Babel: Open Multilingual Large Language Models Serving Over 90% of Global Speakers

Yiran Zhao Chaoqun Liu Yue Deng Jiahao Ying Mahani Aljunied Zhaodonghui Li
Lidong Bing Hou Pong Chan Yu Rong Deli Zhao Wenxuan Zhang[†]

DAMO Academy, Alibaba Group

{zhaoyiran.zyr, royrong.ry}@alibaba-inc.com, wxzhang@sutd.edu.sg

Project page: https://babel-llm.github.io/babel-llm/

*People built the **Tower of Babel** to reach heaven and achieve unity,*
but God confused their language and scattered them across the earth.

– Story from Genesis, Old Testament

Abstract

Large language models (LLMs) have transformed natural language processing (NLP), yet open-source multilingual LLMs remain scarce, with existing models often limited in language coverage. Such models typically prioritize well-resourced languages like French and German, while widely spoken but under-resourced languages such as Hindi, Bengali, and Urdu are overlooked. To address this disparity, we introduce `Babel`, a multilingual LLM that covers the top 25 languages by number of speakers, supports over 90% of the global population, and includes many languages neglected by other open multilingual LLMs. Unlike traditional continue pretraining approaches, Babel expands its parameter count through a layer extension technique that elevates Babel's performance ceiling. We introduce two variants: `Babel-9B`, designed for efficient single-GPU inference and fine-tuning, and `Babel-83B`, which sets a new standard for open multilingual LLMs. Extensive evaluations on multilingual tasks demonstrate its superior performance compared to open LLMs of comparable size. In addition, using existing supervised fine-tuning datasets, Babel achieves remarkable performance, with Babel-9B-Chat leading among 10B-sized LLMs and Babel-83B-Chat setting a new standard for open LLMs, performing comparably to GPT-4o on certain tasks.

1 Introduction

Large language models (LLMs) (Achiam et al., 2023; Reid et al., 2024; Dubey et al., 2024; Team et al., 2024) have revolutionized the field of natural language processing (NLP), emerging as powerful tools that drive innovation and improve various aspects of human life (Li et al., 2023a; Roziere et al., 2023; Li et al., 2023b; Yang et al., 2024; Alves et al., 2024). However, multilingual LLMs remain relatively rare, particularly in the open-source domain (Hurst et al., 2024; Anthropic, 2024), where the supply of such models falls short of the growing demand for broader language support. Even among the existing open-source multilingual LLMs, the range of languages they support is often constrained. These models, such as

[†]Wenxuan Zhang is the corresponding author.

GLM-4 (GLM et al., 2024) and Qwen2.5 (Qwen et al., 2025), tend to prioritize languages with extensive training resources—typically those spoken in developed countries, such as French, Arabic, and German, where high-quality datasets are readily available. In contrast, languages spoken in less developed regions, such as Hindi, Bengali, and Urdu—despite having millions of speakers, often outnumbering those of French or German (Eberhard et al., 2024)—receive considerably less attention.

To bridge this gap and make LLMs more accessible to a wider global audience, we introduce `Babel` - a new open-source multilingual large language model, aiming to serve over 90% of speakers worldwide. Specifically, we focus on the top 25 languages by number of speakers, including English, Chinese, Hindi, Spanish, Arabic, French, Bengali, Portuguese, Russian, Urdu, Indonesian, German, Japanese, Swahili, Filipino, Tamil, Vietnamese, Turkish, Italian, Javanese, Korean, Hausa, Persian, Thai, and Burmese. Notably, more than half of these languages, despite being widely spoken, have been largely neglected by existing open-source multilingual LLMs. Given the limited availability of high-quality training data for many of these languages, we place significant emphasis on optimizing the data-cleaning pipeline to ensure the highest possible data quality. To this end, we collect data from diverse sources and employ an LLM-based quality classifier to curate clean, high-quality content for training.

