Hengyu Luo, Zihao Li, Joseph Attieh, Sawal Devkota, Ona de Gibert, Xu Huang, Shaoxiong Ji, Peiqin Lin, Bhavani Sai Praneeth Varma Mantina, Ananda Sreenidhi, Raúl Vázquez, Mengjie Wang, Samea Yusofi, Fei Yuan, Jörg Tiedemann

GlotEval: A Test Suite for Massively Multilingual Evaluation of Large Language Models

2025

https://doi.org/10.18653/v1/2025.emnlp-demos.43

Publisher's PDF

Luo, Hengyu, et al. "GlotEval: A Test Suite for Massively Multilingual Evaluation of Large Language Models." Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations [Suzhou, China], 2025, pp. 602–14. https://doi.org/10.18653/v1/2025.emnlp-demos.43.

GlotEval: A Test Suite for Massively Multilingual Evaluation of Large Language Models

Hengyu Luo[1], Zihao Li[1], Joseph Attieh[1†], Sawal Devkota[2†], Ona de Gibert[1†], Xu Huang[4†], Shaoxiong Ji[2,5,6†], Peiqin Lin[3†], Bhavani Sai Praneeth Varma Mantina[2†], Ananda Sreenidhi[2†], Raúl Vázquez[1†], Mengjie Wang[2†], Samea Yusofi[2†], Fei Yuan[7], Jörg Tiedemann[1,5]

[1]University of Helsinki, Finland [2]Technical University of Darmstadt, Germany
[3]University of Munich, Germany [4]Nanjing University, China [5]ELLIS Institute Finland
[6]University of Turku, Finland [7]Shanghai Artificial Intelligence Laboratory, China

Correspondence: hengyu.luo@helsinki.fi & shaoxiong.ji@utu.fi [†] Equal contribution.

Abstract

Large language models (LLMs) are advancing at an unprecedented pace globally, with regions increasingly adopting these models for applications in their primary languages. Evaluating these models in diverse linguistic environments, especially in low-resource languages, has become a major challenge for academia and industry. Existing evaluation frameworks suffer from inconsistency across different benchmarks, being disproportionately focused on English and a handful of high-resource languages, thereby overlooking the realistic performance of LLMs in multilingual and lower-resource scenarios. To address this critical challenge of fragmented and inconsistent multilingual evaluation, we introduce GlotEval, a unified and lightweight framework that systematically integrates 27 benchmarks under a standardized ISO 639-3 language identifier system, allowing for seamless incorporation of new benchmarks. Supporting nine key tasks (machine translation, text classification, summarization, open-ended generation, reading comprehension, sequence labeling, intrinsic evaluation, instruction following and reasoning), spanning over dozens to hundreds of languages, GlotEval uniquely enables language-specific, cross-benchmark analysis and non-English-centric evaluations at a scale previously less practical for many researchers. This enables a precise diagnosis of model strengths and weaknesses in diverse linguistic contexts. A multilingual translation case study demonstrates GlotEval's applicability for multilingual and language-specific evaluations.

GlotEval: github.com/MaLA-LM/GlotEval

1 Introduction

In recent years, driven by rapid progress in natural language processing and deep learning, large language models (LLMs) such as GPT-4 (OpenAI, 2023) and DeepSeek-R1 (DeepSeek-AI et al., 2025) have shown remarkable reasoning and generation capabilities across multiple languages and tasks. Although these models approach or surpass expert-level performance in certain high-resource languages (e.g., English), they often exhibit substantial performance fluctuations in other linguistic environments (Zhang et al., 2024). This discrepancy partially arises from the imbalance and scarcity of training data of low-resource languages, and partially from the limited multilingual coverage of current evaluation frameworks: many were originally designed for English or a few widely spoken languages, making it difficult to extend them efficiently to more diverse linguistic tasks or to adapt custom prompts and configurations for each language. Meanwhile, as LLMs proliferate worldwide and different regions rely on their respective local languages, large-scale (massively) multilingual evaluation involving numerous low-resource languages has emerged as a critical research direction.

Recent developments in LLM evaluation toolkits such as EleutherAI's LM Evaluation Harness (Gao et al., 2023) and UltraEval (He et al., 2024) have facilitated automatic evaluation. However, significant gaps persist in language coverage, task diversity, and evaluation flexibility (Chang et al., 2024), especially in evaluating multilingual LLMs in a massively multilingual scenario. To address these issues, we present GlotEval, an evaluation framework designed to provide *systematic* support for a *broad range of languages*, with a strong focus on low-resource ones. Building on the core processes of LLM evaluation—data preparation, model inference, post-processing, and metric computation—GlotEval introduces three novel features.

