Tagengo: A Multilingual Chat Dataset

Peter Devine
Lightblue Inc. (Tokyo, Japan)
peter@lightblue-tech.com

Abstract

Open source large language models (LLMs) have shown great improvements in recent times. However, many of these models are focused solely on popular spoken languages.

We present a high quality dataset of more than 70k prompt-response pairs in 74 languages which consist of human generated prompts and synthetic responses. We use this dataset to train a state-of-the-art open source English LLM to chat multilingually.

We evaluate our model on MT-Bench chat benchmarks in 6 languages, finding that our multilingual model outperforms previous state-of-the-art open source LLMs across each language. We further find that training on more multilingual data is beneficial to the performance in a chosen target language (Japanese) compared to simply training on only data in that language.

These results indicate the necessity of training on large amounts of high quality multilingual data to make a more accessible LLM.

1 Introduction

Recently, open source large language models (LLMs) have grown drastically in both popularity and performance. Models such as Llama 3 (AI@Meta, 2024b) have exceeded the performance of previous state-of-the-art proprietary models like GPT3.5 (Ouyang et al., 2022) on popular robust benchmarks including the Chatbot Arena leaderboard (Chiang et al., 2024). These open source LLMs are also increasingly being used in commercial AI chat products such as the Meta AI assistant (AI@Meta, 2024a).

However, many current LLMs exhibit lower performance on languages outside of English (Achiam et al., 2023). Indeed, Llama 3 itself is currently an English-only LLM, meaning that even when it is prompted in a language besides English, it often replies in English. This limits the potential user base of these LLMs due to the fact that less than 1.5 billion of the world's more than 8 billion population can speak English (Central Intelligence Agency, 2021; Eberhard et al., 2024). Therefore, we set out to train a state-of-the-art open source LLM (Llama 3) to be able to chat not only in English, but in many languages.

In order to make English-focused LLMs accessible in other languages, previous work has fine-tuned these models on non-English data (Sasaki et al., 2023; Sengupta et al., 2023; Nguyen et al., 2023).

Many multilingual chat datasets such as MultiAlpaca (Wei et al., 2023) and Aya (Singh et al., 2024) cover many languages and tasks but can also lack natural prompts and high quality responses.

For this reason, we created a large, diverse, high quality multilingual dataset using more than 70k human generated prompts in 74 languages and generated responses from these using state-of-the-art proprietary chat models. We used this dataset to train two models, a multilingual LLM and a Japanese-only LLM, both supervised fine-tuned models based on the Llama 3 8B Instruct model.

We found that our model achieved better evaluation scores on multilingual chat benchmarks compared to the similarly sized state-of-the-art open source models, indicating the high quality and diversity of our training dataset. We also find that our multilingual-trained LLM performs better on Japanese chat benchmarks compared to our Japanese-only-trained LLM, indicating that transfer learning from training on other languages is beneficial for training even monolingual models outside of English.

Our findings combine to inform the community of exactly how to fine-tune monolingual LLMs to create a strong multilingual model.

We make our training data (Tagengo)[1], train-

[1] https://huggingface.co/datasets/lightblue/tagengo-gpt4

ing code[2], evaluation benchmark (multilingual MT-Bench)[3], and trained models (Suzume)[4][5] publicly available for free use online.

