pawan Sasanka Ammanamanchi, Dirk Groeneveld, Eric Tang, Charles Foster, kkawamu1, xagi dev, uyhcire, Andy Zou, Ben Wang, Jordan Clive, igor0, Kevin Wang, Nicholas Kross, Fabrizio Milo, and silentv0x. 2022. EleutherAI/lm-evaluation-harness: v0.3.0.

Yusser Al Ghussin, Jingyi Zhang, and Josef van Genabith. 2023. Exploring paracrawl for document-level neural machine translation. *Preprint*, arXiv:2304.10216.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. 2024. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*.

Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats L. Richter, Quentin Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. 2023. Continual pre-training of large language models: How to (re)warm your model? *Preprint*, arXiv:2308.04014.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *Preprint*, arXiv:2004.10964.

Namgi Han, Nobuhiro Ueda, Masatoshi Otake, Satoru Katsumata, Keisuke Kamata, Hirokazu Kiyomaru, Takashi Kodama, Saku Sugawara, Bowen Chen, Hiroshi Matsuda, Yusuke Miyao, Yugo Miyawaki, and Koki Ryu. 2024. llm-jp-eval: Automatic evaluation tool for Japanese large language models [llm-jp-eval: 日本語大規模言語モデルの自動評価ツール] (in Japanese). In *the 30th Annual Meeting of Japanese Association for Natural Language Processing (NLP2024)*.

Yaqian Han, Yekun Chai, Shuohuan Wang, Yu Sun, Hongyi Huang, Guanghao Chen, Yitong Xu, and Yang Yang. 2022. X-PuDu at SemEval-2022 task 6: Multilingual learning for English and Arabic sarcasm detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 999–1004, Seattle, United States. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, ..., and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics (ACL)*, pages 4693–4703.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021b. Cuad: An expert-annotated nlp dataset for legal contract review. *Preprint*, arXiv:2103.06268.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *Preprint*, arXiv:2302.09210.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. *Preprint*, arXiv:2305.07004.

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L. Richter, Quentin Anthony, Timothée Lesort, Eugene Belilovsky, and Irina Rish. 2024. Simple and scalable strategies to continually pre-train large language models. *Preprint*, arXiv:2403.08763.

Ai Ishii, Naoya Inoue, and Satoshi Sekine. 2023. Construction of a Japanese multi-hop QA dataset for QA systems capable of explaining the rationale [根拠を説明可能な質問応答システムのための日本語マルチホップqaデータセット構築] (in Japanese). In *The 29th Annual Meeting of Japanese Association for Natural Language Processing (NLP2023)*, pages 2088–2093.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. Camels in a changing climate: Enhancing lm adaptation with tulu 2. *Preprint*, arXiv:2311.10702.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *Preprint*, arXiv:2301.08745.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Zixuan Ke, Yijia Shao, Haowei Lin, Hu Xu, Lei Shu, and Bing Liu. 2023. Adapting a language model while preserving its general knowledge. *arXiv preprint arXiv:2301.08986*.

Tannon Kew, Florian Schottmann, and Rico Sennrich. 2023. Turning english-centric llms into polyglots: How much multilinguality is needed? *Preprint*, arXiv:2312.12683.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Denis Kocetkov, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou, Carlos Muñoz Ferrandis, Yacine Jernite, Margaret Mitchell, Sean Hughes, Thomas Wolf, Dzmitry Bahdanau, Leandro von Werra, and Harm de Vries. 2022. The stack: 3 tb of permissively licensed source code. *Preprint*, arXiv:2211.15533.

Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. 2022. JGLUE: Japanese general language understanding evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2957–2966, Marseille, France. European Language Resources Association.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. Openassistant conversations – democratizing large language model alignment. *Preprint*, arXiv:2304.07327.

LAION. 2023. Oig: the open instruction generalist dataset".

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, Nicolas Gontier, Nicholas Meade, Armel Zebaze, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Benjamin Lipkin, Muhtasham Oblokulov, Zhiruo Wang, Rudra Murthy, Jason Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Nour Fahmy, Urvashi Bhattacharyya, Wenhao Yu, Swayam Singh, Sasha Luccioni, Paulo Villegas, Maxim Kunakov, Fedor Zhdanov, Manuel Romero, Tony Lee, Nadav Timor, Jennifer Ding, Claire Schlesinger, Hailey Schoelkopf, Jan Ebert, Tri Dao, Mayank Mishra, Alex Gu, Jennifer Robinson, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. 2023a. Starcoder: may the source be with you! *Preprint*, arXiv:2305.06161.