1. **Consistent Cross-benchmark Multilingual Evaluation.** We integrate 27 existing multilingual benchmarks into a unified pipeline, by standardizing all ISO 639-3 language codes

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 602–614
November 4-9, 2025

in the different benchmarks,[1] which is an accepted standard with a good coverage of the world's languages. By aligning benchmark language identifiers with ISO 639-3 codes, we enable evaluations for specific languages or language groups (e.g., Bantu, Dravidian, or Uralic languages), allowing the framework to automatically search among integrated benchmarks to find matching test sets. This mapping also makes it easier to incorporate new large-scale benchmarks that target mid- or low-resource languages, ensuring flexibility for future expansions.

2. **Language-Specific Prompt Templates.** Users can configure prompts for each language individually, thereby enabling more precise assessments of a model's instruction-following ability across diverse linguistic settings. All templates are maintained in a centralized prompt library that supports multilingual benchmarks, allowing easy customization as needed. In this way, each task within a benchmark can be run potentially using prompts in the task's original language, rather than defaulting to English prompts. To simplify cross-lingual adaptation, we also implemented Microsoft Translator integration that automatically propagates user-defined prompt templates from one single language to 130+ supported languages.[2]

3. **Non-English-Centered Machine Translation Evaluation.** GlotEval is designed to break away from the traditional English-centric paradigm. Thanks to translation benchmarks featuring fully or partially multi-aligned datasets, GlotEval enables non-English-centered translation evaluations by allowing any supported language to serve as the pivot: users simply update the pivot language in the configuration to assess "any-to-pivot" / "pivot-to-any" translation directions. This flexibility ensures that GlotEval breaks from the traditional "English $\leftrightarrow$ other language" paradigm and adapts seamlessly to diverse, potentially low-resource, language pairs.

By bringing all these capabilities together in a cohesive framework, GlotEval aims to facilitate large-scale, in-depth evaluations of multilingual LLMs across both widely spoken and underrepresented languages, ultimately driving forward more inclusive LLM evaluation. Thus, GlotEval's primary contribution is not the collection of new tasks, but the synergistic integration and standardization of existing benchmarks, which can be a robust tool for researchers and developers conducting massively multilingual LLM evaluation.

2 Related Work

Several evaluation toolkits and benchmarks have been developed to systematically assess LLMs. EleutherAI's LM Evaluation Harness (Gao et al., 2023) is a widely adopted framework covering over 60 tasks, including multilingual datasets such as XNLI (15 languages) and Belebele (122 languages). UltraEval (He et al., 2024) improves modularity and supports FLORES-200 for multilingual translation. OpenAI Evals provides a highly flexible, community-driven framework,[3] and OpenCompass (Contributors, 2023) offers a comprehensive platform with broad support for datasets and models. MEGA (Ahuja et al., 2023) evaluates generative LLMs across diverse languages, with a focus on standard NLP benchmarks. LightEval (Fourrier et al., 2023) developed a flexible LLM evaluation framework that supports different backends.

Despite these advancements, significant gaps remain in language coverage, task diversity, and evaluation flexibility (Chang et al., 2024). Specifically, most toolkits rely on static task definitions and rarely adopt standardized language identifiers across benchmarks, making it difficult to conduct language-specific evaluations in a cross-benchmark setting. As a result, evaluations for a given language (group) must often be performed in isolation for each benchmark, limiting scalability and linguistic granularity. Furthermore, support for language-specific prompt customization is limited—most toolkits default to using English prompts regardless of the task language, which failed to take both goals of languages in multilingual evaluation, i.e., task performance versus language understanding, into consideration (Poelman and de Lhoneux, 2024).

[1] https://iso639-3.sil.org/about

[2] https://learn.microsoft.com/en-us/azure/ai-services/translator/language-support