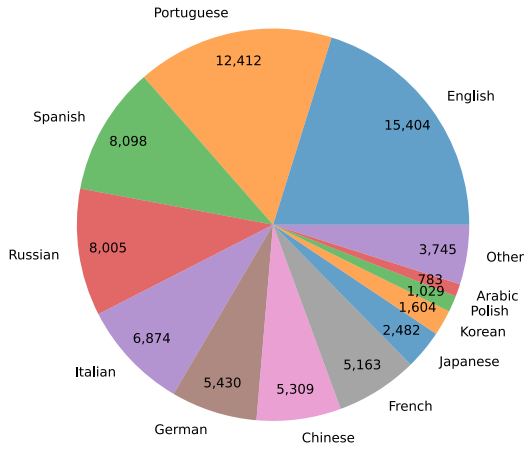

Figure 1: Distribution of the languages found in the Tagengo dataset

2 Related Work

In the literature, strong foundation models such as Llama 2 (Touvron et al., 2023) and Gemma (Team et al., 2024) have been subsequently fine-tuned on data from a specific language or languages, including Japanese (Sasaki et al., 2023), Arabic (Sengupta et al., 2023), and South-East Asian languages (Nguyen et al., 2023). Fine-tuning has often shown to improve the accuracy of the resultant LLM on tasks in that language. However, the training dataset of these models are often not shared, making it difficult to create a truly multilingual LLM across many languages.

Some multilingual chat datasets do exist that can be used for training LLMs. MultiAlpaca (Wei et al., 2023) is a multilingual dataset of 133K prompt-response pairs covering 11 languages that were generated in a similar manner to Alpaca (Taori et al., 2023). This dataset was created by generating synthetic prompts from a small number of English seed prompts and then answering these prompts using an large-scale LLM, GPT3.5 (Ouyang et al., 2022). Because these prompts are generated synthetically, this data may not reflect the sorts of prompts that real users may use with an LLM, potentially limiting the ability of models trained on this data to be used practically. Moreover, the prompts and responses for this dataset were generated using GPT3.5, meaning that the quality of the data may not be as high as if a state-of-the-art LLM was used, like GPT4 (Achiam et al., 2023).

xP3 (Crosslingual Public Pool of Prompts) (Muennighoff et al., 2022) is a dataset of more than 78 million examples covering 46 languages. This dataset was generated by templating other datasets (e.g. translation datasets, classification datasets) into a prompt-response format. While this dataset is large, the templating process limits the usefulness of the dataset as it results in prompts that are not necessarily similar to what an actual user of an LLM would ask. The templating process can also result in unnatural answers, with single word answers being given where a fuller answer may be more appropriate from an LLM.

Aya (Singh et al., 2024) is a dataset of 204k human-annotated prompt-completion pairs covering 65 languages. The majority of this dataset was generated by first translating and templating datasets from various languages, which were then corrected and annotated by human labellers. While the human labelling process will prevent as many unnatural utterances enter the dataset, the templating of datasets means that the prompts will still not necessarily be the kind of prompts that an end-user of LLMs would use. Hence, the usefulness of this dataset in training multilingual LLMs is limited by its data-generation process.

The ShareGPT dataset used by models such as Vicuna (Chiang et al., 2023) and OpenChat (Wang et al., 2023) contain approximately 70k open source conversations between users and GPT3.5 (Ouyang et al., 2022) and 6k conversations between users and GPT4 (Achiam et al., 2023), meaning that the prompts used in these datasets are often much more naturalistic to a real LLM use-case. However, the majority of these prompts are in English, meaning that this dataset is limited in its use in training multilingual models. Moreover, due to the fact that that majority of this dataset contains data generated from GPT3.5, its usefulness in training is limited as many other models have now surpassed the performance of GPT3.5 in English (Zhu et al., 2023; AI@Meta, 2024b). The amount of multi-

[2] https://github.com/lightblue-tech/suzume/tree/main/tagengo

[3] https://github.com/lightblue-tech/multilingual-mt-bench

[4] https://huggingface.co/lightblue/suzume-llama-3-8B-multilingual

[5] https://huggingface.co/lightblue/suzume-llama-3-8B-japanese

lingual data in the higher quality GPT4 subset of the ShareGPT dataset is small, meaning that its usefulness in training is constrained by its size.

To address the shortcomings in existing public datasets, we created a large, diverse, high quality multilingual dataset using more than 70k human generated prompts in 74 languages and generated responses from these using state-of-the-art proprietary chat models.

3 Method

In this section, we detail how we generated our training dataset, our training method, and finally our evaluation techniques.

3.1 Tagengo Dataset Creation

First, to generate our dataset, we sampled prompts from the million row LMSYS-Chat-1M dataset (Zheng et al., 2023). These prompts were collected from users speaking to one of 25 LLMs on the Vicuna demo and Chatbot Arena website[6].