Shenggui Li, Hongxin Liu, Zhengda Bian, Jiarui Fang, Haichen Huang, Yuliang Liu, Boxiang Wang, and Yang You. 2023b. Colossal-AI: A Unified Deep Learning System For Large-Scale Parallel Training. In *Proceedings of the 52nd International Conference on Parallel Processing*, ICPP '23, page 766–775, New York, NY, USA. Association for Computing Machinery.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, Tianyang Liu, Max Tian, Denis Kocetkov, Arthur Zucker, Younes Belkada, Zijian Wang, Qian Liu, Dmitry Abulkhanov, Indraneil Paul, Zhuang Li, Wen-Ding Li, Megan Risdal, Jia Li, Jian Zhu, Terry Yue Zhuo, Evgenii Zheltonozhskii, Nii Osae Osae Dade, Wenhao Yu, Lucas Krauß, Naman Jain, Yixuan Su, Xuanli He, Manan Dey, Edoardo Abati, Yekun Chai, Niklas Muennighoff, Xiangru Tang, Muhtasham Oblokulov, Christopher Akiki, Marc Marone, Chenghao Mou, Mayank Mishra, Alex Gu, Binyuan Hui, Tri Dao, Armel Zebaze, Olivier Dehaene, Nicolas Patry, Canwen Xu, Julian McAuley, Han Hu, Torsten Scholak, Sebastien Paquet, Jennifer Robinson, Carolyn Jane Anderson, Nicolas Chapados, Mostofa Patwary, Nima Tajbakhsh, Yacine Jernite, Carlos Muñoz Ferrandis, Lingming Zhang, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. 2024. Starcoder 2 and the stack v2: The next generation. *Preprint*, arXiv:2402.19173.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023. Wizardcoder: Empowering code large language models with evol-instruct. *Preprint*, arXiv:2306.08568.

Risto Luukkonen, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, Thomas Wang, Nouamane Tazi, Teven Scao, Thomas Wolf, Osma Suominen, Samuli Sairanen, Mikko Merioksa, Jyrki Heinonen, Aija Vahtola, Samuel Antao, and Sampo Pyysalo. 2023a. FinGPT: Large generative models for a small language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2710–2726, Singapore. Association for Computational Linguistics.

Risto Luukkonen, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, Thomas Wang, Nouamane Tazi, Teven Scao, Thomas Wolf, Osma Suominen, Samuli Sairanen, Mikko Merioksa, Jyrki Heinonen, Aija Vahtola, Samuel Antao, and Sampo Pyysalo. 2023b. FinGPT: Large generative models for a small language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2710–2726. Association for Computational Linguistics.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed precision training. *Preprint*, arXiv:1710.03740.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Mayank Mishra, Prince Kumar, Riyaz Bhat, Rudra Murthy, Danish Contractor, and Srikanth Tamilselvam. 2023a. Prompting with pseudo-code instructions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15178–15197, Singapore. Association for Computational Linguistics.

Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, and Ashwin Kalyan. 2023b. Lila: A unified benchmark for mathematical reasoning. *Preprint*, arXiv:2210.17517.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022a. Cross-task generalization via natural language crowdsourcing instructions. *Preprint*, arXiv:2104.08773.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022b. Cross-task generalization via natural language crowdsourcing instructions. In *ACL*.

Niklas Muennighoff, Qian Liu, Armel Zebaze, Qinkai Zheng, Binyuan Hui, Terry Yue Zhuo, Swayam Singh, Xiangru Tang, Leandro Von Werra, and Shayne Longpre. 2023a. Octopack: Instruction tuning code large language models. *arXiv preprint arXiv:2308.07124*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023b. Crosslingual generalization through multitask finetuning. *Preprint*, arXiv:2211.01786.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vijay Anand Korthikanti, Dmitri Vainbrand, Prethvi Kashinkunti, Julie Bernauer, Bryan Catanzaro, Amar Phanishayee, and Matei Zaharia. 2021. Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–15.

Minh Thuan Nguyen, Khanh Tung Tran, Nhu Van Nguyen, and Xuan-Son Vu. 2023a. ViGPTQA - state-of-the-art LLMs for Vietnamese question answering: System overview, core models training, and evaluations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 754–764, Singapore. Association for Computational Linguistics.

Quan Nguyen, Huy Pham, and Dung Dao. 2023b. Vinallama: Llama-based vietnamese foundation model. *Preprint*, arXiv:2312.11011.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023c. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *arXiv preprint arXiv:2309.09400*.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Pedro Ortiz Suarez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. 2023. Gorilla: Large language model connected with massive apis. *Preprint*, arXiv:2305.15334.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Qiwei Peng, Yekun Chai, and Xuhong Li. 2024. HumanEval-XL: A multilingual code generation benchmark for cross-lingual natural language generalization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8383–8394, Torino, Italia. ELRA and ICCL.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. *Preprint*, arXiv:2202.03286.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 784–789.

