The UNLP 2024 Shared Task on Fine-Tuning Large Language Models for Ukrainian

Oleksiy Syvokon, Mariana Romanyshyn, Roman Kyslyi
Microsoft, Grammarly, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
Kyiv, Ukraine
osyvokon@microsoft.com, mariana.romanyshyn@grammarly.com, kyslyi.roman@lll.kpi.ua

Abstract

This paper presents the results of the UNLP 2024 shared task, the first Shared Task on Fine-Tuning Large Language Models for the Ukrainian language. The goal of the task was to facilitate the creation of models that have knowledge of the Ukrainian language, history, and culture, as well as common knowledge, and are capable of generating fluent and accurate responses in Ukrainian. The participants were required to use models with open weights and reasonable size to ensure the reproducibility of the solutions. The participating systems were evaluated using multiple-choice exam questions and manually crafted open questions. Three teams submitted their solutions before the deadline, and two teams submitted papers that were accepted to appear in the UNLP workshop proceedings and are referred to in this report. The Codabench leaderboard is left open for further submissions.

Keywords: Large Language Models, LLM, Fine-Tuning, LLM Benchmarking

1. Introduction

The emergence of large language models (LLMs) marked a significant step forward in the field of natural language processing (NLP), providing a single solution for the tasks of generating human-like text. Creative writing, text evaluation, controlled text generation have suddenly become available to everyone, causing both a surge in popularity of LLM-based tools like ChatGPT (OpenAI, 2022) and discussions about the limitations and ethical implications of using them (Borji, 2023; Kocoń et al., 2023).

However, training an LLM requires significant computational resources, which may be expensive to obtain, and substantial amounts of text data, which is not readily available for most natural languages, including Ukrainian. With the UNLP 2024 shared task, our goal was to facilitate the creation of LLMs better adapted to the Ukrainian language, history, and cultural context with reasonable computational resources.

The remainder of the paper is organized as follows. Section 2 gives an overview of LLM benchmarks and methods of LLM language adaptation. Section 3 describes the UNLP 2024 shared task setup. Section 4 reviews the datasets available in this shared task. Section 5 explains how the competing systems were evaluated and ranked. Section 6 presents the results of the shared task and provides an overview of the submitted solutions. Section 7 mentions how the competing systems compare to GPT-4. Finally, Section 8 summarizes the contribution, and Section 9 provides an ethics statement.

2. Related Work

LLM Benchmarks. Evaluation methods for LLMs fall into two broad categories. Firstly, there are static, *ground-truth-based* benchmarks. These feature a predefined collection of tasks along with correct answers, and an automated metric. Such benchmarks have been the standard for assessing models before the advent of LLMs. Over time, numerous datasets of this kind have been created, and many have been adapted for LLM evaluation: GSM-8k (Cobbe et al., 2021), EXAMS (Hardalov et al., 2020), MMLU (Hendrycks et al., 2020), and AgiEval (Zhong et al., 2023), among many others. These benchmarks are cost-effective, reproducible, and can be executed automatically. However, they are restricted to a limited range of tasks and are often unsuitable for evaluating complex capabilities like open-ended text generation and subjective aspects such as humor and engagement. Consequently, these benchmarks do not fully capture the intricacies of LLM performance.

The second category involves benchmarks that measure *human preferences* for LLM-generated content. Typically, this involves a blind comparison between pairs of LLM responses to the same prompt. These comparisons are then translated into model rankings through systems such as Elo, the Bradley-Terry model, or TrueSkill (Boubdir et al., 2023; Bai et al., 2022; Bradley and Terry, 1952; Herbrich et al., 2007). Some preference-based benchmarks utilize a static set of prompts (Zheng et al., 2024; Li et al., 2023), while others permit open-ended interactions with the models (Chiang et al., 2024; Kiela et al., 2021). Recently, there has been a trend towards using advanced LLMs to

The Third Ukrainian Natural Language Processing Workshop (UNLP) @LREC-COLING-2024, pages 67–74

replace human evaluators (Li et al., 2023; Chiang and Lee, 2023; Zheng et al., 2024).

This shared task employs two complementary benchmarks: an automated metric on multiple-choice exam questions for testing LLM knowledge and human ratings for the subjective evaluation of open-ended text generation tasks.