Unlike conventional continue pretraining approaches (Nguyen et al., 2023b; Zhao et al., 2024a), we improve Babel's performance ceiling by increasing its parameter space through model expansion. Specifically, we employ layer extension, a structured approach that adds new layers identical in architecture to the original ones. Balancing accessibility and state-of-the-art performance, we present two model variants: `Babel-9B` and `Babel-83B`. `Babel-9B` is optimized for open multilingual LLM inference and fine-tuning on a single GPU, while `Babel-83B` establishes a new benchmark as the leading open multilingual LLM. A comprehensive evaluation on multilingual datasets highlights Babel's superior performance compared to open LLMs of a similar size. Furthermore, due to constraints on both human and computational resources, we leverage open-source supervised fine-tuning (SFT) datasets—WildChat (Zhao et al., 2024b) and Everything Instruct Multilingual (rombodawg, 2025)—to construct an SFT training pool of 1 million conversations without creating additional training data. This pool is then used to train Babel-9B-Base and Babel-83B-Base. Surprisingly, with limited training, Babel chat models demonstrate strong task-solving capabilities. Notably, Babel-9B-Chat achieves state-of-the-art performance among 10B-sized LLMs, while Babel-83B-Chat sets a new benchmark for open LLMs and even performs comparably to state-of-the-art commercial models such as GPT-4o on certain tasks.

2 Languages

Current LLMs increasingly support non-English languages such as GLM-4 (GLM et al., 2024) and Qwen2.5 (Qwen et al., 2025). However, these models primarily focus on languages with extensive training corpora, which are often spoken in developed countries, such as French and German, where numerous research institutions curate and process high-quality data. In contrast, languages spoken in less developed countries, such as Hindi (700 million speakers), Bengali (300 million speakers), and Urdu (230 million speakers), receive comparatively less attention. For context, Spanish is spoken by 600 million people, French by 300 million, and German by 375 million. To make LLMs more accessible to a broader audience, we selected languages based on the number of speakers. Specifically, we included a total of 25 languages, with detailed statistics provided in Table 1 (Eberhard et al., 2024). Altogether, `Babel` serves over 7 billion speakers globally, covering more than 90% of the world's population.

3 Data Preparation

3.1 Data Collection

Building on the foundation of prior work (Team et al., 2024; Dou et al., 2024; Zhang et al., 2024a), we have diversified the range of data sources. In particular, we have incorporated essential knowledge from resources like Wikipedia (Foundation) and textbooks (Ben Al-

Language	Speakers	Language Family	Macroarea
English	1.5B	Germanic	Worldwide
Chinese (Mandarin)	1.4B	Sinitic	Asia
Hindi	700M	Indo-Aryan	Asia
Spanish	595M	Romance	Americas, Europe
Standard Arabic	400M	Semitic	Asia, Africa
French	300M	Romance	Europe, Africa, Americas
Bengali	300M	Indo-Aryan	Asia
Portuguese	270M	Romance	Americas, Europe, Africa
Russian	260M	Slavic	Europe, Asia
Urdu	230M	Indo-Aryan	Asia
Indonesian	200M	Malayo-Polynesian	Asia
Standard German	135M	Germanic	Europe
Japanese	130M	Japonic	Asia
Swahili	100M	Bantu	Africa
Filipino (Tagalog)	100M	Malayo-Polynesian	Asia
Tamil	90M	Dravidian	Asia
Vietnamese	86M	Vietic	Asia
Turkish	85M	Turkic	Asia, Europe
Italian	85M	Romance	Europe
Javanese	83M	Malayo-Polynesian	Asia
Korean	80M	Koreanic	Asia
Hausa	80M	Chadic	Africa
Iranian Persian	80M	Indo-Iranian	Asia
Thai	80M	Kra-Dai	Asia
Burmese	50M	Tibeto-Burman	Asia

Table 1: Languages supported by Babel sorted by the number of speakers (B = Billion, M = Million). Highlighted languages are those underexplored by previous multilingual LLMs.

lal et al., 2024), journalistic content from CC-News (Crawl), web-based corpora such as CulturaX (Nguyen et al., 2023a), and the MADLAD-400 dataset (Kudugunta et al., 2023).

3.2 LLMs-based Data Cleaning and Processing

Due to the limited availability of high-quality training data for many of these languages, we place significant emphasis on optimizing the data-cleaning pipeline to ensure the highest possible data quality. The detailed procedures are outlined as follows.

(1) **Normalization.**

We apply predefined rules to filter out low-quality data, such as documents with fewer than 100 characters or those containing more than 30% digits.

(2) **LLMs-based quality classifier.**

We train the classifier based on the Qwen-2.5-0.5B-Instruct model (Qwen et al., 2025), leveraging a method that combines strong model-based labeling with expert linguistic refinement to construct the training dataset. Specifically, we adopt the "LLM-as-a-judge" approach, where GPT-4o is prompted to evaluate and score potential training data across various dimensions. These initial scores are then carefully reviewed by linguistic experts to ensure that only high-quality data is selected for training the evaluator.