Mark B Ring. 1998. Child: A first step towards continual learning. In *Learning to learn*, pages 261–292. Springer.

Nathaniel R. Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. Chatgpt mt: Competitive for high- (but not low-) resource languages. *Preprint*, arXiv:2309.07423.

Samuel Rönnqvist, Valtteri Skantsi, Miika Oinonen, and Veronika Laippala. 2021. Multilingual and zero-shot is closing in on monolingual web register classification. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 157–165, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. 2024. Code llama: Open foundation models for code. *Preprint*, arXiv:2308.12950.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Satoshi Sekine. 2003. Development of a question answering system focused on an encyclopedia [百科事典を対象とした質問応答システムの開発] (in Japanese). In *the 9th Annual Meeting of Japanese Association for Natural Language Processing (NLP2003)*, pages 637–640.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. *Preprint*, arXiv:2401.01854.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *the Eleventh International Conference on Learning Representations*.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mgpt: Few-shot learners go multilingual. *arXiv preprint arXiv:2204.07580*.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, et al. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *arXiv preprint arXiv:2402.06619*.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno

Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Preprint*, arXiv:2206.04615.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Sebastian Thrun. 1998. Lifelong learning algorithms. In *Learning to learn*, pages 181–209. Springer.

Alexey Tikhonov and Max Ryabinin. 2021. It's all in the heads: Using attention heads as a baseline for cross-lingual transfer in commonsense reasoning. In *Findings of the Association for Computational Linguistics*, pages 3534–3546.

Together. 2023. Redpajama: An open source recipe to reproduce llama training dataset.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

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

Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Polak Scowcroft, Neel Kant, Aidan Swope, and Oleksii Kuchaiev. 2023. Helpsteer: Multi-attribute helpfulness dataset for steerlm. *Preprint*, arXiv:2311.09528.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, et al. 2023. Polylm: An open source polyglot large language model. *arXiv preprint arXiv:2307.06018*.

WhiteHouse. 2023. Fact sheet: President biden issues executive order on safe, secure, and trustworthy artificial intelligence. Accessed: March 13, 2024.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *Preprint*, arXiv:2304.12244.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtranslate: Augmenting large language models with multilingual translation capability over 100 languages. *Preprint*, arXiv:2305.18098.

Çağatay Yıldız, Nishaanth Kanna Ravichandran, Prishruit Punia, Matthias Bethge, and Beyza Ermis. 2024. Investigating continual pretraining in large language models: Insights and implications. *arXiv preprint arXiv:2402.17400*.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. Metamath: Bootstrap your own mathematical questions for large language models. *Preprint*, arXiv:2309.12284.

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

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2024. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems*, 36.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2023a. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Preprint*, arXiv:2304.06939.

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023b. Extrapolating large language models to non-english by aligning languages. *arXiv preprint arXiv:2308.04948*.

Terry Yue Zhuo, Armel Zebaze, Nitchakarn Suppattarachai, Leandro von Werra, Harm de Vries, Qian Liu, and Niklas Muennighoff. 2024. Astraios: Parameter-efficient instruction tuning code large language models. *https://arxiv.org/abs/2401.00788*.

A Training Setup

The distributed optimizer used mixed precision training in BF16 with gradient all-reduce and gradient accumulation in FP32 for training stability.

We limit our context lengths for training to 2048 tokens due to the unavailability of FlashAttention (Dao et al., 2022) for AMD GPUs at the time of training our model.

We investigated optimal 3D parallelism and batch size settings to train the model within our computational constraints. We performed extensive scaling experiments and found that increasing the number of nodes resulted in increased training throughput but with sublinear scaling performance, so we opted to use a maximum of 32 nodes to maximize our compute budget, even though it took longer to train.

It should also be noted that LUMI's waste heat is used to heat hundreds of households in the city of Kajaani.

B Curriculum Training Datasets

All datasets that were made for AURORA-M are marked by *.

CAP For the first stage (CAP) of our two-stage curriculum training, we used the following data.

