The Bitter Lesson Learned from 2,000+ Multilingual Benchmarks

Anonymous ACL submission

Abstract

As large language models (LLMs) increasingly bridge linguistic boundaries, robust multilingual evaluation has become critical to ensuring equitable technological development. This position paper analyzes over 2,000 multilingual (non-English) benchmarks from 148 countries, published between 2021 and 2024, to assess past, present, and future practices in multilingual benchmarking. Historically, we observe that a large amount of resources has been invested in creating multilingual benchmarks, yet English is still the most dominant language and many fail to address the needs of underrepresented languages and domains. Currently, we identify the most common use cases of LLMs and observe that benchmarks often lack real-world applicability and fail to align with human judgments, highlighting a disconnect between evaluation frameworks and practical utility. To address these gaps, we propose six essential characteristics for effective benchmarks, including accuracy, resistance to contamination, appropriate challenge levels, practical relevance, linguistic diversity, and cultural authenticity. We also outline five critical research directions, including improving natural language generation evaluation, expanding low-resource language coverage, and developing culturally authentic benchmarks. Our findings highlight a concerning trend: while significant resources are invested in multilingual benchmarks, many fail to reflect real-world applications or align with human judgments. Through this structured analysis, we advocate for evaluation frameworks that better represent global linguistic diversity, ensuring that language technologies serve all communities equitably.

1 Introduction

The remarkable capabilities of large language models (LLMs) have transformed natural language processing (NLP), with applications spanning diverse domains and languages worldwide (Ouyang et al., 2022; Sanh et al., 2022; OpenAI, 2023; Touvron et al., 2023; Anil et al., 2023; Mesnard et al., 2024; Yang et al., 2024; DeepSeek-AI et al., 2025). As these technologies increasingly serve users across linguistic boundaries, robust multilingual evaluation becomes not merely academic but essential (Zhu et al., 2024; Qin et al., 2025). This paper presents a comprehensive analysis of multilingual benchmarking practices to understand past approaches, assess present correlations with human judgments, and chart future directions for more equitable, representative, and effective multilingual evaluation of language technologies.

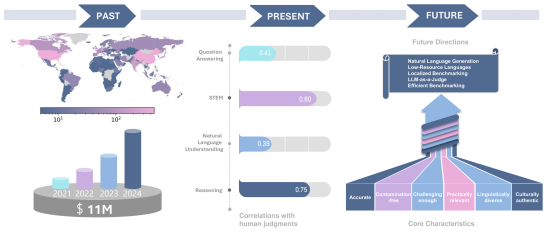


Figure 1: The overview of this work. We examine over 2,000 multilingual (non-English) benchmarks, published between 2021 and 2024, to evaluate past, present, and future practices in multilingual benchmarking.

To provide a comprehensive analysis, we first establish a dataset of multilingual benchmarks by collecting and annotating papers from the arXiv `cs.CL` category (2021–2024) (Section 3). After filtering 370,000 papers through both automated LLM-based screening and expert review, we identify 2,024 relevant studies containing multilingual benchmarks from 148 countries. Our analysis follows a temporal framework organized around three key questions:

1. **PAST: What benchmarks do we currently have? (Section 4)** We document historical trends, revealing the persistent dominance of English and high-resource languages, the underrepresentation of low-resource languages,

and a strong focus on discriminative tasks (66.5%) over generative ones (23.5%). Most benchmarks (61.4%) use original language content, with human translations comprising just 13.2%. Text classification remains the leading task, while question answering has grown rapidly since the rise of LLMs. Dataset sizes have expanded substantially and are mainly built from public domains (e.g., news, social media), with high-value domains like healthcare and law still being rare. Development is concentrated in China, India, Germany, UK, and USA, with Europe emphasizing academic research and China/USA showing more academia-industry collaboration.

2. **PRESENT: What is the current status of multilingual evaluation? (Section 5)** We identify two main findings. First, user interests are consistent across languages (English, Chinese, French, German, Spanish, Russian), with writing tasks dominating (30–45%), followed by commonsense reasoning and programming. Second, benchmark–human judgment correlations vary: STEM tasks (ARC and MGSM) show strong alignment (0.70–0.85), while others are weaker (0.11–0.30). Notably, localized benchmarks (e.g., CMMLU for Chinese) correlate better with human judgments (0.68) than translated ones (0.47 and 0.49), underscoring the need for culturally and linguistically authentic evaluations.

3. **FUTURE: What do we need, and what should we do next? (Section 6)** Based on our analysis, we outline key principles for effective multilingual benchmarks, emphasizing the need for accurate, contamination-free, challenging, practically relevant, linguistically diverse, and culturally authentic evaluations. It identifies critical research directions, including addressing the imbalance in natural language generation (NLG) tasks, improving representation for low-resource languages, creating localized benchmarks that reflect cultural nuances, leveraging LLMs as multilingual judges while addressing biases, and developing efficient benchmarking methods to manage growing complexity.