(3) **Deduplicate.**

We identify and remove duplicate documents by hashing, pairing duplicates, constructing graphs, and recording duplicates for removal.

4 Model Description

4.1 Model Extension

To improve the model's performance upper bound, we increase its parameter count through model expansion. Specifically, as illustrated in Figure 1, we use layer extension, a structured method that directly adds new layers with the same structure as the original ones. This approach does not affect critical components of the model, such as attention heads, hidden embeddings, or the embedding layer, etc. Furthermore, inspired by the observation that the middle and back layers are less sensitive to editing (Kim et al., 2023; Men et al., 2024; Zhang et al., 2024b), we choose to extend the layers in the second half of the model.

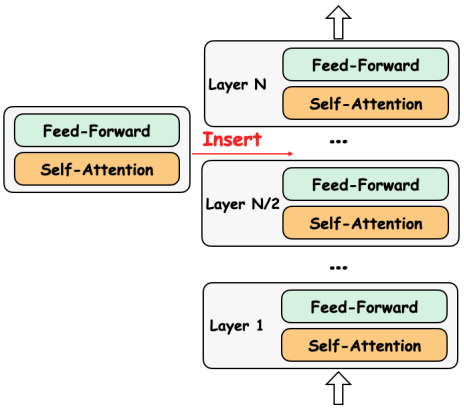


Figure 1: Layer extension for Babel.

We explore various layer extension settings, including different positions for adding layers and methods for parameter initialization. Specifically, we experiment with inserting layers between the original layers or appending them after the original model. For parameter initialization, we consider duplicating the original parameters, initializing with a Gaussian distribution, or adding noise to the original parameters. We select the optimal layer extension methods based on the models' initial performance, prioritizing those that minimally affect performance while also considering their impact on further training.

4.2 Analysis

We conduct an ablation analysis of initialization methods using Qwen2.5-72B-Base (Qwen et al., 2025) as the backbone model and evaluate performance on the MMMLU (OpenAI, 2024) and MGSM (Shi et al., 2023) benchmarks. Specifically, we examine two key aspects: (1) the layer insertion position, which can be either among existing layers or directly appended to the final layer of the original model, and (2) the initialization method, which includes copying the original parameters or introducing noise.

	No-noise	Gaussian ($\mu = 0.1$)	Gaussian ($\mu = 0.01$)	Gaussian ($\mu = 0.0001$)
Among Layers	73.1	13.5	43.1	72.8
After Model	9.4	3.1	3.1	5.2

Table 2: Layer extension initialization analysis. The original performance is 79.5.

Table 2 presents the initialization results for different layer extension methods. Our findings indicate that directly appending new layers to the model leads to a significant decline in performance, suggesting that abrupt structural modifications may disrupt the learned representations. In contrast, inserting new layers within the existing architecture introduces only a minor performance degradation, implying that gradual expansion is less disruptive to the model's stability. Additionally, we observe that duplicating layers without introducing noise achieves the highest performance, as it maintains the integrity of the original feature representations. On the other hand, adding Gaussian noise with a high mean substantially impacts performance, likely due to excessive perturbation of the initialized parameters. Given that adding noise during initialization has the potential to improve training outcomes, we opt for an initialization method that applies Gaussian noise with a mean of 0.0001, striking a balance between stability and adaptability.

5 Model Training

5.1 Model Architecture

Taking into account both accessibility and state-of-the-art performance, we select two model sizes: approximately 10B and 80B. Leveraging Qwen2.5B-7B and Qwen2.5-72B, we employ the model extension method described in Table 3 to initialize `Babel-9B` and `Babel-83B`.

Model	Initialization	Layer Inserting Position
Babel-9B	Duplicate + Gaussian Noise	{14, 16, 18, 20, 22, 24}
Babel-83B		{40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62}

Table 3: Layer extension method details.

5.2 Pre-training Strategy

Stage 1-Recovery. When we modify the parameters and disrupt the well-trained parameter collaboration, Babel's initial performance deteriorates compared to the Qwen2.5 models. Consequently, in the first stage of pre-training, a large and diverse general training corpus encompassing all languages is crucial for recovery. Therefore, we sample a corpus for each language as equally as possible from the pre-training data, although achieving perfect equality can be challenging due to the limited availability of corpora for some languages. Additionally, to accelerate the performance recovery, we combine the English and Chinese training corpora during Stage 1 pre-training. For English, we leverage widely adopted, well-curated pre-training datasets such as RedPajama (Weber et al., 2024) and Proof-Pile-2 (Paster et al., 2023), while for Chinese, we employ YAYI 2 (Luo et al., 2023).