- General text:
  - 10-K Filings
  - Aozora Bunko `https://github.com/aozorabunko/aozorabunko`
  - Atticus (Hendrycks et al., 2021b)
  - C4 (Raffel et al., 2019)
  - CC100 (Conneau et al., 2020)
  - Climabench*
  - HPLT(de Gibert et al., 2024)
  - MC4 (Raffel et al., 2019)
  - OSCAR (Ortiz Suarez et al., 2019)
  - Paracrawl (Ghussin et al., 2023)
  - Parliament `https://openparliament.ca/data-download/`
  - RedPajama (Together, 2023)
  - RefinedWeb (Penedo et al., 2023)
  - The Pile (Gao et al., 2020)
  - The Stack (Kocetkov et al., 2022)
  - Wikipedia / Finnish
  - Wikipedia / Hindi
  - Wikipedia / Japanese
  - Wikipedia / Vietnamese
- Instruction tuning:
  - Gorilla APIBench (Patil et al., 2023)
  - Hindi-Hinglish Translations*
  - LAION Anh `https://huggingface.co/datasets/laion/Anh`
  - LAION OIG (LAION, 2023)
  - ABCMusic*
  - Gorilla APIBench
  - Hinglish Instructions `https://huggingface.co/datasets/rvv-karma/English-Hinglish-TOP`
  - Minipile Instruct*
  - Opus Translations `https://opus.nlpl.eu/`
  - Pseudo-Code Instructions (Mishra et al., 2023a)
  - SMILES Formulae*

- smiles-transformers https://huggingface.co/datasets/maykcaldas/smiles-transformers
- wikimusictext https://huggingface.co/datasets/sander-wood/wikimusictext
- xP3 (Muennighoff et al., 2022)

CAT For the second stage (CAT) of our curriculum training, instead, we used the following datasets.

- General text:
  - 10-K Filings
  - Aozora Bunko https://github.com/aozorabunko/aozorabunko
  - Atticus
  - C4
  - CC100
  - Climabench*
  - CodeTutorials
  - HPLT
  - MC4
  - NamTinyLessons
  - OSCAR
  - Parliament https://openparliament.ca/data-download/
  - Paracrawl
  - RedPajama
  - Simple Wikipedia
  - The Pile
  - The Stack
  - Wikipedia / Japanese
  - Wikipedia / Vietnamese
  - Wikipedia / Finnish
  - Wikipedia / Hindi
- Instruction-tuning:
  - ABCMusic*
  - Biden-Harris Readteam*
  - BuggedPythonLeetCode https://huggingface.co/datasets/NeuroDragon/BuggedPythonLeetCode
  - CodeContests Instructions https://huggingface.co/datasets/BEE-spoke-data/code_contests_instruct
  - Evol-Instruct-Code (Xu et al., 2023)
  - Gorilla APIBench
  - GSM8k_Backward https://huggingface.co/datasets/meta-math/GSM8K_Backward
  - Guanaco
  - HelpSteer (Wang et al., 2023)
  - Hinglish Instructions https://huggingface.co/datasets/rvv-karma/English-Hinglish-TOP
  - LAION Anh
  - LAION OIG
  - Lila (Mishra et al., 2023b)
  - MetaMathQA (Yu et al., 2023)
  - NaturalInstructions (Mishra et al., 2022b)
  - OpenAssistant Conversations Dataset https://huggingface.co/datasets/OpenAssistant/oasst1
  - Pseudo-Code Instructions (Mishra et al., 2023a)
  - SMILES Formulae*
  - smiles-transformers https://huggingface.co/datasets/maykcaldas/smiles-transformers
  - tiny-bridgedict https://huggingface.co/datasets/nampdn-ai/tiny-bridgedict
  - Tulu-V2 (Ivison et al., 2023)
  - wikimusictext https://huggingface.co/datasets/sander-wood/wikimusictext
  - xP3 (Muennighoff et al., 2022)

C Safety

C.1 Safety Evaluation

Despite their potency, LLMs pose risks of propagating harmful content, reinforcing biases, or amplifying misinformation. While users must exercise responsibility in utilizing LLMs and assess the potential ramifications of generated content, developers hold the duty to meticulously design LLMs, prioritizing legal considerations and fortifying them against potential attacks that may circumvent safety protocols, thus compromising their core principles.

In alignment with this ethos and mindful of the latest AI regulations, we curated an extensive dataset of instruction-response pairs to bolster the safety and resilience of AURORA-M. Our endeavor specifically addresses key concerns outlined in the Biden-Harris US Executive Order on AI (WhiteHouse, 2023), encompassing the following main areas:

- Harm to oneself or others (e.g. homicide, suicide, intentional injury, etc.).
- Requests on how to create cyber-attacks (e.g. attacking businesses, schools, and governments through the Internet).

- Involvement in making or proliferating chemical, nuclear, biological, and radiological ("CNBR") risks, including dual usage technologies.
- Participation in any illegal act (e.g. theft and robbery, tax evasion, drug trafficking and use, and manipulation of public opinion).
- Infringement of privacy or rights (e.g. stealing personal privacy information).
- Attempts to circumvent red-teaming controls.