We cleaned this dataset by first removing all prompts which contain an OpenAI Moderation Endpoint[7] flag in order to remove explicit, sexual, or illegal material.

We then removed all prompts which were listed as a non-recognised or fictional language (unknown, Klingon, xx, zp, and zzp).

We removed any prompts which contained the string "name" when lower-cased, as NAME0, NAME1 etc. was used as the placeholders for anonymised material. Effectively, this removed any anonymised prompts from our dataset.

We then removed any prompts which contained the following keywords: "gpt", "vicuna", "alpaca", "llama", "koala", "claude", "guanaco". This was done to remove prompts which referred explicitly to the LLMs that were being tested in the Chatbot Arena as many prompts asked about the LLM specifically, which we theorize is less useful in a more general context.

We then used the FastText (Joulin et al., 2016) LangDetect library[8] to determine the confidence level of classifying a particular language. We filtered out all prompts in which the confidence level of the language indicated in the original LMSYS-Chat-1M paper was less than 80%. This was done to filter out ambiguous language examples, as we later sample per-language.

Finally, we analysed the number of tokens of both the first prompt and LLM response, and removed any prompts in which the combined token total of the first prompt and LLM response amounted to more than 512 tokens. This was done to prevent very long prompts or prompts which elicited very long responses being used in our dataset in order to minimise costs when generating data with these prompts using GPT4.

We then sampled a maximum of 25,000 prompts from each language, which effectively meant we sampled the English prompts in this dataset as only English (380,138) had more than 25,000 examples, while the next most popular language Chinese (21,057) had less than 25,000. This was done to counteract the outweighed prevalence of English within this dataset.

For each language, we then embedded each prompt using the BGE M3 embedding model (Chen et al., 2024), which is a state-of-the-art embedding model that supports more than 100 languages. We then compared the prompt embeddings pairwise using the dot product to obtain a similarity score for each prompt pair. We perform fuzzy de-duplication by removing one of any prompt pairs which have a similarity score of greater than 0.8 in order to bolster the diversity of our dataset. The amount of data removed from each language varied widely with languages such as Chinese having a very high rate of de-duplication ($\sim$75%) and other such as Portuguese having a lower rate of de-duplication ($\sim$40%). This may be due to the biases of the embedding model or due to the kind of prompts submitted to the original dataset in different languages.

A table of the number of prompts filtered at each stage of our cleaning process can be found in Table 1.

We used these prompts to generate responses using an Azure OpenAI deployment of a state-of-the-art proprietary LLM, GPT4 (0125-Preview), with the generation temperature set to 0 and setting a maximum number of response tokens to be 2,048.

Due to the fact that generating high quality responses for all of these prompts manually for each language would be prohibitively expensive, we decided to generate these responses using a state-of-the-art model. We hypothesize that using an LLM much larger - rumoured to be 1.8 trillion parame-

[6] https://chat.lmsys.org/

[7] https://platform.openai.com/docs/guides/moderation

[8] https://github.com/zafercavdar/fasttext-langdetect

Stage	Number of prompts
Start	1,000,000
OpenAI Moderation check	964,464
Remove unknown languages	936,468
Remove anonymised data	753,731
Remove references to models	735,390
Language detection confidence score >80%	556,368
Remove prompt plus responses with more than 512 tokens	513,011
Random sampling of 25,000 prompts per language	157,873
Fuzzy de-duplication	78,057
Remove uncompleted/ unanswered prompts	76,338

Table 1: Table describing the number of prompts after each cleaning stage.

ters (Schreiner, 2023) - than nearly all other open source models to generate responses will lead to high quality responses that can then be used to improve existing open source models. When viewed in this way, this training can be viewed as a form of model distillation (Buciluǎ et al., 2006; Hinton et al., 2015).

We finally removed any responses which GPT4 did not answer or was not able to complete within the 2,048 token limit. The number of prompts in our resultant Tagengo dataset can be found in Table 1 and a breakdown of the prompts by language can be found in Fig 1.