Our contributions in this position paper are multifaceted. First, we provide a large-scale and comprehensive analysis to date of multilingual benchmarking trends, documenting historical patterns and identifying critical gaps in language coverage. Second, we quantitatively evaluate how well current benchmarks align with human judgments across multiple languages, offering insights into which evaluation approaches best reflect the true model quality. Third, we propose concrete strategies and a call to action for developing the next generation of multilingual benchmarks, balancing practical constraints with the need for greater linguistic diversity and cultural authenticity. By critically examining existing practices and charting clear directions forward, we aim to catalyze more equitable, representative, and meaningful evaluation methodologies that can better guide the development of truly multilingual language technologies serving the global community.

2 Related Work

Multilingual Large Language Models LLMs have revolutionized the landscape of natural language processing (NLP) and artificial intelligence (AI) (Brown et al., 2020; Ouyang et al., 2022; Bai et al., 2022; OpenAI, 2023; Anil et al., 2023; Rivière et al., 2024), extending advanced language understanding and generation to multiple languages (Scao et al., 2022; Wu et al., 2024; Wang et al., 2024b). Early LLMs often centered on English, but multilingual capabilities now play a vital role (Touvron et al., 2023; Jiang et al., 2023). For instance, Llama-1 mainly targeted English (Touvron et al., 2023), whereas Llama-3.1 supports eight languages (Dubey et al., 2024); the Qwen series began with Chinese and English (Bai et al., 2023) but, by the Qwen 3 release, covers more than 100 languages.[1] Meanwhile, researchers have assembled large multilingual corpora to fuel pre-training (Ortiz Suárez et al., 2020; Laurençon et al., 2022; Nguyen et al., 2024), supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF)) (Li et al., 2023; Lai et al., 2023; Singh et al., 2024b). A recent survey offers a systematic overview of multilingual LLMs (Zhu et al., 2024).

Multilingual Evaluation As large language models (LLMs) evolve, evaluating their multilingual performance becomes ever more urgent to address global linguistic diversity (Guo et al., 2023; Chang et al., 2024; Wang et al., 2024a). Two

[1] https://qwenlm.github.io/blog/qwen3/

main strategies have emerged: translating existing English-based evaluation sets into other languages (Shi et al., 2023; Lai et al., 2023; Singh et al., 2024a) or developing fresh benchmarks specific to each target language. Some efforts rely on local exam-style questions (Koto et al., 2023; Li et al., 2024; Yüksel et al., 2024), while others stress culturally relevant content. Recent examples include CulturalBench (Chiu et al., 2024), ArtELingo-28 (Mohamed et al., 2024), and CVQA (Romero et al., 2024), which integrate region-specific images, questions, and opinions to gauge how well LLMs handle real-world cultural nuances.

Ours In this position paper, we analyse over 2,000 multilingual evaluation studies published between 2021 and 2024, focusing on the era shaped by LMs. We highlight emerging trends in multilingual evaluation, examine how current benchmarks align with human judgments, and suggest directions for future research. Our work extends a previous survey of 156 multilingual evaluation studies spanning 2008 to 2021 (Yu et al., 2022), offering an up-to-date perspective on the impact of LLMs.

3 Scope, Collection, and Annotation

In this section, we outline our approach to determining the scope of datasets included in our study, detail our collection process from arXiv submissions, and describe our annotation methodology.

Scope Our work follows the approach of Yu et al. (2022), focusing exclusively on labeled datasets in which a system is tasked with generating an output label y from an input text x. It is important to note that the output label is not limited to a single categorical value but may also consist of generated text, allowing for the production of more complex outputs. To maintain this focus on clear input-output relationships and ensure that the generated labels remain meaningful and contextually relevant, we deliberately exclude English-only benchmarks, training datasets, unlabeled datasets, machine translation datasets, language identification datasets, and multi-modal datasets from our study. Furthermore, we also exclude the programming languages from our study.

Collection In this work, we collect papers under the `cs.CL` category of arXiv from January 1, 2021, to December 31, 2024 using the arXiv API.[2] The arXiv API provides programmatic access to metadata and abstracts of papers, enabling efficient data collection. From this process, we initially retrieved a total of 370K papers. To refine the dataset, we utilize QWEN2.5-MAX to analyze the abstracts of each paper and filter out those irrelevant to our study. Following this automated step, we conduct a manual review to ensure the suitability of each paper for inclusion in our study. This rigorous process resulted in a final dataset of 2,024 papers.

Annotation Besides utilizing metadata from the arXiv API, three authors manually annotate the collected papers following the annotation scheme presented in Table 3 (Appendix B), including publication date, covered languages, tasks and their categories, dataset size, affiliation information, translation use, and domains. These authors have at least one year of experience in NLP research and proficiency in multiple languages.

Takeaways for PAST (Section 4)

Multilingual benchmarks remain dominated by English and high-resource languages, with limited progress in linguistic diversity, domain coverage, and industry involvement. Dataset sizes are rapidly increasing, but most resources originate from a few countries and academic institutions, highlighting persistent gaps in linguistic and contextual representation.