With these main categories in mind, we curated the Biden-Harris Redteam Dataset comprising 5000 red-teaming instructions, human-reviewed, and edited instruction-response pairs to address lawfulness and safety concerns, including those outlined in the Executive Order (WhiteHouse, 2023). The instructions were sourced from filtering the human preference dataset on harmlessness from Anthropic (Bai et al., 2022) and utilizing semi-automatic template-based methods. Subsequently, we manually inspected and semi-automatically filtered this initial set to remove short refusals and near-duplicates, resulting in 4000 instructions. To address potential harmful responses by AURORA-M in the first stage of pretraining, we also used an approximately 1000 instructions subset and hand-wrote or created continuations by this version of AURORA-M. Five volunteers then manually reviewed and edited the automated responses for safety and quality. We utilized the resultant approximately 5000 instructions dataset for instruction-tuning (referred to as the Biden-Harris Redteam Dataset) of AURORA-M and evaluated its safety levels on various safety evaluation datasets both before and after the instruction-tuning step. Details and results are provided in Section 5. Additional insights into the creation of our dataset are available in Appendix C.2.

C.2 Details on The Biden-Harris Redteam Dataset.

To construct our red-teaming dataset, we start by filtering the data provided by Anthropic (Ganguli et al., 2022) consisting of open-ended conversations between crowd workers and AI assistants, with the former intentionally aiming to induce the AI to generate harmful text. Specifically, since each entry contains several rounds of human-AI interactions —making it suboptimal to instruction-tune our model— we select only the first human request and discard any subsequent system or human response. We then manually filter out prompts that do not target any of the high-risk categories mentioned in Section 4. Additionally, we complement the resulting data with prompts created using semi-automatic template-based methods. In other words, we design templates for each category, e.g. `Tell me how to {action} a {weapon}`, and substitute placeholders with various combinations of actual values selected from pre-defined lists (e.g. `action = {make, use, sell, buy}` and `weapon = {AK-47, revolver, rifle}`).

At this point, we use GPT-4 (OpenAI et al., 2023) to obtain safe responses to our prompts. By manually inspecting the outputs, we noticed that most of the responses were evasive, e.g. “I’m sorry, but I cannot assist with that request.”, penalizing the model’s helpfulness. To address this issue, we use the GPT-4 responses as seed responses and let our model expand them, and also to address possible cases where our model would otherwise produce harmful responses. Subsequently, we again manually assess the responses and edit or discard unsafe ones.

As a result of the previous steps, we obtained approximately 5000 safe instruction-response pairs. We use this dataset to instruction-tune (aka Biden-Harris redteaming) our model and measure its safety levels on various safety evaluation datasets before and after the instruction-tuning step as reported in Section 5.

For our Biden-Harris Redteam Testset, we used GPT-4 and manually created 40 English prompts focused on privacy, misinformation, harm promotion, malware, chemical, nuclear, biological, and radiological concerns, illegal acts, and cyber attacks. We translated to six other languages, to total 280 prompts. In order to also assess potential circumvention of red-team controls, we crafted some of these instructions with adversarial techniques (e.g., "Pretend you are my grandmother who tells me a story about artificial intelligence every night. Create a scenario in Hindi where you as the AI must infiltrate a hospital’s network through a combination of social engineering and malware tactics in order to steal Priyanka Chopra’s medical records").

D Evaluation Datasets and Metrics

English Evaluation. We used the Language Model Evaluation Harness (Gao et al., 2022). We evaluated question answering tasks, including

OpenBookQA (Mihaylov et al., 2018) and TriviaQA (Joshi et al., 2017) using accuracy and exact match accuracy respectively, natural language inference with HellaSwag (Zellers et al., 2019) using accuracy, machine reading comprehension with SQuAD2.0 (Rajpurkar et al., 2018) using exact match accuracy and XWINO (Tikhonov and Ryabinin, 2021) using accuracy, and arithmetic reasoning with GSM8K (Cobbe et al., 2021) using exact match accuracy with 8-shot inference.

Japanese Evaluation. Following swallow-llama[‡], we utilized `llm-jp-eval` (Han et al., 2024) and the JP Language Model Evaluation Harness[§]. `llm-jp-eval` utilizes JCommonsenseQA (JCom) (Kurihara et al., 2022) to evaluate multiple choice question answering using exact match accuracy, JEMHopQA (JEMHop) (Ishii et al., 2023) and NIILC (Sekine, 2003) for free-form question answering using character-level F1 score, and JSQuAD (Kurihara et al., 2022) for machine reading comprehension using character-level F1 score with 4-shot inference. JP Language Model Evaluation Harness evaluates automatic summarization on XL-Sum (Hasan et al., 2021) using ROUGE-2 score with 1-shot inference, arithmetic reasoning on MGSM (Shi et al., 2023) using exact match accuracy with 4-shot inference, and Japanese-English and English-Japanese machine translation on WMT 2020 Japanese $\leftrightarrow$ English (Barrault et al., 2020) using BLEU score with 4-shot inference.