We share our dataset creation code and training dataset on Huggingface[9].

3.2 Training

For training data, we add two more datasets to the Tagengo dataset which we regard as high quality chat datasets. The first is the Megagon Instruction dataset (Hayashibe, 2023), a manually annotated dataset of 669 Japanese prompt-response pairs. The second is the 6k GPT4 subset of the ShareGPT dataset[10], which has a majority of prompts in English but also includes responses in other languages. We combined and randomly shuffled these three datasets to use as a 83,213 prompt-response pair training dataset for the multilingual model.

We used our training data to train a Llama 3 8B Instruct model[11] with the Axolotl LLM training package[12]. We trained for one epoch using full fine tuning, using sample packing (Brown et al., 2020) and a context length of 8,096. We name this model Suzume 8B multilingual and the full training configuration for this model can be found on our model card[13].

We also prepared a subset of the above three datasets that only included Japanese data from each dataset, which amounted to 3,318 prompt-response pairs. This was prepared to isolate the effect of monolingual training compared to multilingual training on our data. We trained our model in the same manner as the multilingual model with the name Suzume 8B Japanese. Full details for how the training was conducted can be found on our model card[14].

3.3 Evaluation

We tested our models by using a forked version of the original MT-Bench evaluation suite (Zheng et al., 2024). The MT-Bench evaluation benchmark is a set of prompts and responses in English that cover 8 broad categories of prompts: writing, role-play, extraction, reasoning, math, coding, STEM knowledge, and humanities knowledge. Responses to these prompts are generated using an LLM, and those responses are then evaluated using an evaluation model such as GPT4.

We added publicly available translated versions of the original MT-Bench dataset in Chinese, French, German, Japanese, and Russian that had been human-verified by a native speaker of that language.

Note that the Russian translation did not contain reference answers for the math, coding, and reasoning questions, so our evaluation does not include math, coding, and reasoning problems in Russian.

Finally, we added the phrase "Your evaluation

[9] https://huggingface.co/datasets/lightblue/tagengo-gpt4

[10] https://huggingface.co/datasets/openchat/openchat_sharegpt4_dataset/blob/main/sharegpt_gpt4.json

[11] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[12] https://github.com/OpenAccess-AI-Collective/axolotl

[13] https://huggingface.co/lightblue/suzume-llama-3-8B-multilingual

[14] https://huggingface.co/lightblue/suzume-llama-3-8B-japanese

	Llama 3 8B Instruct	Suzume 8B multilingual	Suzume 8B Japanese	Starling 7B beta	GPT3.5 Turbo
Chinese	-	7.11	-	6.97	7.55
French	-	7.66	-	7.29	7.74
German	-	7.26	-	6.99	7.68
Japanese	-	6.56	6.24	6.22	7.84
Russian	-	8.19	-	8.28	7.94
English	7.98	7.73	-	7.92	8.26

Table 2: Average MT-Bench scores across 6 languages for each LLM evaluated.

should also consider whether the prompt responded in the correct language and the fluency and naturalness of this response." to the original MT-Bench evaluation criteria to ensure that the LLM judge would not simply evaluate factually correct responses in English to non-English prompts as correct. We conducted these evaluations using the "gpt-4-turbo model" from OpenAI as the judge LLM.

We make our evaluation code freely available online[15].

As baselines, we also evaluate the original Llama 3 8B Instruct model (AI@Meta, 2024b), GPT3.5-Turbo (Ouyang et al., 2022), and the Starling 7B Beta (Zhu et al., 2023) which is the highest rated similarly sized multilingual model on the Chatbot Arena leaderboard (Chiang et al., 2024) and has been trained on the ShareGPT dataset amongst other data.

4 Results

The MT-Bench scores for each model evaluated can be found in Table 2.

We first found that we were able to train Llama 3 8B Instruct to output responses in the same language as the prompt. This means that we achieved our base objective of enabling a monolingual model (Llama 3) to be able to output multilingual chat.

