What Determines Multilingual LLM Performance? Revisiting the Roles of Linguistic Distance and Resource Availability

Qingan Guo Ryohei Sasano
Nagoya University
guo.qingan.z1@s.mail.nagoya-u.ac.jp sasano@i.nagoya-u.ac.jp

Abstract

This paper investigates whether linguistic distance from English and resource availability affect language-wise performance on modern multilingual tasks, when using recent large language models (LLMs) reported to achieve strong multilingual performance. Our evaluation of six recent LLMs using two modern multilingual benchmarks and round-trip translation reveals that both linguistic distance and resource availability continue to significantly impact LLM performance. We also observe that syntactic distance and resource availability complement each other in predicting language-wise performance of modern LLMs.

1 Introduction

Large language models (LLMs) are increasingly being employed in multilingual systems, but their performance varies substantially across languages. Prior work [1, 2, 3] demonstrates that such differences are associated with the linguistic distance from the dominant language in pre-training, which in most cases is English, and target-language resource availability in the pre-training data. It has also been reported that combining multiple factors enables more accurate prediction of language-wise performance. However, it is unclear whether these findings remain valid when recent LLMs reported to have strong multilingual capabilities are applied to increasingly challenging multilingual tasks that require advanced language understanding and reasoning.

In this study, we revisit the relationship between linguistic distance, resource availability, and language model performance. Specifically, we consider six recent open-source language models and leverage two multilingual benchmarks: Belebele [4] for reading comprehension and MMLU-ProX [5] for knowledge and reasoning, together with round-trip translation (RTT), where an English sentence is translated into a target language and translated back into English. The RTT task differs from the other two benchmarks in that it does not require human-annotated data, making it applicable even to languages for which no human-curated benchmark datasets are available.

Methodologically, we quantify linguistic distance using URIEL+ feature vectors [6] and approximate language resource availability using Common Crawl language distribution statistics.[1)] Focusing on 37 high-frequency languages, we evaluate their correlations with three multilingual tasks to investigate, for recent LLMs and multilingual tasks: (1) whether linguistic distance from English still affects performance, (2) which types of linguistic distance are most influential when such effects persist, (3) whether the influence of resource availability remains, and (4) whether linguistic distance and resource availability are complementary.

2 Methods

2.1 Tasks and Performance Measures

Multilingual benchmarks. We use Belebele and MMLU-ProX, as manually curated multilingual benchmarks. Both benchmarks are multiple-choice question answering (QA) evaluations.[2)] Rather than directly comparing raw scores across tasks, we analyze how language-wise performance variation within each task correlates with linguistic distance and resource availability.

Round-Trip Translation. RTT is a procedure in which an English sentence is translated into a target language and then translated back into English. For RTT inputs, we use sentence-level segments from the WMT News

1) https://commoncrawl.github.io/cc-crawl-statistics/plots/languages.
2) See Appendix A for details of each benchmark.

Crawl 2024 corpus [7]. We measure RTT performance by comparing the back-translated text with the original English source using automatic sentence similarity metrics, BERTScore [8]. Details of the RTT generation settings are provided in Appendix B.

2.2 Linguistic Distance

We use URIEL+ [6] to quantify linguistic distance. URIEL+ is a multilingual database that provides typological, genealogical, and geographic features for cross-lingual NLP analysis. It extends the original URIEL framework [9] by expanding typological feature coverage and improving the robustness and usability of distance calculations. We compute pairwise language distances using cosine distance over URIEL+ feature vectors.

We consider four distance types, each corresponding to a distinct linguistic domain: syntactic distance (SYN), phonological distance (PHON), inventory distance (INV) and morphological distance (MOR). Among these, SYN and PHON were also used in the analysis by Lauscher et al. [1], who reported that they exhibit a strong correlation with the performance of language models. Since English is the dominant language in LLM training and is also used as the pivot language in RTT, we compute and use linguistic distances with respect to English in this study.