LLM language adaptation. Despite the rapid development of open LLMs, many of these models primarily focus on English and offer limited support for other languages. A few notable exceptions (Üstün et al., 2024; Lin et al., 2021; Xue et al., 2020; Liang et al., 2023) just underscore this trends. Training a large language model from scratch demands substantial resources, making it an impractical option for many researchers. A feasible alternative is to adapt existing models to one or more languages by fine-tuning a model with a smaller set of language-specific data. This adaptation process may involve selecting a strong base LLM, curating language-specific datasets, expanding the vocabulary, conducting continual pretraining (whether full or adapter-based), translating instruction-tuning datasets, generating synthetic data, clustering languages based on their similarities, among other strategies (Lin et al., 2024; Csaki et al., 2024; Yong et al., 2022; ImaniGooghari et al., 2023; Ebrahimi and Kann, 2021; Blevins et al., 2024; Yang et al., 2023; Zhu et al., 2023).

3. Task description

The UNLP 2024 shared task required participants to fine-tune a large language model that can answer questions about the Ukrainian language, history, and culture, as well as perform text-generation tasks, all by producing fluent and factually accurate text in Ukrainian.

To ensure fair competition with reproducible results, we enforced the following limitations:

1. Only LLMs with open weights such as Llama 2 (Touvron et al., 2023), Mistral 7B (Jiang et al., 2023), Phi-2 (Javaheripi and Bubeck, 2023), Gemma (Mesnard et al., 2024), Aya 101 (Üstün et al., 2024), etc. were allowed to be used in the shared task.

2. The models had to run on GPU with 16GB VRAM and CUDA compute capacity 8.6. The type and amount of compute used for training were not limited, but the model weights and activations had to fit and stay in the GPU memory entirely. CPU memory and disk offloading were not allowed.

3. The weights of the final model had to be published on the Hugging Face Hub[1] or a similar open platform.

The participants were allowed to complement the fine-tuning with various prompting strategies, like few-shot learning or chain-of-thought reasoning, or use retrieval-augmented generation (RAG) from open data sources.

We split the evaluation of the submitted models into two tracks: multiple-choice exam questions and open questions. We provided the participants with a set of multiple-choice exam questions for training and validation and set up a Codabench[2] environment to test the systems on a hidden test set. For open questions, we shared sample questions with the participants and ran a human evaluation task to test the systems on a manually crafted test set. See Section 5 for details.

Additionally, we highly encouraged the participants to use any external data of their choice, released a script that loads the provided dataset and generates a sample prompt, and prepared sample submission files for Codabench.

4. Data

We provided the participants of the shared task with two datasets: multiple-choice exam questions and manually crafted open questions. The dataset statistics can be found in Table 1.

Task	Split	Size
Exam questions	train	3,063
	test	751
Open questions	dev	20
	test	100

Table 1: The sizes of the datasets provided in the UNLP 2024 shared task.

Both datasets can be accessed through the repository of the shared task[3].

4.1. Exam Questions

This dataset contains machine-readable questions and answers taken from the Ukrainian External Independent Evaluation[4] called ЗНО (transl: ZNO) in Ukrainian. External Independent Evaluation is a standard set of exams taken by schoolchildren in Ukraine when they apply to higher educational institutions. The dataset contains exam questions from the years 2006-2023 and covers two subjects only: History of Ukraine and Ukrainian language and literature.

[1] https://huggingface.co/

[2] https://www.codabench.org/competitions/2046/

[3] https://github.com/unlp-workshop/unlp-2024-shared-task/tree/main/data

[4] https://zno.osvita.ua/

We filtered the dataset by extracting only multiple-choice questions with one correct answer. We removed questions that referenced images (maps, portraits, photos, etc.). The final dataset was published in the .jsonl format. The training set contained 3,063 questions/answers from the years 2006-2019. The test set contained 751 questions and hidden answers, spanning the years 2020 to 2023.

4.2. Open Questions

The dataset of open questions was crafted by two native speakers of Ukrainian and comprised instruction prompts for text-generation tasks and common-knowledge question-answering. The dataset contained an equal distribution of the following:

- common knowledge questions on the topics of Ukrainian literature, music, history, geography, and culture;
- composition tasks that asked the model to write messages across a set of formality levels, lengths, and topics;
- rewrite tasks that asked the model to correct the input text, simplify it, add humor, add more details, or add emotions;
- evaluation tasks that asked the model to outline ways of improving the input text, analyze emotions in text, answer follow-up questions, brainstorm ways to complete the text, or find an odd word in a row.

The final dataset was published as a .jsonl file in the Alpaca dataset format[5]. The dev set contained 20 questions. The hidden test set contained 100 questions.

5. Evaluation

The competing LLM solutions for Ukrainian were evaluated on two hidden test sets: exam questions and open questions.