Secondly, English performance of the multilingual trained model only dropped slightly compared to the base Llama 3 8B Instruct model. This indicates that English chat performance does not considerably drop even when training on a majority of non-English data.

Thirdly, we found that the multilingual trained model performs better compared to the Starling 7B Beta across 5 out of 6 non-English languages tested. However, also we found that our multilingual model achieved lower evaluation scores compared to the proprietary GPT3.5 on 5 of 6 non-English languages. This indicates that our model has achieved state-of-the-art performance in multilingual chat for open-source models of its size, but has not achieved state-of-the-art performance more generally.

Finally, the Suzume 8B multilingual model achieves higher MT-Bench scores on the Japanese benchmark compared to the Suzume 8B Japanese model, indicating that transfer learning from training on other languages is beneficial for training even monolingual models outside of English.

5 Discussion

Our results indicate the need for large, high quality, multilingual datasets when training multilingual models. We find that with such a dataset, we can train a state-of-the-art monolingual model such as Llama 3 to achieve state-of-the-art multilingual performance.

We also found that training on additional non-Japanese data improves the performance of our LLM on Japanese benchmarks when compared to training solely on Japanese data, indicating that there is a collective improvement effect between languages when training using multilingual data. This adds to the body of work that indicates that training on multiple languages enables the LLM to better generalise to other languages (Nguyen and Chiang, 2017; Schuster et al., 2018). This suggests that generating an even larger, more diverse dataset in the future could further aid the performance of LLMs on low-resource languages.

[15] https://github.com/lightblue-tech/multilingual-mt-bench

6 Future Work

Our work could be built upon and improved in the following ways.

Our training dataset mainly consisted of single prompt-response pairs, but many chats between users and LLMs extend beyond a single conversation turn. Therefore, future work could include creating a dataset that contains multiple turns of conversation, with the prompts either generated by humans or by high quality LLMs.

Future work could also include adding more languages to our dataset. Our dataset only included 74 languages, and crucially omits any languages in the Niger–Congo language family, one of the most diverse language families in the world (Good, 2017). Therefore, future work could involve sampling initial prompts from a wider range of sources (possibly by advertising free chatbot access to people in areas with many speakers of underrepresented languages) and generating responses based on these prompts. This would help to both improve an LLMs linguistic understanding of these low-resource languages as well as improve their understanding of the topics and questions that people from that language and culture may ask.

Finally, future work could include generating preference data, such as was done in English in the Nectar dataset (Zhu et al., 2023), for use with contrastive learning techniques such as Direct Preference Optimisation (Rafailov et al., 2024) and Odds Ratio Preference Optimisation (Hong et al., 2024). These techniques have been shown to further improve the accuracy of LLMs, suggesting that training using these techniques may also improve the performance of LLMs in multilingual chat.

7 Conclusion

In this study, we successfully trained a state-of-the-art monolingual Llama 3 LLM to chat multilingually using a new, diverse dataset comprising over 70k human-generated prompts in 74 languages paired with high-quality synthetic responses.

Our multilingual model showcased superior performance across multiple languages compared to similar-sized open-source models on various chat benchmarks.

Interestingly, training using a multilingual dataset also enhanced the performance on specific monolingual tasks, implying beneficial cross-linguistic transfer effects.

These outcomes underline the importance of using rich, diverse multilingual data for improving the capabilities of LLMs in global, multilingual applications.

Limitations

The three main limitations of this paper concern our prompt diversity, our data generation methodology, and our model evaluation methodology.

Firstly, as stated in Section 6, our training data has a paucity of low-resource languages represented within it. While we try to focus on non-English data in our work by sampling a maximum of 25,000 prompts per language, this still does not counteract the fact that the prompts in the LMSYS-Chat-1M dataset (Zheng et al., 2023) are disproportionately from a small set of languages. These prompts are collected from users on the Chatbot Arena LLM demo site, meaning that the speakers of low-resource languages may be too few, unable, unaware, or unwilling to talk to an LLM chatbot in their native language. This means that current open source LLMs will continue to have lower performance on low-resource languages if this problem is not resolved.