Task	Benchmark	Languages	Domain	Open Source	Metrics
Text Classification	Taxi-1500 (Ma et al., 2024)	1507	Bible text	Yes (GitHub)	Acc., F1
	SIB-200 (Adelani et al., 2024)	205	News topics	Yes (HF, GitHub)	Acc., F1
Token Classification	WikiANN (Pan et al., 2017)	282	Wikipedia NER	Yes (HF)	F1
	UD treebank v2.15 (de Marneffe et al., 2021)	148	POS tagging	Yes (UD website)	F1
Machine Translation	FLORES-200 (NLLB Team et al., 2022)	200+	General web	Yes (HF)	BLEU, ChrF++, COMET
	FLORES+	212	Gen. web, low-resource focus	Yes (HF)	BLEU, ChrF++, COMET
	NTREX-128 (Federmann et al., 2022)	128	News	Yes (GitHub)	BLEU, ChrF++, COMET
	AmericasNLP (de Gibert et al., 2025)	14	Short sentences, court proceedings, books.	Yes (GitHub)	BLEU, ChrF++
	TICO-19 (Anastasopoulos et al., 2020)	37	COVID-19 medical	Yes (GitHub, OPUS)	BLEU, ChrF++
	IN22 (Gala et al., 2023)	23	Indian langs., news+conv.	Yes (GitHub)	BLEU, ChrF++
	NTEU (Bié et al., 2020)	24	EU formal (gov)	Partial (Upon request)	BLEU, ChrF++
	MAFAND (Adelani et al., 2022)	22	News	Yes (GitHub)	BLEU, ChrF++
	Tatoeba Challenge v2023 (Tiedemann, 2020)	500+	Mixed short sents.	Yes (GitHub)	BLEU, ChrF
	OpenSubtitles v2024 (Lison and Tiedemann, 2016)	93	Subtitles	Yes (GitHub)	BLEU, ChrF
	MMHB (Tan et al., 2024)	9	Multilingual bias detection	Yes (GitHub)	ChrF with gender
Open-Ended Generation	Aya (Singh et al., 2024b)	119	Instruction-following	Yes (HF)	self-BLEU
	PolyWrite (Ji et al., 2024)	240	Creative writing	Yes (HF)	self-BLEU
Intrinsic Evaluation	PBC (Mayer and Cysouw, 2014)	372+	Bible text	Partial (Upon request)	NLL
	MaLA (Ji et al., 2024)	546	General web	Yes (HF)	NLL
Comprehension	MMMLU (Hendrycks et al., 2021)	14+	General knowledge QA	Yes (HF)	Acc.
	Global-MMLU (Singh et al., 2024a)	42	Culture-aware QA	Yes (HF)	Acc.
Summarization	XLSum (Hasan et al., 2021)	44	News	Yes (HF, GitHub)	ROUGE
	MassiveSumm Long (Varab and Schluter, 2021)	55	News	Yes (HF)	ROUGE
	MassiveSumm Short (Varab and Schluter, 2021)	88	News	Yes (HF)	ROUGE
Instruction Following	BenchMAX Rule-based (Huang et al., 2025)	17	Verifiable instructions	Yes (HF)	Instruction-level Acc. etc.
Reasoning	BenchMAX Math (Huang et al., 2025)	17	Grade School Math	Yes (HF)	Accuracy
	BenchMAX Science (Huang et al., 2025)	17	Graduate-level Scientific QA	Yes (HF)	Accuracy

Table 1: Overview of multilingual LLM evaluation benchmarks, with typical metrics used in each.

3 GlotEval

3.1 Benchmarks, Languages and Metrics

As shown in Table 1, GlotEval integrates publicly available multilingual benchmark datasets, covering machine translation, text classification, summarization, open-ended generation, reading comprehension, sequence labeling, intrinsic evaluation, instruction following and reasoning, spanning a wide range of languages from high-resource to low-resource. In total, GlotEval comprises 9 tasks and 27 benchmarks, evaluates in over 1500 languages, and utilizes diverse metrics. Refer to Appendix A for more details of supported benchmark datasets.

3.2 Workflow

As shown in Figure 1, the workflow of GlotEval proceeds from specifying which benchmarks and languages to use, to producing final metrics and visualization.

First, users specify their choices and through command-line arguments. Users can specify the language(s) and the benchmark task(s) to evaluate. Besides, as for prompting strategy choice, GlotEval supports two prompting strategies: Setting prompting strategy as *single* along with a chosen prompt language (e.g., `eng_Latn`) applies the same prompt in one single language for every dataset in one benchmark. This is useful for controlling variables or using a single reference prompt style; Setting prompting strategy as *multi* makes GlotEval search for a language-specific template in the prompt library, which corresponds to the tested language, falling back to English if not found. Especially in machine translation tasks, the source language typically determines the prompt's language by default. Further, users can freely modify or expand the prompt library with a built-in multilingual prompt builder.