In the **first track**, we evaluated the models on a hidden test set of 751 multiple-choice exam questions, where each question had one correct answer. This setting allowed us to use accuracy as our primary metric to rank the competing LLM solutions. The registered participants submitted their system results using Codabench, which automatically compared their results with the hidden answers, returned the score, and placed the systems on the leaderboard.

In the **second track**, we evaluated the models on 100 manually crafted open questions. We asked the participants to send us their models' answers to the questions in the predefined format. We then set up a side-by-side evaluation task in the Hugging Face Spaces[6]. For that, we created a simple space with the Gradio application that displayed a question from the test set and two randomly chosen anonymized model outputs. The user could then vote which answer is better; if neither, declare a tie. The model answers to all the questions and voting logs were stored in a Firebase DB[7].

We crowdsourced annotations from 63 native speakers of Ukrainian from the Ukrainian NLP community who volunteered to join the annotation task. The annotation guidelines for the human evaluation are available in the repository of the shared task in Ukrainian and English[8].

We collected over 300 human judgments for each competing model and used the TrueSkill (Herbrich et al., 2007) ranking algorithm implemented in the trueskill Python library[9] to define the winner. TrueSkill is a statistically based ranking system for multiplayer competitions that infers the relative rankings of players based on their performance against each other. It uses the Bayesian inference algorithm to estimate each player's skill level.

6. Results and System Descriptions

A total of twenty-two teams registered for the UNLP 2024 shared task, but only three teams submitted their solutions before the deadline: Sherlock, CodeKobzar, and UkraineNow. Team Sherlock submitted two distinct solutions, each evaluated independently. Teams Sherlock and UkraineNow submitted papers that were accepted to appear in the UNLP workshop proceedings and are referred to in this report. Team CodeKobzar provided their system description by email.

We briefly review the systems here; for complete descriptions, please see the corresponding papers. Table 2 and Table 3 present the leaderboards for the two tasks.

Rank	Participant	Accuracy
1	Sherlock (RAG)	0.49
2	Sherlock (no RAG)	0.42
3	CodeKobzar	0.39
4	UkraineNow	0.23

Table 2: The **official** UNLP 2024 shared task results for the task of answering multiple-choice exam questions.

[5] https://github.com/tatsu-lab/stanford_alpaca

[6] https://huggingface.co/spaces

[7] https://firebase.google.com/

[8] https://github.com/unlp-workshop/unlp-2024-shared-task/blob/main/annotation

[9] https://github.com/sublee/trueskill

Sherlock (Boros et al., 2024) submitted winning solutions for both tasks. The team used a set of data-augmentation techniques with Mistral 7B.

Notably, the team used an array of diverse data sources for training, including Ukrainian Wikipedia, manually selected books on the target subject, and several translated datasets, both free-text and instruction-formatted. Due to the limiting factors in the standard RAG process (e.g., low performance of embeddings for Ukrainian), the team employed n-gram techniques. This method outperformed conventional similarity scoring approaches, with the LLM itself generating n-grams to enhance the retrieval process.

The tuning process began with the base Mistral 7B model, Mistral-7B-Instruct-v0.2[10]. The team experimented with standard fine-tuning on different datasets, delved into model weight merging, and leveraged direct preference optimization training to refine performance further. The availability of a test set allowed for iterative testing of various method combinations, optimizing the overall system efficacy. The team has made both the source code[11] and the model[12] publicly available.

UkraineNow (Kiulian et al., 2024) fine-tuned the open-source Gemma (gemma-2b-it[13] and gemma-7b-it[14]) and Mistral-7B-Instruct-v0.1[15] LLMs with a combination of instruction datasets, which included 10,000 rows of the UAlpaca dataset[16], 962 rows of their own UKID dataset, and 3,063 rows of the ZNO dataset provided by the organizers of the shared task. Due to resource constraints, the team chose to use the LoRA (Hu et al., 2022) fine-tuning approach, experimenting with various implementations of LoRA adapters. The team put extra effort into the quality evaluation of the models' outputs, dedicating a section of the paper to the phenomenon of code-switching, also known as Azirivka[17].

The fine-tuned gemma-2b-it model was submitted for the competition. The team has made the source code and the model available in their GitHub repository[18].