4 PAST: What Benchmarks Do We Have?

In this section, we present a comprehensive analysis of the current landscape of multilingual benchmarks based on our paper collection. We examine the distribution of languages, the evolution of task types, translation methods, and more across benchmarks collected from 2021 to 2024. Understanding the existing benchmark ecosystem is crucial for identifying gaps in language coverage, tracking shifts in evaluation focus, and recognizing opportunities for more inclusive benchmark development.

Languages Figure 2 illustrates the distribution of the top 50 languages across our collected benchmarks. Notably, even though we deliberately exclude English benchmarks during the collection process, English still tops the chart, peaking near 1000 occurrences. Similarly, other high-resource

[2] https://info.arxiv.org/help/api/index.html

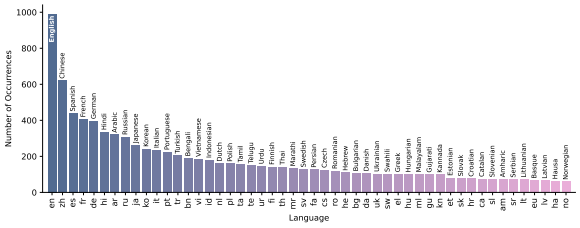


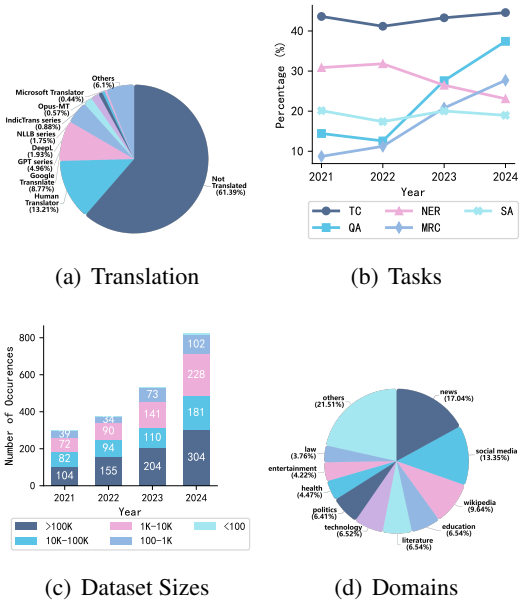
Figure 2: Distribution of the top 50 languages in our multilingual benchmark collection. Although English is deliberately excluded from the collection, it still appears as the most frequent language in the collection. This distribution illustrates the current imbalance in multilingual evaluation benchmarks.

languages (HRLs) occupy the leading positions. In contrast, low-resource languages (LRLs) appear much less frequently. This distribution underscores the dominance of high-resource languages, while highlighting the challenges in achieving broader linguistic representation.

Translations Figure 3(a) illustrates the distribution of translation methods used in benchmark creation. Notably, the majority (61.4%) of benchmarks are not translated, suggesting they are created in their original languages. Human translations account for 13.2% of the benchmarks, representing the highest quality but most resource-intensive approach. Among machine translation tools, Google Translate leads with 8.8%, followed by GPT series models (OpenAI, 2023) (5.0%) and DeepL (1.9%). This distribution highlights both the prevalence of native-language benchmark development and the growth of machine translation technologies in multilingual benchmark creation.

Tasks In our collected benchmarks, 66.5% of the papers focus on discriminative tasks, 23.5% on generative tasks, and 10.0% on both. Figure 3(b) shows the percentage distribution of the top 5 tasks from 2021 to 2024. Text classification remains the dominant task, while question answering and machine reading comprehension have grown significantly, especially since the emergence of LLMs in 2023. Named entity recognition is declining, and sentiment analysis remains stable.

Dataset Sizes Figure 3(c) illustrates that dataset sizes in multilingual benchmarks have consistently increased from 2021 to 2024, with significant growth in larger datasets. For example, very large datasets (>100K examples) have nearly tripled from 104 to 304. This trend reflects the emphasis on large-scale evaluation resources in the era of foundation models.