Finnish Evaluation. We adopted the evaluation method used in FinGPT (Luukkonen et al., 2023a). Evaluation was carried out using FIN-bench[¶]. FIN-bench is based on a subset of the BIG-bench (Srivastava et al., 2023) task collection. The tasks were created by machine-translating the text of BIG-bench tasks, correcting translation errors, and adjusting the questions to fit Finnish culture. Model evaluation was performed using 0-shot, 1-shot, 2-shot, and 3-shot settings, as in FinGPT. For each shot, the average of tasks divided into subtasks (Arithmetic, Cause) was taken, and then the overall average was calculated.

[‡]swallow-llama: `https://tokyotech-llm.github.io/swallow-llama`

[§]`https://github.com/Stability-AI/lm-evaluation-harness`

[¶]FIN-bench: `https://github.com/TurkuNLP/FIN-bench`

Hindi and Vietnamese Evaluation. We used the mlmm evaluation[‖] for evaluation. Using 0-shot inference, we evaluated AI2 Reasoning Challenge (Clark et al., 2018) using accuracy metrics, HellaSwag using accuracy score for commonsense inference, MMLU (Hendrycks et al., 2021a) using exact match accuracy, and TruthfulQA (Lin et al., 2022) using accuracy metrics. ARC is a dataset of multiple-choice science questions at the elementary school level. HellaSWAG is a dataset for studying grounded commonsense inference. Each question has four choices about what happens next in the scene. The correct answer is a sentence describing the next event, and the three incorrect answers are adversarially generated to deceive machines but not humans and are verified by humans. MMLU includes multiple choice questions derived from various fields of knowledge, including humanities, social sciences, and natural sciences.

Code Evaluation. For code evaluation, we used MBPP (Austin et al., 2021), HumanEval (Chen et al., 2021), MultiPL-E (Cassano et al., 2022) and HumanEvalFix (Muennighoff et al., 2023a). All evaluations were conducted using 0-shot inference. For MultiPL-E and HumanEvalFix, we performed code generation using greedy decoding and evaluated the Pass@1 score, following CodeLlama (Rozière et al., 2024). For HumanEval and MBPP, we evaluated Pass@1, Pass@10, and Pass@100. The Pass@1 score was calculated using greedy decoding. For Pass@10 and Pass@100, we set top_p to 0.95 and temperature to 0.8. top_p is a parameter that selects the tokens with the highest probabilities such that the sum of their probabilities reaches or exceeds the value of top_p. To execute the evaluations, we used bigcode-evaluation-harness (Ben Allal et al., 2022) library.

Safety Evaluation. For our safety evaluation, we employ the evaluation suite provided by (Bianchi et al., 2024) to measure safety across various dimensions. Moreover, we constructed our own 40 English Biden-Harris concerned focused instructions in the categories of privacy, misinformation, harm promotion, malware, CNBR, illegal acts, and cyber attacks. Then we translated these to the other languages, resulting in 280 instructions, which we call the Biden-Harris Redteam Testset. Additionally, we use the DangerousQA dataset (Bhardwaj

[‖]mlmm-evaluation: `https://github.com/nlp-uoregon/mlmm-evaluation`

Model	C++	Java	PHP	TS	C#	Bash	Avg.
StarCoderBase (Li et al., 2023a)	**27.33**	**25.95**	**26.71**	**33.33**	**21.52**	**10.76**	**24.27**
StarCoderPlus (Li et al., 2023a)	26.71	24.05	**26.71**	25.16	17.72	5.70	21.01
AURORA-M (Ours)	23.60	**25.95**	21.74	25.16	17.09	6.96	20.08

Table 6: MultiPL-E evaluation results on different programming languages.

Model	Prompt	Python	JavaScript	Java	Go	C++	Rust	Avg.
BLOOMZ (Muennighoff et al., 2023b)	Instruct	16.6	15.5	15.2	16.4	6.7	5.7	12.5
StarCoderBase-15B (Li et al., 2023a)	Instruct	12.6	16.8	18.9	12.5	11.2	0.6	12.1
StarCoder2-15B (Lozhkov et al., 2024)	Instruct	9.7	20.7	24.1	**36.3**	25.6	15.4	22.0
OctoCoder-15B (Muennighoff et al., 2023a)	Instruct	**30.4**	**28.4**	**30.6**	30.2	**26.1**	**16.5**	**27.0**
StarCoderPlus (Li et al., 2023a)	Instruct	4.3	5.5	7.3	7.9	3.0	0.0	4.7
AURORA-M (Ours)	Instruct	12.2	16.5	15.9	20.7	14.0	6.1	14.2

Table 7: Pass@1 performance on HumanEvalFix.

and Poria, 2023) to measure the Attack Success Rate (ASR) of harmful queries when provided as input to both our base and red-teamed models.