2.3 Language Resource Availability

Recent LLMs often do not report per-language proportions of their pre-training corpora, making it difficult to directly quantify resource availability for individual languages. Accordingly, we adopt Common Crawl language distribution statistics as an external proxy for language resource availability. Common Crawl publishes monthly language estimates based on sampled web pages rather than exhaustive counts. To reduce snapshot-specific variance, we aggregate twelve consecutive monthly crawls from Week 33 of 2024 through Week 30 of 2025 by summing the estimated page counts for each language across months. We then use the base-10 logarithm of this aggregated count as our indicator of language resource availability.[3)]

2.4 Language Selection

Figure 1 shows the languages with the highest occurrence frequencies in Common Crawl, along with the proportion of pages for each language. We consider 37 languages with occurrence rates above 0.1% in the Common Crawl for our experiments.[4)] This threshold allows us to cover a wide range of languages while excluding extremely low-resource cases for which Common Crawl statistics may be unreliable.

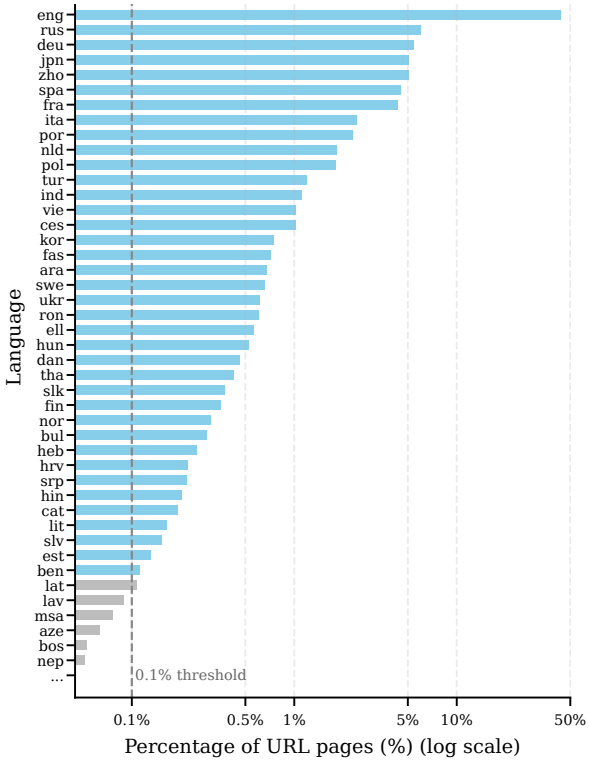


Figure 1: Language codes and page proportions of frequently occurring languages in Common Crawl.

2.5 Language Models

We use four widely used open-source large language model families: Llama-3.1 [10], Mistral-v0.3 [11], Gemma-3 [12], and Qwen3 [13]. For the first three model families, we specifically use Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3, and gemma-3-12b-it, respectively. Hereafter, we refer to these models as Llama-3.1-8B, Mistral-7B, and Gemma-3-12B for simplicity. For Qwen3, to examine the effect of parameter size, we use three models with different parameter sizes: Qwen3-8B, Qwen3-14B, and Qwen3-32B, resulting in a total of six LLMs in our experiments.

3) The values used in our experiments are provided in Appendix C.

4) We exclude Latin (`lat`), which is not a contemporary living language.