(a) Translation (b) Tasks

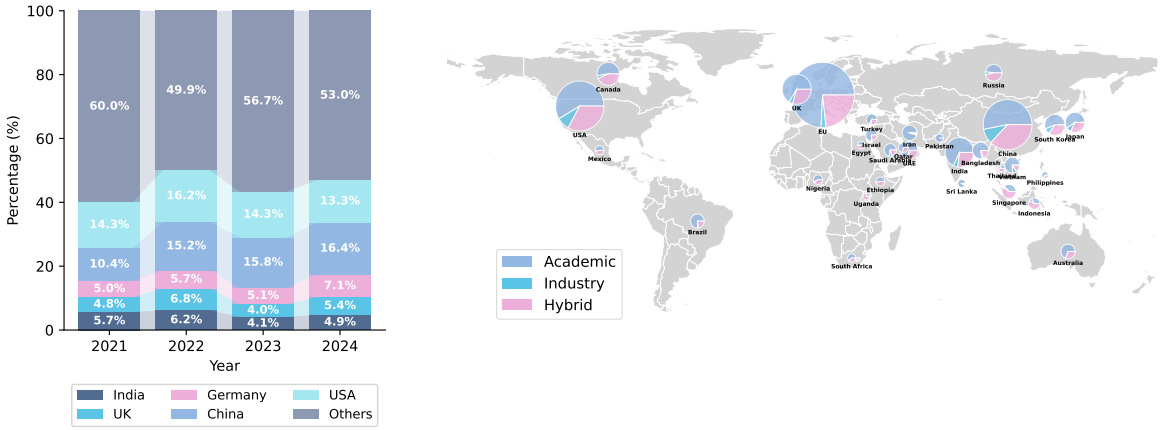
(c) Dataset Sizes (d) Domains

Figure 3: Analysis of multilingual benchmarks. Figure 3(a): 61.4% of benchmarks originate in their native languages with varying translation methods; Figure 3(b): Task trends (2021-2024) show dominance of text classification and growth in question answering; Figure 3(c): Dataset sizes consistently grow; Figure 3(d): Domains are dominated by public sources like news.

Domains Figure 3(d) demonstrates that multilingual benchmarks predominantly utilize publicly accessible sources such as news, social media, and Wikipedia-derived content, which constitute a significant portion of the datasets. This concentration highlights a clear trend: multilingual benchmarks predominantly leverage publicly accessible sources rather than specialized, high-value domains. This distribution reflects the convenience of using public data and the challenges in obtaining high-quality data from specialized domains.

Countries and Institutions Figure 4(a) highlights the countries leading multilingual benchmark development from 2021 to 2024. In this paper, we introduce the term "G5 countries" to refer to the China, India, Germany, UK, and USA, which together account for at least 40% of multilingual benchmark development. Among them, only China shows steady growth from 2021 (10.4%) to 2024 (16.4%). Figure 4(b) illustrates the institutional distribution across the top 50 countries:[3] Europe leads

[3]We use publicly available geographical data for visualization purposes only. The map representation does not imply any



(a) Country Distributions

(b) Affiliation Type Distributions of Top 50 countries

Figure 4: (a) Top 5 countries in multilingual benchmark creation from 2021 to 2024. (b) Affiliation type distributions of the top 50 countries in multilingual benchmark creation. We merge the countries in European Union (EU) into one category for better visualization.

with a strong academic focus, while China and the USA feature more balanced academia-industry collaborations. The predominantly academic-driven nature of these benchmarks points to a gap between research and real-world application, suggesting opportunities for greater industry engagement in multilingual benchmark creation.

Takeaways for PRESENT (Section 5)

Multilingual evaluation reveals that users across languages share remarkably similar interests, with writing tasks (30-45%) dominating across all languages. However, benchmarks vary significantly in their alignment with human judgments: STEM-related tasks (ARC and MGSM) consistently show strong correlation, while translated benchmarks perform inconsistently and often poorly. Localized benchmarks demonstrate superior correlation compared to translated ones, highlighting the critical importance of culturally and linguistically authentic evaluations.

political stance or territorial claims. Data Source: https://www.naturalearthdata.com/downloads/110m-cultural-vectors/

```
[Text]
$text

[Instruction]
Please identify the category of the text above
↪ from Greeting, Writing, Translation, Math,
↪ or Programming. If the text does not belong
↪ to any of these categories, you may add a
↪ new category.
```

Figure 5: The prompt used for categorizing the user interests.

5 PRESENT: What is the Current Status of Multilingual Evaluation?

In this section, we examine the present state of multilingual evaluation from two critical perspectives: the actual interests of multilingual users (Section 5.1), and the alignment between multilingual benchmarks and human judgments (Section 5.2).

5.1 What Are the Multilingual Users Interested in?

Setup To understand the interests of multilingual users, we analyze the distribution of user instructions in Chatbot Arena (Chiang et al., 2024) and WildChat (Zhao et al., 2024b). We analyze six languages, including English, Chinese, French, German, Spanish, and Russian, with 10K instructions for each language. We employ QWEN2.5-MAX to categorize the instructions. We provide 5 seed

	Type	Chinese	French	German	Spanish	Russian
Discriminative						
XNLI	Natural Language Inference	0.233	0.235	0.410	0.483	0.588
ARC	STEM Question Answering	0.818	0.735	0.767	**0.801**	0.803
HellaSwag	Commonsense Reasoning	—	0.684	0.745	0.772	**0.811**
TruthfulQA	Question Answering	0.547	0.613	0.614	0.624	0.773
MMLU	Understanding	0.473	0.398	0.371	0.345	0.303
GlobalMMLU	Understanding	0.487	0.422	0.395	0.349	0.331
Generative						
XQuAD	Question Answering	0.110	—	0.301	0.225	0.154
MGSM	Mathematics	**0.855**	**0.814**	**0.848**	0.798	0.711

Table 1: The Spearman's ρ for various benchmarks across 5 languages. The highest correlation for each language is highlighted in **bold**. Type indicates the capability type that the benchmark is testing.