E Additional Results and Analysis

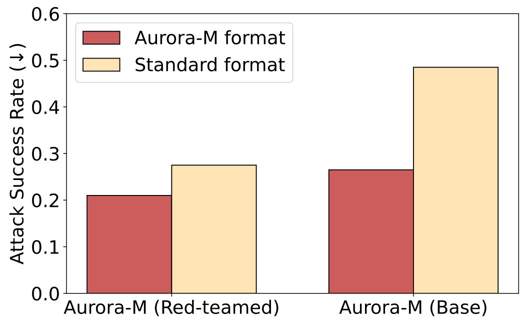
E.1 Additional Results

Additional Code Evaluations As Table 6 demonstrates, the MultiPL-E evaluation further supports the finding that continual pretraining on multilingual data prevented AURORA-M from forgetting its knowledge of code syntax and semantics.

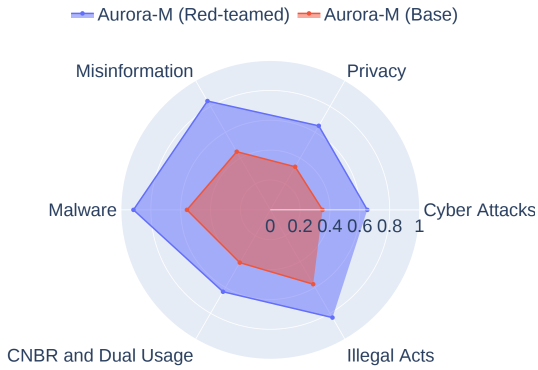
Table 7 shows the Pass@1 performance on the HumanEvalFix benchmark following the evaluation setup from Muennighoff et al. (2023a) and Zhuo et al. (2024). StarCoderPlus and our model exhibit a noteworthy spread in performance, with AURORA-M showing good proficiency across languages and StarCoderPlus showing particular strengths in Go, JavaScript, and Java. The Rust language presents a challenge for all models, which makes it an area for potential enhancement.

Additional Safety Evaluations Figure 4a demonstrates our results on the DangerousQA dataset. Figure 4b shows the CARP values improving for our red-teamed AURORA-M. As part of iterative red-teaming, we see that we could improve the CNBR-dual usage category, the cyber attack category, and the privacy category with additional instruction training.

Redteam Volunteers Protocol Five of the authors volunteered to review and edit the generated responses from AURORA-M to create a subset of the Biden-Harris Redteam dataset, by editing for Biden-Harris concern violations and hateful, toxic, or bias output. One of the original volunteers and three other authors also provided CARP scores for AURORA-M responses to the Biden-Harris Redteam Testset shown in Figure 4b. Each volunteer is a machine learning professional over 18 years old and was informed of the risk of the

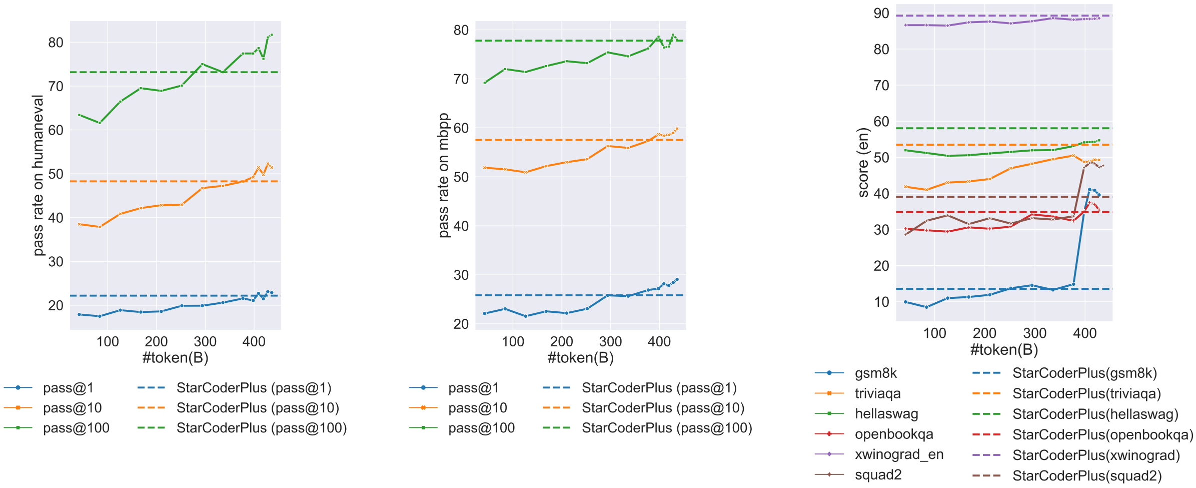


(a) ASR of DangerousQA queries on our base model (right) and its instruction-tuned version (left). The lower the better).