Upon selecting the desired benchmarks, languages, and prompt strategy, the user triggers GlotEval's data loader to automatically locate each dataset and load the relevant language subsets. It then initializes the appropriate model backend depending on the task type. Specifically, for non-generative tasks, we employ the HuggingFace Transformers backend (Wolf et al., 2020) to ensure more efficient use of computational resources. For generation tasks, such as machine translation, summarization, and open-ended text generation, we prioritize the vLLM backend (Kwon et al., 2023) to ensure high throughput, while retaining the HF Transformers generation interface for compatibility purposes.

After model inference is completed, GlotEval automatically computes evaluation metrics according to the task-specific settings listed in Table 1. Optionally, as mentioned before, by appending `-store_details`, users can export each sample's prompt, model output, reference, and corresponding scores to a JSONL file, which allows researchers to work outside the framework and conduct custom error analysis and result visualization. This ensures that our framework is not just an eval-

[3]https://github.com/openai/evals

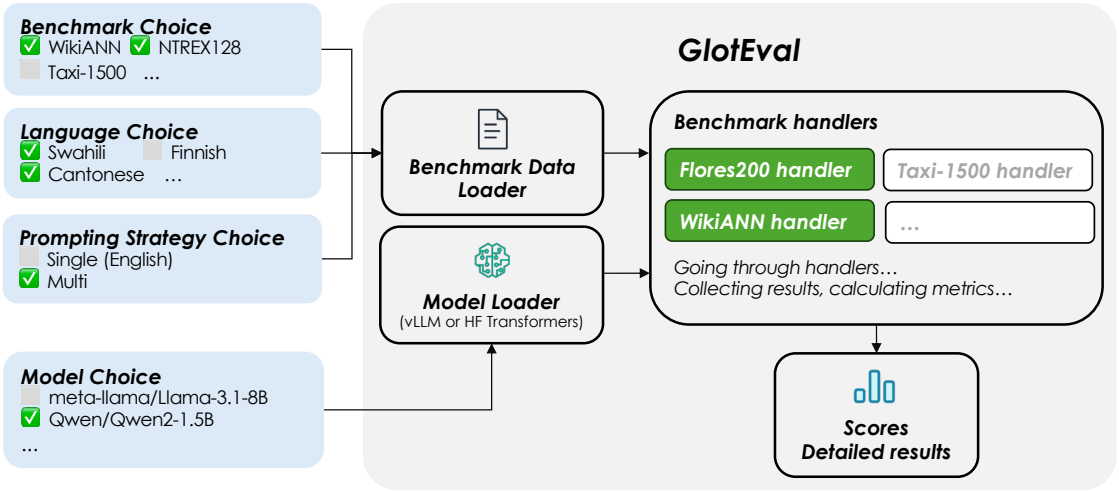


Figure 1: Workflow of GlotEval

uation executor, but also a starting point for more fine-grained analysis.

3.3 A Deeper Look at Benchmark Data Loader

Figure 2 illustrates the overall workflow within GlotEval's data loading and prompt preparation pipeline. At its core, GlotEval aligns language identifiers from various benchmarks to a unified `ISO 639-3_Script` format. Once the alignment is complete, the standardized language codes serve as the central connection for downstream operations. When the user queries GlotEval with a target language (e.g., `zho` for Chinese or `spa` for Spanish), the system consults the language-to-code dictionary and retrieves all benchmark-specific subsets whose original language codes map to the same standardized form. These subsets are then included in the evaluation process. Moreover, if a language-specific prompting strategy is selected, GlotEval uses the same aligned codes to retrieve the appropriate prompt templates from the multilingual prompt library. For example, as shown in Figure 2, querying `zho` and `spa` will automatically select the corresponding benchmark subsets and load their respective prompts (`zho_Hans`, `spa_Latn`) for evaluation. This workflow builds on both the language code alignment mechanism and the multilingual prompt builder described in the following sections.

Language Code Alignment to ISO 639-3

Different benchmarks often use inconsistent codes for the same language (e.g., `zh`, `zho`, `cmn`, `Chinese`, `Mandarin-CN` etc. for Mandarin Chinese). Before reading benchmark datasets via dedicated data loaders, GlotEval unifies these language identifiers used across different benchmarks, to enable cross-benchmark language-specific evaluation and prompting. Figure 3a visualizes this process.