TASK	MODEL	SYN P / S	PHON P / S	INV P / S	MOR P / S	RES P / S	S&R P / S
BELEBELE	Llama-3.1-8B	**-0.41 / -0.50**	-0.01 / -0.05	0.19 / 0.15	0.12 / 0.07	0.76 / 0.73	0.82 / 0.83
	Mistral-7B	**-0.48 / -0.56**	-0.11 / -0.09	0.10 / 0.10	-0.04 / -0.08	0.60 / 0.58	0.72 / 0.75
	Gemma-3-12B	**-0.60 / -0.75**	-0.28 / -0.21	0.12 / 0.02	-0.33 / -0.27	0.28 / 0.23	0.63 / 0.75
	Qwen3-8B	**-0.45 / -0.61**	0.07 / -0.06	-0.00 / -0.02	-0.07 / -0.10	0.46 / 0.57	0.60 / 0.79
	Qwen3-14B	**-0.49 / -0.58**	-0.20 / -0.25	0.08 / 0.06	-0.07 / -0.07	0.55 / 0.53	0.69 / 0.74
	Qwen3-32B	**-0.46 / -0.47**	-0.08 / -0.07	0.06 / -0.00	0.03 / 0.12	0.58 / 0.59	0.69 / 0.71
MMLU-ProX	Llama-3.1-8B	**-0.60 / -0.54**	0.12 / 0.18	-0.18 / -0.16	-0.13 / -0.22	0.50 / 0.51	0.65 / 0.64
	Mistral-7B	**-0.73 / -0.91**	-0.19 / -0.25	0.13 / 0.04	-0.46 / -0.62	0.69 / 0.62	0.84 / 0.90
	Gemma-3-12B	**-0.78 / -0.83**	-0.30 / -0.29	0.01 / 0.07	-0.38 / -0.44	0.61 / 0.58	0.84 / 0.89
	Qwen3-8B	**-0.63 / -0.71**	-0.09 / -0.07	0.19 / 0.23	-0.25 / -0.36	0.82 / 0.78	0.87 / 0.86
	Qwen3-14B	**-0.65 / -0.79**	-0.13 / -0.17	0.16 / 0.13	-0.28 / -0.45	0.80 / 0.73	0.87 / 0.85
	Qwen3-32B	**-0.59 / -0.77**	-0.27 / -0.37	0.18 / 0.03	-0.26 / -0.44	0.57 / 0.50	0.68 / 0.74
RTT	Llama-3.1-8B	**-0.63 / -0.74**	-0.36 / -0.38	0.15 / 0.03	-0.39 / -0.33	0.40 / 0.45	0.70 / 0.79
	Mistral-7B	**-0.63 / -0.71**	-0.33 / -0.26	-0.01 / 0.01	-0.24 / -0.23	0.40 / 0.43	0.70 / 0.77
	Gemma-3-12B	**-0.72 / -0.73**	-0.39 / -0.35	-0.12 / -0.04	-0.27 / -0.25	0.20 / 0.21	0.72 / 0.74
	Qwen3-8B	**-0.61 / -0.65**	-0.27 / -0.26	-0.08 / -0.08	-0.13 / -0.16	0.47 / 0.47	0.73 / 0.75
	Qwen3-14B	**-0.65 / -0.65**	-0.33 / -0.32	-0.11 / -0.10	-0.20 / -0.21	0.41 / 0.41	0.72 / 0.70
	Qwen3-32B	**-0.60 / -0.63**	-0.23 / -0.26	-0.07 / -0.06	-0.11 / -0.15	0.51 / 0.49	0.74 / 0.74

Table 1: Pearson (P) and Spearman (S) correlations between language-wise task performance and five factors: syntactic (SYN), phonological (PHON), inventory (INV), morphological (MOR) distances, and resource availability (RES). In addition, S&R denotes the metric that integrates syntactic distance and resource availability introduced in Section 3.2. Bold values indicate the strongest absolute correlation among SYN, PHON, INV, and MOR.