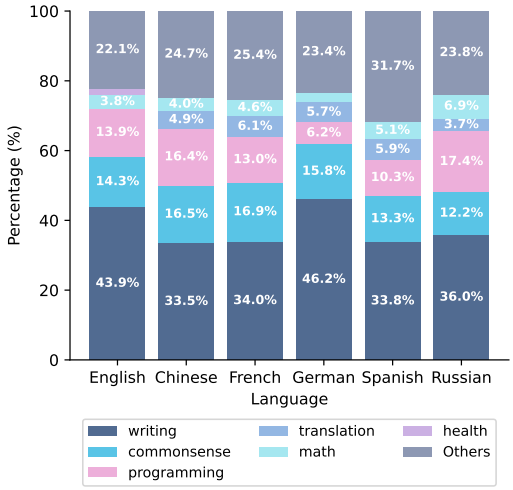


Figure 6: Distribution of user instruction categories across six languages. We discard the "Greetings" category, as it is not a task-oriented instruction.

categories: *Greeting*, *Writing*, *Translation*, *Math*, and *Programming*, but allow the model to introduce new categories, with the prompt shown in Figure 5.

Users from different countries share common interests. We present the distribution of user instructions in Figure 6. Our analysis reveals striking similarities in user interests across different languages. Writing tasks dominate user interactions across all six languages, comprising 30-45% of all instructions. This is followed by commonsense reasoning and programming tasks, which consistently appear among the top three categories in almost all languages. Interestingly, while translation tasks are present in non-English languages (ranging from 4-6% of instructions), they are understandably absent in English. Mathematical tasks appear consistently across all languages but at lower frequencies (3-7%). These patterns suggest that despite linguistic and cultural differences, users across different languages primarily use LLMs for similar purposes, with content creation and practical problem-solving being universal priorities. Furthermore, it is important to note that the user instructions are collected from Chatbot Arena and WildChat, which are primarily used for research purposes. Therefore, the distribution of user instructions may not accurately reflect the general population's interests.

5.2 Do These Benchmarks Correlate Well with Human Judgments?

Setup To assess correlation between benchmarks and human judgments, we compare 30 LLMs' performance on 8 multilingual benchmarks (XNLI (Conneau et al., 2018), ARC (Clark et al., 2018), HellaSwag (Zellers et al., 2019), TruthfulQA (Lin et al., 2022), MMLU (Hendrycks et al., 2021), GlobalMMLU (Singh et al., 2024a), XQuAD (Artetxe et al., 2020), and MGSM (Shi et al., 2023)) against their Chatbot Arena Elo rankings as of December 30 2024. For ARC, HellaSwag, TruthfulQA, and MMLU, we use Google Translate translations from Lai et al. (2023). We analyze five languages (Chinese, French, German, Spanish, Russian) using Spearman's ρ. The details of LLMs are in Appendix A.

STEM-related tasks are more aligned with human judgments. As shown in Table 1, benchmarks that focus on STEM-related capabilities consistently exhibit stronger correlations with human judgments across all languages. ARC, which tests commonsense and scientific reasoning, shows strong correlations ranging from 0.735 in French to 0.818 in Chinese. Similarly, MGSM, which evaluates mathematical problem-solving abilities, demonstrates exceptionally high correlations across all languages (0.711-0.848). In contrast, conventional NLP tasks, such as XNLI and

XQuAD, show weaker correlations (0.110-0.588). We believe this discrepancy is due to the fact that reasoning capabilities are language agnostic and less affected by translation quality. Referring to our analysis in Section 4, these findings highlight a concerning trend: despite significant investments, many multilingual benchmarks fail to align well with human judgments.

Translation is NOT all you need. Table 1 reveals that translated English benchmarks perform inconsistently across languages. XNLI shows weak correlation in Chinese and French (0.233 and 0.235) but moderate correlation in Russian (0.588), while XQuAD performs poorly in Chinese (0.110) compared to German (0.301). Notably, human-translated GlobalMMLU exhibits stronger correlations with human judgments than machine-translated MMLU, emphasizing the importance of high-quality translations.

Localized benchmarks are crucial. In addition to the results presented in Table 1, we also evaluate these LLMs on CMMLU (Li et al., 2024), which includes authentic exam questions from various Chinese exams. CMMLU demonstrates a correlation of 0.682 with human judgments in Chinese, significantly higher than the correlation of translated MMLU in Chinese (0.473 and 0.487). This finding underscores the importance of localized benchmarks specifically designed to capture these nuances and contexts.

Takeaways for FUTURE (Section 6)

Effective multilingual benchmarks should be accurate, contamination-free, challenging, relevant, linguistically diverse, and culturally authentic. Future research should prioritize: expanding language generation evaluation, including low-resource languages, developing culturally-specific benchmarks, exploring LLM-as-a-judge approaches, and creating efficient evaluation methods.