Specifically, we process each benchmark-provided language code—which may appear in the form of ISO 639-1, 639-2/B (bibliographic), 639-2/T (terminological), ISO 639-3 codes, or even language names—by utilizing the `iso639-lang` Python package.[4] This allows us to retrieve all available mappings from the ISO 639-3 standard, including ISO 639-3 identifiers, ISO 639-2/B, 639-2/T, and ISO 639-1 codes. Using both exact and fuzzy matching strategies, we attempt to automatically identify the corresponding ISO 639-3 code for each language. A report is generated that documents, for each benchmark language, whether the match was exact or fuzzy, and whether it corresponds to an individual language or a macrolanguage in the ISO 639-3 standard.

We further identify the script used by each dataset, using `GlotScript` (Kargaran et al., 2024) to detect the dominant script.[5] Here we assume each dataset is primarily in one script. We randomly select up to 100 lines and attempt script recognition into ISO 15924 script code. This ensures each dataset obtains a `<language>_<script>` label, such as `eng_Latn`. The final ISO 639-3 code, along with the script code, is stored as the value in a language-to-code dictionary within the benchmark's configuration

[4] https://pypi.org/project/iso639-lang/
[5] https://pypi.org/project/GlotScript/

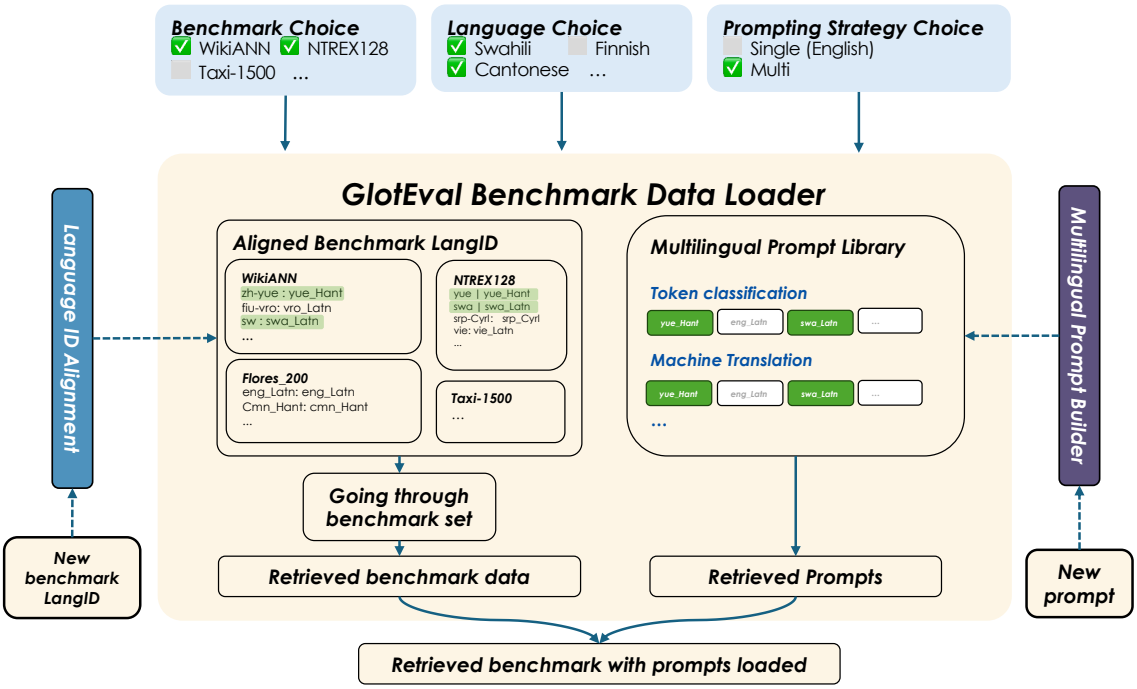


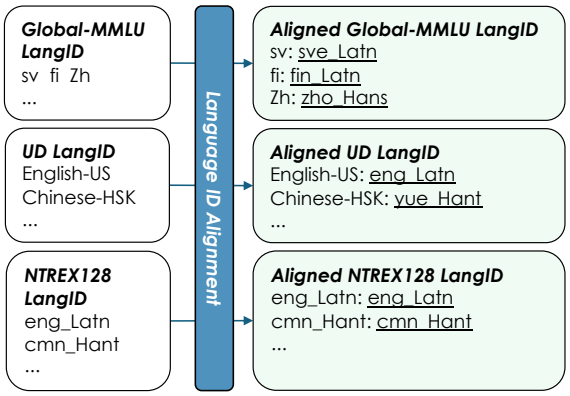
Figure 2: GlotEval benchmark data loader

file. Hence, each language + script combination is standardized in GlotEval for consistent usage across benchmarks.