3 Results and Analysis

3.1 Correlation between Individual Metrics and Performance

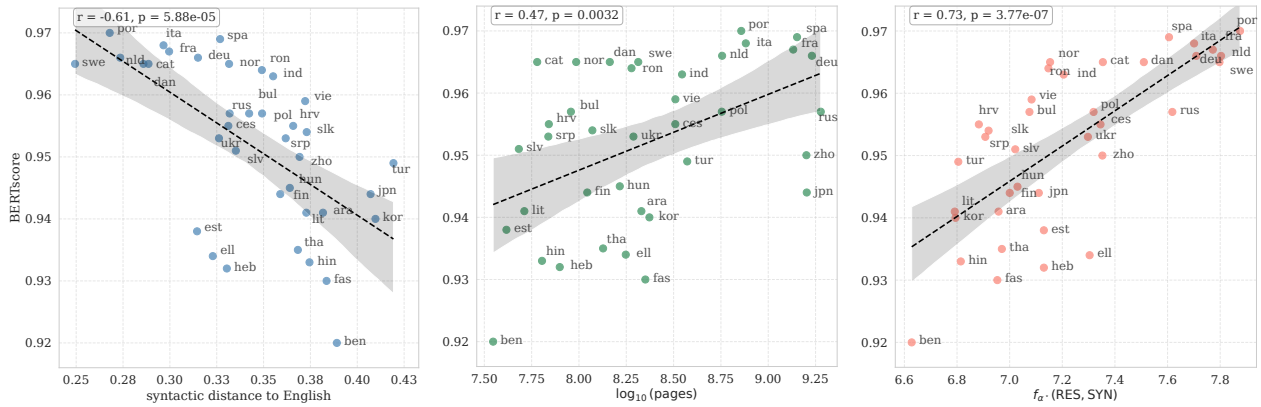
Table 1 summarizes the results. Columns 3 to 6 report the correlations between four types of linguistic distance based on URIEL+ and task performance, while Column 7 shows the correlation between resource availability and task performance. We report both Pearson correlation and Spearman's rank correlation.

First, syntactic (SYN), phonological (PHON), and morphological (MOR) distances are negatively correlated with task performance in almost all settings. As a result, even for recent LLMs and multilingual tasks, languages that are linguistically more distant from English tend to exhibit lower performance. Regarding which types of linguistic distance have the strongest impact, our findings diverge slightly from those reported by Lauscher et al. [1], who studied models such as XLM-R [14] and mBERT [15] and focused on relatively basic tasks including dependency parsing, as well as XNLI [16] and XQuAD [17]. Specifically, they reported that syntactic (SYN) and phonological (PHON) distances have the strongest effects on performance; however, in our experiments, across all three tasks, SYN shows the strongest association with LLM performance.

Shifting our focus to the relationship between language resource availability and LLM performance, we observe consistently positive correlations, suggesting that higher-resource languages tend to exhibit better performance, even for recent LLMs and multilingual tasks. This effect is most pronounced in the MMLU-ProX task; notably, for Qwen3-8B, the Pearson correlation reaches 0.82. Taken together across all experiments, including analyses of linguistic distance, we observe little effect of either language model family or model size. This aligns with the trends reported by Bagheri Nezhad et al. [3].

3.2 Complementarity between Linguistic Distance and Resource Availability

To investigate whether linguistic distance and resource availability serve as complementary factors in language-wise performance prediction, we examine the correlation between LLM performance and a linearly interpolated metric combining linguistic distance and resource availability scores. For linguistic distance, we adopt syntactic distance (SYN) and use $f_\alpha(l)$, computed by the following equation,



(a) Syntactic distance $\textsc{Syn}(l)$. (b) Resource availability $\textsc{Res}(l)$. (c) Combined indicator $f_\alpha(l)$.

Figure 2: Scatter plots of per-language BERTScore against (a) syntactic distance to English, (b) resource availability, and (c) the combined indicator $f_\alpha(l)$.

as the metric:

$$f_\alpha(l) = \alpha \cdot 10 \cdot (1 - \textsc{Syn}(l)) + (1-\alpha) \cdot \textsc{Res}(l),$$

where $\textsc{Syn}(l)$ and $\textsc{Res}(l)$ denote the syntactic distance from English and resource proxy for language l, respectively. The parameter α controls the relative weights of the two metrics and is varied from 0 to 1 in increments of 0.01. The range of variation of $\textsc{Res}(l)$ is approximately an order of magnitude larger than that of $\textsc{Syn}(l)$; therefore, we scale the former by a factor of 10 before tuning the parameter.