Stage 2-Continuous Training. After recovery, the next step is to enhance multilingual capabilities, particularly for languages overlooked by previous models. To achieve this, we increase the proportion of low-resource languages in the pre-training corpus and continue training the model.

6 Evaluations

We evaluate Babel against comparably sized open-source and commercial multilingual LLMs across a comprehensive set of multilingual tasks.

6.1 Experiment Setup

Dataset We employ multilingual tasks across several categories: (1) **World Knowledge**: MMMLU (OpenAI, 2024), a human-translated version of MMLU (Hendrycks et al., 2021) available in 14 languages. For languages not covered, we use Google Translate (Google, n.d.) to generate translations. Additionally, we include M3Exam (Zhang et al., 2023), which consists of authentic human exam questions collected from various countries, covering multiple subjects and educational levels. (2) **Reasoning**: MGSM (Shi et al., 2022) and XCOPA (Ponti et al., 2020); (3) **Understanding**: XNLI (Conneau et al., 2018); (4) **Translation**: Flore-200 (Team et al., 2022).

Benchmark For the 10B-size model, we compare Babel-9B with GLM4-9B (GLM et al., 2024), Gemma2-9B (Gemma, 2024), Mistral-Nemo-2407[1] (referred to as Mistral-12B), Llama3.1-8B (Dubey et al., 2024), and Qwen2.5-7B (Qwen et al., 2025), listed in order of their release dates. Furthermore, we compare Babel-83B with Llama3.1-70B (Dubey et al., 2024) and Qwen2.5-72B (Qwen et al., 2025).

[1] `https://huggingface.co/mistralai/Mistral-Nemo-Base-2407`

Dataset	Gemma2-9B	Mistral-12B	Llama3.1-8B	Qwen2.5-7B	GLM4-9B	Babel-9B
MMMLU	**59.8**	52.8	49.4	56.7	55.6	59.4
M3Exam	**61.6**	54.2	52.5	58.8	56.6	61.3
XCOPA	84.6	81.3	75.9	81.1	87.3	**89.2**
MGSM	34.3	26.0	18.0	41.1	39.0	**43.4**
XNLI	61.7	55.0	48.9	70.3	69.9	**71.9**
Flores-200	53.2	50.8	50.9	45.5	46.6	**55.1**
Average	59.5	53.4	49.3	58.9	59.2	**63.4**

Table 4: Performance of 10B-Size Base Models vs. Babel-9B-Base.

Dataset	Llama3.1-70B	Qwen2.5-72B	Babel-83B
MMMLU	69.1	74.7	**76.3**
M3Exam	67.4	71.2	**72.1**
XCOPA	92.6	81.1	**92.8**
MGSM	48.9	**63.9**	62.6
XNLI	66.2	74.9	**76.6**
Flores-200	57.4	53.1	**58.8**
Average	66.9	69.8	**73.2**

Table 5: Performance of Open Large Multilingual LLMs vs. Babel-83B-Base.

Evaluation Details We utilize few-shot prompting methods across all datasets and models. For datasets other than Flore-200, accuracy serves as the evaluation metric, while for Flore-200, we use the chrF++ score, translating between each language and English.