CodeKobzar[19] by Ben Ye and Mariia Ponomarenko is a large language model specifically fine-tuned for Ukrainian language, employing the Chain of LoRA technique (Xia et al., 2024) on the Vicuna-13B pretrained model[20]. The initial dataset[21] comprises articles from the Ukrainian Wikipedia, segmented into 40.6K sentences to facilitate question generation using the Mistral Medium API[22]. These sentence-question pairs served as the basis for the model's first fine-tuning phase, focusing on question generation and answering. The model was trained for one epoch with a maximum sequence length of 2,048 tokens, using the Nvidia A100 GPU. After that, the LoRA layers were integrated with the base model for a second round of training on the same dataset.

In the third fine-tuning phase, the dataset was refined to include only Ukrainian historical, literary, and cultural content, supplemented by grammatical rules from `pravopys.net`. This resulted in a new dataset of 73.6K entries[23], which was divided into prompt-question and question-response pairings. The model was then fine-tuned on the combination of the aforementioned dataset and a corpus of ZNO multiple-answer questions provided with the shared task.

This iterative approach, through the Chain of LoRA, enabled the KodKobzar model to perform an iterative low-rank residual learning procedure to approximate the optimal weight update and thereby improve the model's proficiency in grammatical accuracy and sentence construction in Ukrainian. However, since the training concentrated mainly on question generation and answering, it constrained the model's broader generative abilities, and the restricted dataset limited the model's deeper understanding of Ukrainian culture, literature, and history.

7. Comparison with GPT-4

The participants of the shared tasks were limited in the selection of models for fine-tuning with regard to their size and accessibility. These limitations were needed to ensure that the resulting solutions are reproducible and practically useful to the NLP community. However, we were curious to understand

[10] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

[11] https://github.com/adobe/sherlock-backend/tree/UNLP2024

[12] https://huggingface.co/SherlockAssistant/Mistral-7B-Instruct-Ukrainian

[13] https://huggingface.co/google/gemma-2b-it

[14] https://huggingface.co/google/gemma-7b-it

[15] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1

[16] https://github.com/robinhad/kruk

[17] https://en.wikipedia.org/wiki/Azirivka

[18] https://github.com/PolyAgent/from-bytes-to-borsch

[19] https://huggingface.co/ponoma16/KodKobzar13B

[20] https://huggingface.co/lmsys/vicuna-13b-v1.5

[21] https://huggingface.co/datasets/byebyebye/ukr-wiki-qa-v1

[22] https://mistral.ai/

[23] https://huggingface.co/datasets/byebyebye/ukr-wiki-qa-v2

Rank	Participant	TrueSkill	σ	Number of judgments
1	Sherlock (no RAG)	26.77	0.75	330
2	Sherlock (RAG)	26.27	0.74	329
3	UkraineNow	24.89	0.75	326
4	CodeKobzar	23.79	0.76	311

Table 3: The **official** UNLP 2024 shared task results for the task of answering open questions. The TrueSkill column shows the participant's rating, and σ represents the confidence of the rating.

Rank	Participant	TrueSkill	σ	Number of judgments
1	GPT-4	28.48	0.79	474
2	Sherlock (no RAG)	25.05	0.75	462
3	Sherlock (RAG)	24.14	0.76	439
4	UkraineNow	23.16	0.76	455
5	CodeKobzar	22.17	0.77	414

Table 4: The **non-official** UNLP 2024 shared task results for the task of answering open questions. Here, GPT-4 is included for comparison. The TrueSkill column shows the participant's rating, and σ represents the confidence of the rating.

how these fine-tuned open solutions compare to the proprietary OpenAI models, in particular gpt-4-0613[24], which was the latest OpenAI GPT version at the time when the shared task started.

For the exam task, we used a very simple prompt[25] and the default parameters of GPT-4. The model managed to achieve an accuracy of 0.61. We also ran GPT-4 on the open questions task and included the responses in the human evaluation. Table 4 shows the gap between the winning open-source solution and GPT-4.

This experiment set an ambitious goal for the next iteration of this shared task.

8. Conclusion

We believe that the UNLP 2024 shared task was instrumental in facilitating research on fine-tuning large language models for the Ukrainian language, and we hope that the insights from the teams' research will be useful to the NLP community. All the datasets used in the shared task are available on GitHub, and the competing systems were openly published, which contributes to the reproducibility of the shared task results and the creation of more accessible LLMs.

The best-performing systems were submitted by team Sherlock, scoring 49% accuracy on the exam task (with RAG) and 26.77% rating on the open question task (without RAG). The Codabench environment remains open for further submissions, although any such submissions will be considered outside of the UNLP 2024 competition.