Secondly, we generate our responses to prompts using GPT4, which means that all training data will be in the worldview and within the domain of knowledge that GPT4 exhibits. This biases the model as many LLMs have been shown to have both political (Feng et al., 2023) and cultural biases (Cao et al., 2023) in the text they generate, meaning that what may be deemed acceptable by one user may not be deemed acceptable by another. Moreover, while GPT4 is state-of-the-art and has been shown to generate more accurate information compared to previous models (Achiam et al., 2023), it is still capable of generating incorrect data in response to a prompt, meaning that our training data may contain incorrect statements or otherwise inaccurate data.

Thirdly, we compare our Suzume model results to the Starling LLM (Zhu et al., 2023), with the former being an 8 billion parameter model while the latter is a 7 billion parameter model. This makes for a somewhat unfair comparison as our model is larger than previous open source multilingual models. This was done as the 8 billion parameter size of LLMs was somewhat novel at the time of release, meaning that we did not have a perfect comparison to previous state of the art open source

models. However, future work could isolate the effect of training on the Tagengo dataset by training an existing multilingual model and then comparing the base model to the trained model.

Finally, our evaluation methodology is biased by the fact that our 6 evaluation languages are all within the top 10 most popular languages in our training data. This means that our evaluation does not consider the performance of our models on low resource languages, limiting the usefulness of our results to speakers of low resource languages.

Ethics Statement

Due to the potential for LLMs to be misused for unethical purposes (Derner and Batistič, 2023; Zhuo et al., 2023), we considered the ethical implications of releasing both the training data and final trained model of this work. However, since our training data was made up of human-generated content that was already publicly available, and the synthetic parts of our dataset were generated using a readily available LLM (GPT-4), we consider that the increase in risk profile with our releasing this dataset is marginal. Likewise, due to state-of-the-art models such as GPT-4 being readily available to the public, we believe the increase in risk profile from our model release is similarly minimal.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI@Meta. 2024a. [link].

AI@Meta. 2024b. Llama 3 model card.

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.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *arXiv preprint arXiv:2303.17466*.

Central Intelligence Agency. 2021. *The World Factbook 2021*. Central Intelligence Agency, Washington, DC.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

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

Erik Derner and Kristina Batistič. 2023. Beyond the safeguards: exploring the security risks of chatgpt. *arXiv preprint arXiv:2305.08005*.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2024. *Ethnologue: Languages of the World*, twenty-seventh edition. SIL International, Dallas, Texas. Online version: `http://www.ethnologue.com`.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. *arXiv preprint arXiv:2305.08283*.

Jeff Good. 2017. Niger-congo languages. *The Cambridge handbook of areal linguistics*, pages 471–499.

Yuta Hayashibe. 2023. megagonlabs/instruction_ja: Japanese instructions data for llm.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Reference-free monolithic preference optimization with odds ratio. *arXiv preprint arXiv:2403.07691*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

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*.

Toan Q Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. *arXiv preprint arXiv:1708.09803*.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2023. Seallms - large language models for southeast asia.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Akira Sasaki, Masato Hirakawa, Shintaro Horie, and Tomoaki Nakamura. 2023. Elyza-japanese-llama-2-7b.

Maximilian Schreiner. 2023. Gpt-4 architecture, datasets, costs and more leaked.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2018. Cross-lingual transfer learning for multilingual task oriented dialog. *arXiv preprint arXiv:1810.13327*.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Alham Fikri Aji, Zhengzhong Liu, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models.

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*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*.

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*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric. P Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2023. Lmsys-chat-1m: A large-scale real-world llm conversation dataset.

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.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, Karthik Ganesan, Wei-Lin Chiang, Jian Zhang, and Jiantao Jiao. 2023. Starling-7b: Improving llm helpfulness & harmlessness with rlaif.

Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Red teaming chatgpt via jailbreaking: Bias, robustness, reliability and toxicity. *arXiv preprint arXiv:2301.12867*.