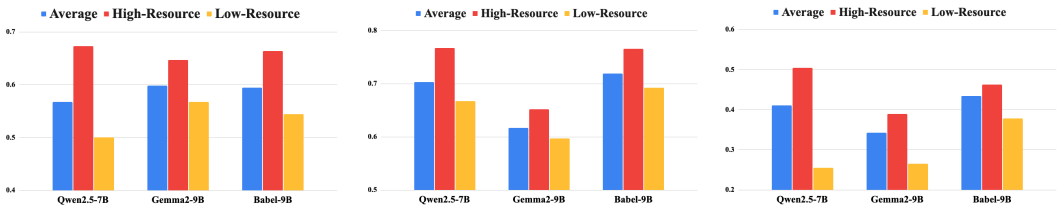
6.2 Main Results

Table 4 shows the results of Babel-9B-Base compared with 10B-size models. We find that Babel-9B-Base achieves the highest overall performance among the evaluated 10B-size base models, with an average score of 63.4, outperforming the closest competitor, Gemma2-9B-Base (59.5), by 3.9 points. Notably, Babel-9B-Base achieves the best results on XCOPA (89.2), MGSM (43.4), XNLI (70.9), and Flores-200 (55.1), demonstrating strong multilingual reasoning, understanding, and translation capabilities. While Gemma2-9B-Base performs competitively on MMMLU and M3Exam, Babel-9B-Base remains consistently strong across all benchmarks. Table 5 illustrates the results of Babel-83B-Base compared with open large multilingual LLMs. We find that Babel-83B achieves the highest overall performance among the evaluated models, with an average score of 73.2, outperforming the closest competitor, Qwen2.5-72B (69.8), by 3.4 points. Notably, Babel-83B achieves the best results on MMMLU (76.3), M3Exam (72.1), XCOPA (92.8), XNLI (76.6), and Flores-200 (58.8), demonstrating strong multilingual reasoning, understanding, and translation capabilities. These results highlight Babel's effectiveness in multilingual understanding and reasoning, positioning it as the most capable open multilingual LLM within its parameter range.

7 Further Analysis

7.1 Performance across Languages

To further analyze Babel's performance across languages, we categorized them into high-resource and low-resource languages based on their scores in Crawl (2025), a statistical measure derived from Common Crawl's monthly archives that reflects the availability of public training corpora. Languages with a score higher than 1 are classified as high-resource languages, including English, Chinese, German, Spanish, French, Indonesian,



(a) MMMLU (b) XNLI (c) MGSM

Figure 2: Performance of Babel-9B-Base comparison across languages.

	English	Multilingual
MMMLU	50.7	**52.1**
M3Exam	55.3	**58.4**
XCOPA	84.2	**83.3**
MGSM	41.8	**42.1**
XNLI	64.5	**67.8**
Flore-200	42.6	**48.1**
Average	56.5	**58.6**

Table 6: Performance comparison of English and multilingual SFT data.

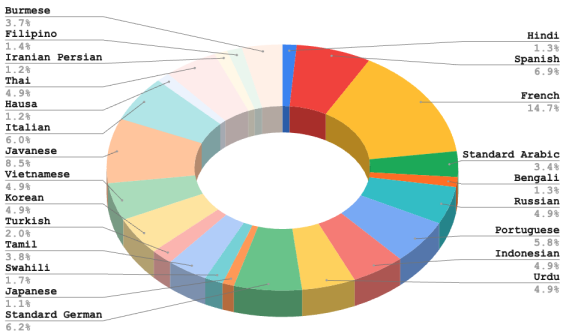


Figure 3: Multilingual SFT data distribution excluding English and Chinese.

Italian, Japanese, Portuguese, Russian, and Vietnamese. In contrast, languages with a score lower than 1 are considered low-resource languages, including Hindi, Standard Arabic, Bengali, Urdu, Swahili, Tamil, Turkish, Korean, Javanese, Hausa, Thai, Iranian Persian, Filipino, and Burmese. We find that low-resource languages are those that have been underexplored by previous multilingual LLMs.

Figure 2 illustrates the performance of Babel-9B-Base across high-resource and low-resource languages, compared to Qwen2.5-7B-Base and Gemma2-9B-Base. Qwen2.5-7B-Base serves as our backbone model, while Gemma2-9B-Base is the second most optimal multilingual LLM. Notably, Babel-9B-Base demonstrates significantly improved performance on low-resource languages compared to Qwen2.5-7B-Base (50.0 vs. 54.4 on MMMLU, 66.7 vs. 69.2 on XNLI, and 25.5 vs. 37.8 on MGSM). Conversely, when compared to Gemma2-9B-Base, Babel-9B-Base achieves higher performance on high-resource languages (64.7 vs. 66.4 on MMMLU, 65.2 vs. 76.6 on XNLI, and 38.9 vs. 46.3 on MGSM). Thus, Babel-9B-Base not only achieves the highest average performance but also strikes a balance between high-resource and low-resource languages.

7.2 Supervised Fine-Tuning (SFT)

SFT Data We primarily leverage open-source multilingual SFT training corpora and translated SFT training data. Specifically, we utilize WildChat (Zhao et al., 2024b), a dataset comprising 1 million user-ChatGPT conversations with over 2.5 million interaction turns. Additionally, we employ Everything Instruct Multilingual (rombodawg, 2025), an extensive Alpaca-instruct-formatted dataset covering a diverse range of topics.