The Pearson and Spearman correlation coefficients between LLM performance and $f_\alpha(l)$ obtained using the values of α that yield the highest correlations for each metric are reported in the rightmost column of Table 1. Note that the purpose of this experiment is to examine whether linguistic distance and resource availability are complementary factors for predicting LLM performance; therefore, we do not tune α using development data.[5)] The experimental results show that combining syntactic distance and resource availability yields correlations that are comparable to or higher than those obtained when each metric is used individually, suggesting that these metrics are complementary for predicting LLM performance.

Figure 2 shows scatter plots of language-wise metric values and performance on the RTT task when using Qwen3-8B. By using a metric that integrates syntactic distance and resource availability, we observe that the degree of outlierness for data points that appear as large outliers under each individual metric is reduced. For example, Korean (kor) exhibits lower RTT performance than expected given its resource availability; this can be attributed to its large syntactic distance from English. When using $f_\alpha(l)$, a metric that accounts for syntactic distance, KOR is no longer observed as an outlier.

5) Since the correlation coefficients vary smoothly as α changes, we consider these results to be not highly sensitive to the choice of α.

4 Conclusion

In this paper, we investigated whether prior findings reported for earlier language models such as XLM-R and mBERT, which show that multilingual language model performance is related to each language's linguistic distance from English and its resource availability, also hold for modern LLMs with strong multilingual capabilities and for recent multilingual benchmark tasks that require advanced language understanding and reasoning. Our evaluation of six modern LLMs on two recent multilingual benchmarks, Belebele and MMLU-ProX, together with round-trip translation, demonstrates that multilingual performance in modern LLMs is still strongly affected by linguistic distance and resource availability. Furthermore, among linguistic distance measures, syntactic distance has the strongest impact on the performance of modern LLMs. We also show that syntactic distance and resource availability are complementary indicators, and that combining them enables more accurate prediction of language-wise LLM performance. In future work, we plan to further expand the set of target languages, introduce additional multilingual evaluations, and verify whether similar conclusions persist under alternative prompting and evaluation conditions.

References

[1] Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 4483–4499, Online, November 2020. Association for Computational Linguistics.

[2] Benjamin Muller, Deepanshu Gupta, Jean-Philippe Fauconnier, Siddharth Patwardhan, David Vandyke, and Sachin Agarwal. Languages you know influence those you learn: Impact of language characteristics on multi-lingual text-to-text transfer. In Alon Albalak, Chunting Zhou, Colin Raffel, Deepak Ramachandran, Sebastian Ruder, and Xuezhe Ma, editors, **Proceedings of The 1st Transfer Learning for Natural Language Processing Workshop**, Vol. 203 of **Proceedings of Machine Learning Research**, pp. 88–102. PMLR, 03 Dec 2023.

[3] Sina Bagheri Nezhad and Ameeta Agrawal. What drives performance in multilingual language models? In Yves Scherrer, Tommi Jauhiainen, Nikola Ljubešić, Marcos Zampieri, Preslav Nakov, and Jörg Tiedemann, editors, **Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)**, pp. 16–27, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[4] Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 749–775, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[5] Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, Felix Juefei-Xu, Foutse Khomh, Osamu Yoshie, Qingyu Chen, Douglas Teodoro, Nan Liu, Randy Goebel, Lei Ma, Edison Marrese-Taylor, Shijian Lu, Yusuke Iwasawa, Yutaka Matsuo, and Irene Li. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation, 2025.

[6] Aditya Khan, Mason Shipton, David Anugraha, Kaiyao Duan, Phuong H. Hoang, Eric Khiu, A. Seza Doğruöz, and En-Shiun Annie Lee. URIEL+: Enhancing linguistic inclusion and usability in a typological and multilingual knowledge base. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, **Proceedings of the 31st International Conference on Computational Linguistics**, pp. 6937–6952, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

[7] Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, **Proceedings of the Ninth Conference on Machine Translation**, pp. 1–46, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[8] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020.

[9] Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers**, pp. 8–14, Valencia, Spain, April 2017. Association for Computational Linguistics.