6 FUTURE: What We Need and What We Should Do Next?

In this section, we firstly identify essential characteristics of effective multilingual benchmarks (Section 6.1). We then propose concrete directions for future research efforts that address persistent gaps in evaluating language models across diverse languages, contexts, and applications (Section 6.2).

6.1 What We Need for Effective Multilingual Benchmarks?

Effective multilingual benchmarks require several key characteristics to meaningfully evaluate LLM capabilities across diverse languages. Drawing inspiration from Reiter (2025), we propose the following key characteristics for good multilingual benchmarks:

- **Accurate:** All the benchmarks must contain reliable ground truth annotations, properly verified by domain experts. Recent research reveals that even widely adopted benchmarks like MMLU contain numerous errors (Gema et al., 2025), undermining evaluation validity.
- **Contamination-free:** Benchmark contamination occurs when evaluation data appears in a model's training corpus, leading to inflated performance metrics that misrepresent a model's true capabilities. Recent research even demonstrate that the data contamination in one language can be transferred to another language (Yao et al., 2024).
- **Challenging enough:** The performance of recent state-of-the-art models has quickly saturated on widely used benchmarks, with scores approaching or exceeding human performance. As shown in KILLED-BY-LLM,[4] the average lifespan of a popular benchmark is only 2.6 years before it is not sufficiently challenging. Therefore, multilingual benchmarks must maintain an appropriate difficulty level that can differentiate between models.
- **Practically relevant:** As shown in Figure 3(d) and Figure 4(b), about 70% of the benchmarks are released by the academic community and these benchmarks are created from the public sources, which may not always reflect real-world applications. Without this practical grounding, benchmarks risk optimizing for capabilities that have limited real-world impact, creating a disconnect between research advancements and actual user needs.
- **Linguistically diverse:** Effective multilingual benchmarks must include languages representing different families, writing systems, and typological features. As shown in Figure 2, our

[4] https://r0bk.github.io/killedbyllm/

analysis reveals severe imbalances favoring high-resource languages, with many language families entirely unrepresented.

- **Culturally authentic:** Multilingual benchmarks must reflect the cultural diversity. Recent research has highlighted the importance of cultural considerations in benchmark design (Son et al., 2024; Zhao et al., 2024a; Chiu et al., 2024). Our results also demonstrate that CMMLU (Li et al., 2024) aligns better with Chinese users' judgments, compared with the translated MMLU (Hendrycks et al., 2021; Singh et al., 2024a).

Advancing toward more comprehensive multilingual benchmarks following these principles is essential for ensuring language technologies serve global populations equitably.

6.2 What We Should Do Next?

Building on our analysis of necessary characteristics for effective multilingual benchmarks, we now outline five critical research directions.

Natural Language Generation While most existing multilingual benchmarks focus on discriminative tasks like classification and multiple-choice problems, natural language generation (NLG) capabilities remain significantly underassessed across diverse languages. As discussed in Section 4, about 66% of the benchmarks are focused on discriminative tasks, while only 23% of the benchmarks are focused on NLG tasks. This imbalance is particularly concerning as generative applications are increasingly prevalent in real-world applications.

Low-Resource Languages As shown in Figure 2, low-resource languages, which lack substantial amounts of digital text data, remain significantly underrepresented in current multilingual benchmarks. This underrepresentation creates a problematic cycle: models perform poorly on these languages, leading researchers to focus on higher-resource languages where improvements are more easily demonstrable, further widening the capability gap. Breaking this cycle requires deliberate effort to develop specialized benchmarks that focus on low-resource languages.

Localized Benchmarking Current evaluation approaches often rely on translated content from English or other high-resource languages. As shown in Section 5.2, the localized benchmarks can achieve better alignment with the target language and culture. Recent work has begun addressing these issues by incorporating more diverse cultural perspectives (Li et al., 2024; Son et al., 2024; Zhao et al., 2024a; Chiu et al., 2024), but there remains significant room for benchmarks that assess models on their ability to handle local applications.

LLM-as-a-Judge Recent research has demonstrated the potential of using LLMs themselves as evaluation tools for assessing the quality of model-generated text in English (Zheng et al., 2023; Dubois et al., 2024). This approach offers promising opportunities for multilingual evaluation by extending these techniques across diverse languages and tasks. However, deploying LLMs as judges in multilingual contexts also introduces unique challenges, such as the potential evaluation biases that mirror the language disparities in the judge models.

Efficient Benchmarking Current benchmarks often include numerous languages and tasks to thoroughly assess model capabilities. The size of benchmarks grows linearly with the number of languages and combinatorially with the number of tasks and evaluation dimensions. As shown in Figure 3(c), the size of the benchmarks is growing rapidly over the years. Future research should aim to develop methods for efficient evaluation, such as identifying representative language-task subsets, employing statistical sampling techniques, or using adaptive testing approaches that maintain evaluation quality while reducing computational costs.