The open LLMs used in the shared task included Mistral 7B, Gemma, and Vicuna-13B. All system descriptions mention the scarcity of open datasets for the task at hand and show the creativity of the researchers in creating new datasets.

In the next iterations of this shared task, we plan to increase the size and variability of the test sets and introduce automated metrics for the open question evaluation, in addition to human evaluation.

9. Ethics Statement

To make sure that the participants of the shared task have equal opportunities and that the resulting solutions can be used by the research community, the organizers of the shared task set strict limitations on the size and accessibility of the models that were allowed in the competition.

Upon entering the competition, all participants of the shared task accepted the following terms and conditions:

- All participants agree to compete in a fair and honest manner in the shared task and not use any illegal, malicious, or otherwise unethical methods to gain an advantage in the shared task.
- All participants agree to not distribute or share the test data obtained during the shared task with any third parties.
- All participants agree to make their solutions publicly available upon the completion of the shared task in order to facilitate knowledge sharing and developments of the Ukrainian language.

[24] https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo

[25] https://github.com/unlp-workshop/unlp-2024-shared-task/blob/main/examples/random_baseline.py

To the best of our knowledge, the shared task participants followed these terms and conditions.

10. Acknowledgements

We extend special gratitude to Nataliia Romanyshyn for her invaluable contribution to the development of the open questions set. We are grateful to the participants of the Ukrainian NLP community who volunteered to help with side-by-side evaluation.

11. Bibliographical References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A Smith, and Luke Zettlemoyer. 2024. Breaking the curse of multilinguality with cross-lingual expert language models. *arXiv preprint arXiv:2401.10440*.

Ali Borji. 2023. A categorical archive of chatgpt failures.

Tiberiu Boros, Radu Chivereanu, Stefan Dumitrescu, and Octavian Purcaru. 2024. Fine-tuning and retrieval augmented generation for question answering using affordable large language models. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop*, Torino, Italy. European Language Resources Association.

Meriem Boubdir, Edward Kim, Beyza Ermis, Sara Hooker, and Marzieh Fadaee. 2023. Elo uncovered: Robustness and best practices in language model evaluation. *arXiv preprint arXiv:2311.17295*.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

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

Zoltan Csaki, Bo Li, Jonathan Li, Qiantong Xu, Pian Pawakapan, Leon Zhang, Yun Du, Hengyu Zhao, Changran Hu, and Urmish Thakker. 2024. Sambalingo: Teaching large language models new languages.

Abteen Ebrahimi and Katharina Kann. 2021. How to adapt your pretrained multilingual model to 1600 languages. *arXiv preprint arXiv:2106.02124*.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. Exams: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. *arXiv preprint arXiv:2011.03080*.

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

Ralf Herbrich, Tom Minka, and Thore Graepel. 2007. Trueskill(tm): A bayesian skill rating system. In *Advances in Neural Information Processing Systems 20*, pages 569–576. MIT Press.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André FT Martins, François Yvon, et al. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. *arXiv preprint arXiv:2305.12182*.

Mojan Javaheripi and Sebastien Bubeck. 2023. Phi-2: The surprising power of small language models. *Microsoft Research Blog*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, and et al. 2023. Mistral 7b.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. 2021. Dynabench: Rethinking benchmarking in nlp. *arXiv preprint arXiv:2104.14337*.

Artur Kiulian, Anton Polishko, Mykola Khandoga, Oryna Chubych, Jack Connor, Raghav Ravishankar, and Adarsh Shirawalmath. 2024. From bytes to borsch: Fine-tuning gemma and mistral for the ukrainian language representation. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop*, Torino, Italy. European Language Resources Association.

Jan Kocoń, Igor Cichecki, and et al. Kaszyca. 2023. Chatgpt: Jack of all trades, master of none. *Information Fusion*, 99:101861.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. Xlm-v: Overcoming the vocabulary bottleneck in multilingual masked language models.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André FT Martins, and Hinrich Schütze. 2024. Mala-500: Massive language adaptation of large language models. *arXiv preprint arXiv:2401.13303*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

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

OpenAI. 2022. Introducing chatgpt.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models.

Hugo Touvron, Louis Martin, Kevin Stone, and et al. 2023. Llama 2: Open foundation and fine-tuned chat models.

Wenhan Xia, Chengwei Qin, and Elad Hazan. 2024. Chain of lora: Efficient fine-tuning of language models via residual learning.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

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

Zheng-Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, et al. 2022. Bloom+ 1: Adding language support to bloom for zero-shot prompting. *arXiv preprint arXiv:2212.09535*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao

Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

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

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model.