Furthermore, we explore two approaches to constructing an SFT data pool with 400k conversations: one exclusively in English and the other in multiple languages. Table 6 compares these two SFT datasets, revealing that while English SFT data enhances the model's instruction-following capability, multilingual SFT data yields significantly better overall performance. Consequently, we construct a larger multilingual SFT data pool. Specifically, our final dataset consists of approximately 1 million multi-turn conversations. Figure 3 illustrates the distribution of SFT data across languages, excluding English and

Dataset	Gemma2-9B	Mistral-12B	Llama3.1-8B	Qwen2.5-7B	GLM4-9B	Babel-9B
MMMLU	59.6	52.0	50.6	56.0	53.9	**59.8**
M3Exam	**63.2**	54.1	54.2	58.0	55.0	62.9
XCOPA	87.4	83.5	82.1	80.4	86.2	**88.9**
MGSM	62.4	41.4	37.2	59.1	52.2	**64.3**
XNLI	66.7	56.1	55.8	68.3	66.2	**72.4**
Flores-200	54.8	48.9	47.3	45.8	50.8	**56.7**
Average	65.7	56.0	54.5	61.3	60.7	**67.5**

Table 7: Performance of 10B-Size Instruct Models vs. Babel-9B-Chat

Dataset	GPT-4o	Qwen2.5-72B	Llama3.1-70B	Babel-83B
MMMLU	77.3	73.0	71.7	76.8
M3Exam	74.9	70.2	69.5	73.2
XCOPA	90.6	89.2	92.2	92.7
MGSM	83.1	75.8	56.7	72.5
XNLI	69.6	72.6	55.8	76.3
Flores-200	54.9	50.4	56.1	54.8
Average	75.1	71.9	67.0	74.4

Table 8: Babel-83B-Chat vs. top open multilingual LLMs and the best commercial model.

Chinese for better visualization. English comprises 40% of the total SFT training data, while Chinese accounts for 10%.

SFT Training During training, conversations are packed together for efficiency, with a maximum token limit of 4096. The learning rate is configured to 4.0×10^{-6}, and a warmup ratio of 0.1 is applied.

Main Results Table 7 shows that Babel-9B-Chat achieves the highest average score (67.5), surpassing Gemma2-9B-Instruct (65.7) and other models. It leads in five out of six benchmarks, excelling in XCOPA (88.9), MGSM (64.3), XNLI (72.4), and Flores-200 (56.7), demonstrating strong multilingual reasoning and problem-solving. While Gemma2-9B-Instruct slightly outperforms it on M3Exam (63.2 vs. 62.9), Babel-9B-Chat remains consistently strong across tasks. Table 8 illustrates that Babel-83B-Chat achieves the highest average performance (74.4) among open multilingual LLMs, closely trailing GPT-4o (75.1) and outperforming Qwen2.5-72B-Instruct (71.9) and Llama3.1-70B-Instruct (67.0). With leading scores in XCOPA and XNLI, Babel-83B-Chat is the closest open multilingual LLM to the best commercial alternative, demonstrating strong multilingual capabilities.

8 Conclusion

In conclusion, Babel represents a significant advancement in the development of multilingual LLMs, addressing critical gaps in NLP for underserved yet widely spoken languages such as Hindi, Bengali, and Urdu. By leveraging an innovative data-cleaning pipeline, a robust combination of diverse pre-training corpora, supervised fine-tuning, and a novel layer extension technique, Babel delivers state-of-the-art performance across 25 languages, covering over 90% of global speakers. Its open-source variants, Babel-9B and Babel-83B, not only push the boundaries of multilingual NLP but also democratize access to cutting-edge technology. This work underscores the importance of inclusivity in NLP development and sets a strong foundation for future research in multilingual language modeling.

Acknowledgments

We would like to thank Guanzheng Chen for assisting with the implementation of the training codebase. Our special thanks go to our professional and native linguists—Tantong Champaiboon, Nguyen Ngoc Yen Nhi, and Tara Devina Putri—who contributed to building, evaluating, and fact-checking our sampled pretraining dataset. We also appreciate Fan Wang, Jiasheng Tang, Xin Li, and Hao Zhang for their efforts in coordinating computing resources.