7 Conclusion

In this position paper, we present a comprehensive analysis of multilingual benchmarking practices by systematically examining over 2,000 studies. Our findings uncover persistent disparities in language representation, evolving task types, dataset sizes, and other critical factors. Through a present-focused investigation, we identify user interests across different languages and highlight significant gaps between benchmark scores and actual human preferences, particularly in translation-based evaluations. Our analysis underscores six key limitations in current multilingual evaluation practices and proposes guiding principles for effective multilingual benchmarking. Additionally, we outline five critical research directions to advance the field.

8 Limitations

In Section 5.1, we employ QWEN2.5-MAX to categorize user interests, acknowledging the potential for annotation errors. The instructions are derived from Chatbot Arena (Chiang et al., 2024) and WildChat (Zhao et al., 2024b), which are predominantly utilized by researchers and developers. As a result, the user interests in our dataset may not fully reflect those of the general population.

In Section 5.2, we analyze correlations between benchmarks and human judgments using 30 open-source LLMs. Due to constraints in computational resources and financial budgets, our evaluation is limited to these models and excludes proprietary LLMs. Expanding the evaluation to include more models, particularly proprietary LLMs, would provide a broader perspective.

References

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, and 33 others. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *CoRR*, abs/2309.16609.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022. Constitutional AI: harmlessness from AI feedback. *CoRR*, abs/2212.08073.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3):39:1–39:45.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2024. Culturalbench: a robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of llms. *CoRR*, abs/2410.02677.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 81 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2024. Length-controlled al-

pacaeval: A simple way to debias automatic evaluators. *CoRR*, abs/2404.04475.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile van Krieken, and Pasquale Minervini. 2025. Are we done with mmlu? *Preprint*, arXiv:2406.04127.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. 2023. Evaluating large language models: A comprehensive survey. *CoRR*, abs/2310.19736.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12359–12374, Singapore. Association for Computational Linguistics.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Sasko, Quentin Lhoest, Angelina McMillan-Major, Gérard Dupont, Stella Biderman, Anna Rogers, Loubna Ben Allal, Francesco De Toni, and 35 others. 2022. The bigscience ROOTS corpus: A 1.6tb composite multilingual dataset. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. Bactrian-x : A multilingual replicable instruction-following model with low-rank adaptation. *CoRR*, abs/2305.15011.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285, Bangkok, Thailand. Association for Computational Linguistics.

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

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, and 30 others. 2024. Gemma: Open models based on gemini research and technology. *CoRR*, abs/2403.08295.

Youssef Mohamed, Runjia Li, Ibrahim Said Ahmad, Kilichbek Haydarov, Philip Torr, Kenneth Church, and Mohamed Elhoseiny. 2024. No culture left behind: ArtELingo-28, a benchmark of WikiArt with captions in 28 languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20939–20962, Miami, Florida, USA. Association for Computational Linguistics.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke

Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2025. A survey of multilingual large language models. *Patterns*, 6(1):101118.

Ehud Reiter. 2025. We need better llm benchmarks.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, and 80 others. 2024. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Santiago Góngora, Aishik Mandal, Sukannya Purkayastha, Jesús-Germán Ortiz-Barajas, Emilio Villa-Cueva, Jinheon Baek, Soyeong Jeong, Injy Hamed, Zheng Xin Yong, Zheng Wei Lim, Paula Mónica Silva, Jocelyn Dunstan, Mélanie Jouitteau, David Le Meur, Joan Nwatu, Ganzorig Batnasan, and 57 others. 2024. CVQA: culturally-diverse multilingual visual question answering benchmark. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, and 21 others. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, and 30 others. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

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, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Madeline Smith, Antoine Bosselut, Alice Oh, André F. T. Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2024a. Global MMLU: understanding and addressing cultural and linguistic biases in multilingual evaluation. *CoRR*, abs/2412.03304.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, and 14 others. 2024b. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2024. KMMLU: measuring massive multitask language understanding in korean. *CoRR*, abs/2402.11548.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Weixuan Wang, Barry Haddow, and Alexandra Birch. 2024a. Retrieval-augmented multilingual knowledge editing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 335–354, Bangkok, Thailand. Association for Computational Linguistics.

Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. 2024b. Bridging the language gaps in large language models with inference-time cross-lingual intervention. *CoRR*, abs/2410.12462.

Minghao Wu, Yulin Yuan, Gholamreza Haffari, and Longyue Wang. 2024. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *CoRR*, abs/2405.11804.

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, and 22 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Feng Yao, Yufan Zhuang, Zihao Sun, Sunan Xu, Animesh Kumar, and Jingbo Shang. 2024. Data contamination can cross language barriers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17864–17875, Miami, Florida, USA. Association for Computational Linguistics.

Xinyan Yu, Trina Chatterjee, Akari Asai, Junjie Hu, and Eunsol Choi. 2022. Beyond counting datasets: A survey of multilingual dataset construction and necessary resources. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3725–3743, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Arda Yüksel, Abdullatif Köksal, Lütfi Kerem Senel, Anna Korhonen, and Hinrich Schuetze. 2024. TurkishMMLU: Measuring massive multitask language understanding in Turkish. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7035–7055, Miami, Florida, USA. Association for Computational Linguistics.