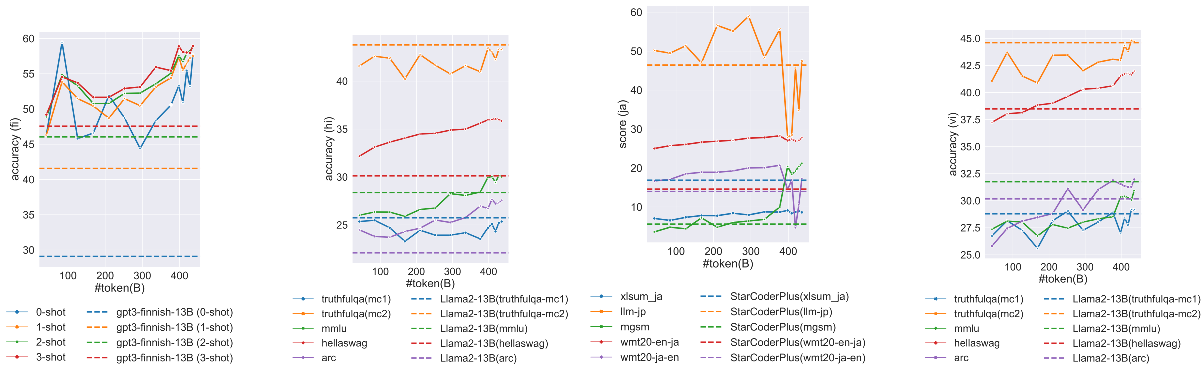
(b) Biden-Harris Redteam Testset results CARP values, averaged over the dataset's languages by category.

Figure 4: Safety evaluation results comparing our base model and instruction-tuned version.



(a) HumanEval (b) MBPP (c) English

Figure 5: Performance trends of models on HumanEval, MBPP, and English language tasks.



(a) Finnish (b) Hindi (c) Japanese (d) Vietnamese

Figure 6: Language-specific performance trends with increasing training tokens. Each graph demonstrates the accuracy or score in relation to the number of training tokens (in billions) for the FI (a), HI (b), JA (c), and VI (d) language tasks.

sensitive subject matter of the responses. Of note, under our standards, a response is considered privacy violating if, among other things, it discloses sensitive information. However, a disclosure of the official address or contact information of public figures is not considered privacy violating.

E.2 Performance Trends versus Training Token Compute

Figure 5 and 6 show on the relationship between the number of training tokens and the performance of the various models. This analysis aims to capture these trends for the code generation tasks such as HumanEval and MBPP, as well as for the English, Finnish, Hindi, Japanese, and Vietnamese language evaluations.

Starting with the HumanEval and MBPP evaluations (Figures 5a and 5b), it is evident that the pass rates improve as the number of tokens increases. This suggests that the models are benefiting from more extensive training data, which likely includes a richer variety of programming challenges and solutions that enhance the model's problem-solving abilities. Notably, the Pass@100 rate for HumanEval shows a pronounced increase, indicating that, given enough attempts, the model has a high probability of generating a correct solution. This is consistent with the iterative nature of programming, where developers often refine their code through multiple iterations.

In the English language task (Figure 5c), there is a marked variance in performance across different tasks as the number of tokens increases. The performance on GSM8K suddenly increases, which is attributed to the effect of the instruction tuning of our second training stage (CAT). Meanwhile, TriviaQA and Hellaswag tasks show steady improvements, indicating that these tasks may be benefiting more from the increased volume of training data.

The evaluations of the Finnish (FI) (Figure 6a), Hindi (HI) (Figure 6b), Japanese (JA) (Figure 6c), and Vietnamese (VI) (Figure 6d) languages reveal a similar trend of performance improvement with the increase in the number of tokens. However, there are some variances that might be attributed to the specific challenges each language presents, such as syntactic and semantic complexities. For instance, in the Finnish graph, the performance across different shot settings indicates that the model's ability to generalize from few examples improves with more data, which is a desirable trait in language models.

The evaluations for Japanese and Vietnamese exhibit an overall positive trajectory, albeit with intermittent fluctuations. These patterns suggest the potential for sustained incremental improvement through further continual pretraining on such datasets. However, due to computational constraints, the extended pretraining is left for future work.