[10] Aaron Grattafiori, et al. The llama 3 herd of models, 2024.

[11] Albert Q. Jiang, et al. Mistral 7b, 2023.

[12] Gemma Team, et al. Gemma 3 technical report, 2025.

[13] An Yang, et al. Qwen3 technical report, 2025.

[14] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[16] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 2475–2485, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[17] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 4623–4637, Online, July 2020. Association for Computational Linguistics.

[18] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024.

[19] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In **Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles**, 2023.

A Benchmark and Evaluation Details

We summarize the multilingual benchmarks used in this study and the evaluation setup in Table 2. Specifically, we evaluate on BELEBELE (4-choice) and MMLU-ProX (10-choice). For both benchmarks, we use `lm-evaluation-harness` (`lm-eval`) [18] and follow the official task implementations as well as their default evaluation and prompting settings.

Benchmark	Type	#Langs	#Used	#Items/lang
BELEBELE	4-choice	122	37	900
MMLU-ProX	10-choice	29	19	11,829

Table 2: Summary of the multilingual benchmarks used in this study. #Used: languages evaluated in this work; #Items/lang: evaluation items per language.

B RTT generation settings.

We implement RTT using vLLM [19]. For the Qwen3 family, we disable the thinking mode in the chat template (`enable_thinking=False`). For each target language, we sample 500 English sentences as RTT inputs. The inference and decoding settings are listed below.

- **Inference:**
  - `dtype=bfloat16`
  - `gpu_memory_utilization=0.7`
  - `max_model_len=8196`
  - `trust_remote_code=True`
- **Decoding:**
  - `temperature=0.5`
  - `top_p=0.8`
  - `top_k=20`
  - `repetition_penalty=1.05`
  - `min_p=0.0`
  - `max_tokens=4096`
- **Prompt:**

  System: You are a translation system. Translate from `{SOURCE_LANGUAGE}` to `{TARGET_LANGUAGE}`. Output only the translation.
  User: `{SENTENCE}`
  Assistant: `{TRANSLATION}`

C Language List and Resource Statistics

Table 3 lists the languages used in our experiments (Common Crawl share > 0.1%) together with their language codes and resource statistics.

Language	Code	CC share	$\log_{10}$(pages)
English	eng	43.72%	10.14
Russian	rus	6.03%	9.28
German	deu	5.41%	9.23
Japanese	jpn	5.08%	9.20
Chinese	zho	5.06%	9.20
Spanish	spa	4.52%	9.15
French	fra	4.31%	9.13
Italian	ita	2.42%	8.88
Portuguese	por	2.29%	8.86
Dutch	nld	1.82%	8.76
Polish	pol	1.81%	8.75
Turkish	tur	1.19%	8.57
Indonesian	ind	1.12%	8.54
Vietnamese	vie	1.03%	8.51
Czech	ces	1.03%	8.51
Korean	kor	0.75%	8.37
Persian	fas	0.71%	8.35
Arabic	ara	0.68%	8.33
Swedish	swe	0.65%	8.31
Ukrainian	ukr	0.62%	8.29
Romanian	ron	0.60%	8.28
Modern Greek	ell	0.56%	8.25
Hungarian	hun	0.52%	8.22
Danish	dan	0.46%	8.16
Thai	tha	0.43%	8.13
Slovak	slk	0.37%	8.07
Finnish	fin	0.35%	8.04
Norwegian	nor	0.31%	7.99
Bulgarian	bul	0.29%	7.96
Hebrew	heb	0.25%	7.90
Croatian	hrv	0.22%	7.84
Serbian	srp	0.22%	7.84
Hindi	hin	0.20%	7.80
Catalan	cat	0.19%	7.78
Lithuanian	lit	0.16%	7.71
Slovenian	slv	0.15%	7.68
Estonian	est	0.13%	7.62
Bengali	ben	0.11%	7.55

Table 3: Details of the languages used in our experiments, with language codes, Common Crawl share, and resource indicator $\log_{10}$(pages).