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

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024a. WorldValuesBench: A large-scale benchmark dataset for multi-cultural value awareness of language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17696–17706, Torino, Italia. ELRA and ICCL.

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, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Shaolin Zhu, Supryadi, Shaoyang Xu, Haoran Sun, Leiyu Pan, Menglong Cui, Jiangcun Du, Renren Jin, António Branco, and Deyi Xiong. 2024. Multilingual large language models: A systematic survey. *CoRR*, abs/2411.11072.

A LLMs in Evaluations

In this work, we evaluate 30 LLMs on 8 multilingual benchmarks across 5 languages. We present all the LLMs used in this work in Table 2.

B Annotation

Our annotation schema is presented in Table 3.

Model	Chinese	French	German	Spanish	Russian
google/gemma-1.1-7b-it	1118.6	1018.5	1049.9	1052.1	1076.6
CohereForAI/aya-expanse-32b	1267.1	1199.8	1196.6	1199.7	1249.8
google/gemma-7b-it	1095.5	979.7	978.9	983.3	1014.7
meta-llama/Llama-3.1-8B-Instruct	1211.2	1142.8	1138.8	1176.5	1187.2
meta-llama/Llama-2-7b-chat-hf	1031.8	925.1	956.4	989.9	1015.4
microsoft/Phi-3-small-8k-instruct	1122.9	1091.9	1075.4	1110.6	1138.4
Qwen/Qwen2.5-Coder-32B-Instruct	1277.2	1182.5	1192.8	1219.7	1250.3
meta-llama/Meta-Llama-3-8B-Instruct	1135.2	1113.1	1101.6	1174.5	1138.8
ibm-granite/granite-3.0-8b-instruct	1130.5	1027.3	983.2	1034.0	1102.5
microsoft/Phi-3-medium-4k-instruct	1165.1	1070.4	1100.7	1098.7	1169.8
google/gemma-2-27b-it	1278.8	1190.1	1206.0	1223.9	1255.9
google/gemma-1.1-2b-it	1076.4	963.9	947.8	991.4	1020.1
microsoft/Phi-3-mini-128k-instruct	1076.3	994.2	1009.0	1056.1	1039.0
meta-llama/Llama-3.2-1B-Instruct	1023.0	1021.4	1010.9	1030.2	972.9
CohereForAI/aya-expanse-8b	1241.2	1166.1	1180.3	1161.4	1228.4
meta-llama/Llama-2-13b-chat-hf	1055.3	992.3	998.2	1076.9	1075.9
meta-llama/Llama-3.2-3B-Instruct	1084.7	1031.0	1053.6	1095.7	984.0
google/gemma-2-2b-it	1190.3	1129.3	1105.1	1144.6	1142.1
HuggingFaceH4/zephyr-7b-beta	1017.9	989.5	975.5	1040.4	1067.1
microsoft/Phi-3-mini-4k-instruct	1081.8	1033.2	1038.5	1094.3	1056.5
google/gemma-2b-it	1049.3	852.2	909.9	985.9	964.9
mistralai/Mistral-7B-Instruct-v0.2	1068.4	983.9	979.2	1025.7	1045.8
HuggingFaceTB/SmolLM2-1.7B-Instruct	1106.5	1001.9	948.5	941.4	1033.0
Qwen/Qwen1.5-14B-Chat	1202.6	1068.4	1042.3	1079.2	1073.3
google/gemma-2-9b-it	1243.3	1142.9	1180.2	1199.7	1220.1
Qwen/Qwen1.5-4B-Chat	1083.6	929.8	904.9	1013.6	977.5
mistralai/Ministral-8B-Instruct-2410	1256.8	1133.4	1128.0	1131.1	1222.9
ibm-granite/granite-3.0-2b-instruct	1130.3	1003.1	988.4	1037.0	1081.4
Qwen/Qwen1.5-7B-Chat	1196.1	1017.3	1022.4	1012.4	1035.8
allenai/OLMo-7B-Instruct	1071.7	879.6	885.3	975.9	970.9

Table 2: LLMs used for evaluation and their Elo scores on 5 languages up to December 30, 2024.

Aspect	Description
Year and Month	The publication year and month of the paper.
Languages	The languages covered by the dataset.
Task Category	The task types discussed in the paper (e.g., discriminative, generative, or both).
Tasks	The specific tasks covered (e.g., sentiment analysis, question answering, summarization, etc.).
Dataset Size	The approximate size of the dataset, categorized as: $<$100, 100–1K, 1K–10K, 10K–100K, or $>$100K.
Affiliation Type	The Affiliation type of the creator of the dataset (e.g. academic, industry, or both).
Affiliation	The affiliations that create the dataset.
Country	The countries of the affiliations that create the dataset.
Translation	The method used for dataset translation (e.g., not translated, human translation, Google Translate, etc.).
Domain	The domains of the dataset (e.g., news, social media, etc.).

Table 3: Annotation scheme for the collected paper.