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*.

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Anthropic. 2024. Introducing claude 3.5 sonnet. Accessed: 2025-02-27.

Loubna Ben Allal, Anton Lozhkov, Guilherme Penedo, Thomas Wolf, and Leandro von Werra. 2024. Cosmopedia.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations.

Common Crawl. Common crawl news.

Common Crawl. 2025. Statistics of common crawl monthly archives: Distribution of languages. Accessed: 2025-02-27.

Longxu Dou, Qian Liu, Guangtao Zeng, Jia Guo, Jiahui Zhou, Wei Lu, and Min Lin. 2024. Sailor: Open language models for south-east asia. *CoRR*, abs/2404.03608.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2024. *Ethnologue: Languages of the World*, 27th edition. SIL International, Dallas, Texas.

Wikimedia Foundation. Wikimedia downloads.

Gemma. 2024. Gemma 2: Improving open language models at a practical size.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Google. n.d. Google translate api. Accessed: 2025-02-26.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, et al. 2023. Solar 10.7 b: Scaling large language models with simple yet effective depth up-scaling. *arXiv preprint arXiv:2312.15166*.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset.

Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. 2023a. Huatuo-26m, a large-scale chinese medical qa dataset.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023b. Large language models in finance: A survey. In *Proceedings of the fourth ACM international conference on AI in finance*, pages 374–382.

Yin Luo, Qingchao Kong, Nan Xu, Jia Cao, Bao Hao, Baoyu Qu, Bo Chen, Chao Zhu, Chenyang Zhao, Donglei Zhang, Fan Feng, Feifei Zhao, Hailong Sun, Hanxuan Yang, Haojun Pan, Hongyu Liu, Jianbin Guo, Jiangtao Du, Jingyi Wang, Junfeng Li, Lei Sun, Liduo Liu, Lifeng Dong, Lili Liu, Lin Wang, Liwen Zhang, Minzheng Wang, Pin Wang, Ping Yu, Qingxiao Li, Rui Yan, Rui Zou, Ruiqun Li, Taiwen Huang, Xiaodong Wang, Xiaofei Wu, Xin Peng, Xina Zhang, Xing Fang, Xinglin Xiao, Yanni Hao, Yao Dong, Yigang Wang, Ying Liu, Yongyu Jiang, Yungan Wang, Yuqi Wang, Zhangsheng Wang, Zhaoxin Yu, Zhen Luo, Wenji Mao, Lei Wang, and Dajun Zeng. 2023. Yayi 2: Multilingual open-source large language models.

Xin Men, Mingyu Xu, Qingyu Zhang, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. 2024. Shortgpt: Layers in large language models are more redundant than you expect. *arXiv preprint arXiv:2403.03853*.

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

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2023b. Seallms - large language models for southeast asia. *CoRR*, abs/2312.00738.

OpenAI. 2024. Multilingual massive multitask language understanding (mmmlu). Hugging Face.

Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, and Jimmy Ba. 2023. Openwebmath: An open dataset of high-quality mathematical web text.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

rombodawg. 2025. Everything instruct multilingual. `https://huggingface.co/datasets/rombodawg/Everything_Instruct_Multilingual`. A multilingual dataset for instruction fine-tuning, covering a wide variety of topics and languages.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

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. 2022. Language models are multilingual chain-of-thought reasoners.

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, ICLR 2023*. OpenReview.net.

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*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Maurice Weber, Daniel Y. Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. 2024. Redpajama: an open dataset for training large language models. *NeurIPS Datasets and Benchmarks Track*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*.

Wenxuan Zhang, Hou Pong Chan, Yiran Zhao, Mahani Aljunied, Jianyu Wang, Chaoqun Liu, Yue Deng, Zhiqiang Hu, Weiwen Xu, Yew Ken Chia, et al. 2024a. Seallms 3: Open foundation and chat multilingual large language models for southeast asian languages. *arXiv preprint arXiv:2407.19672*.

Yang Zhang, Yawei Li, Xinpeng Wang, Qianli Shen, Barbara Plank, Bernd Bischl, Mina Rezaei, and Kenji Kawaguchi. 2024b. Finercut: Finer-grained interpretable layer pruning for large language models. *arXiv preprint arXiv:2405.18218*.

Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024a. Llama beyond english: An empirical study on language capability transfer.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024b. Wildchat: 1m chatgpt interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*.