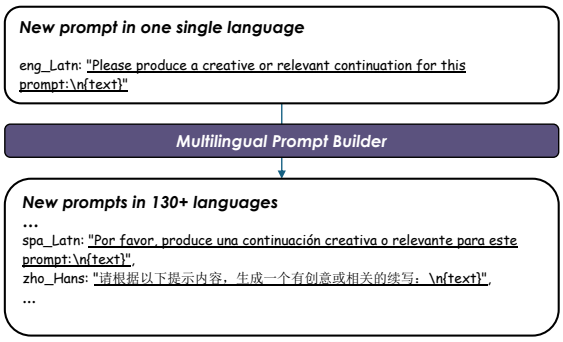
Multilingual Prompt Builder

We constructed a dedicated command-line prompt builder to automatically prepare or adapt prompt templates for multilingual tasks. Figure 3b visualizes this process. In particular, the builder leverages Microsoft Translator to convert an instruction and/or few-shot prompt template from a given source language into 130+ target languages, while ensuring that placeholders (e.g., `{src_text}`) remain intact during translation. These newly created multilingual prompts, are stored in the updated prompt library. As a result, each dataset's prompts are aligned with the same `<language>_<script>` language taxonomy, enabling consistent, language-specific evaluation.

Note that the automatic translation of prompts is intended as a convenience feature to support rapid, large-scale multilingual evaluation. While translation quality may vary—particularly for low-resource languages—this approach offers a practical starting point for exploratory analysis with language-specific prompts at scale. The framework remains fully customizable: users are able to provide their own human-written or verified prompts in the prompt library for languages of interest.



(a)



(b)

Figure 3: Benchmark data loader components: (a) Language ID alignment process and (b) multilingual prompt generation.

4 Evaluation

4.1 Efficiency Analysis

We benchmark GlotEval's inference speed on six tasks: FLORES-200, Aya, and XLSum for generative tasks, and SIB-200, Global-MMLU, and WikiANN for non-generative tasks. All evaluations are conducted on 19 languages spanning diverse writing systems (e.g., Latin, Arabic, Cyrillic, Devanagari, Chinese, etc.). For each language, we sample 10 examples per task for evaluation. We choose Qwen2-1.5B model (Yang et al., 2024) for evaluation. For generative tasks, we measure generation throughput (prefilling and decoding) with vLLM backend. For non-generative tasks, we measure classification throughput (prefilling only) using HF Transformers.

We consider two GPU environments:

- **AMD MI250X 64GB** (BF16, single GPU, batch size set as 1)
- **NVIDIA A100 40GB** (BF16, single GPU, batch size set as 1)

For detailed throughput performance, Appendix B shows statistics on both GPU environments. They demonstrate that in general, NVIDIA A100 consistently achieves higher throughput than AMD MI250X across both generative and non-generative tasks. Besides, this gap may also reflect the different backends between vLLM and HF Transformers. We further observe that scripts such as Devanagari or Amharic (`amh_Ethi`) often have lower throughput, potentially due to more complex tokenization. Lastly, summarization tasks like XLSum typically involve longer inputs and outputs than sentence-level translation tasks (e.g., FLORES-200), which increases the prefilling overhead and thus reduces the overall tokens/s.

4.2 Case Study on Multilingual Translation

To further illustrate GlotEval's capabilities, we conducted a detailed case study comparing EMMA-500 (Ji et al., 2024), a large-scale multilingual language model designed to enhance multilingual performance, with the base Llama-2-7B model (Touvron et al., 2023) across various multilingual translation scenarios. This study aimed to investigate performance differences under different prompting strategies and diverse language-centric translation tasks. We designed a factorial experiment with the following variables:

- **Models**: EMMA-500 vs. Llama-2-7B
- **Prompting strategies**: multilingual prompting (source language-specific), Chinese prompting (zho_Hans), Finnish prompting (fin_Latn), and English prompting (eng_Latn)
- **Translation directions**: six configurations with different central languages (X→eng-US, eng-US→X, zho-CN→X, X→zho-CN, fin→X, X→fin)

A demonstration of prompt templates is shown in Table 2. For evaluation, we utilized NTREX-128, a multi-aligned benchmark containing parallel texts across 128 languages, which is supported in GlotEval. In the multilingual prompting condition, we used built-in prompt builder in GlotEval, with the support of Microsoft Translator service, to automatically translate prompts into 134 languages supported by their platform. In our case study, 106 of these languages overlap with NTREX-128 languages, allowing us to test performance across this diverse language set.

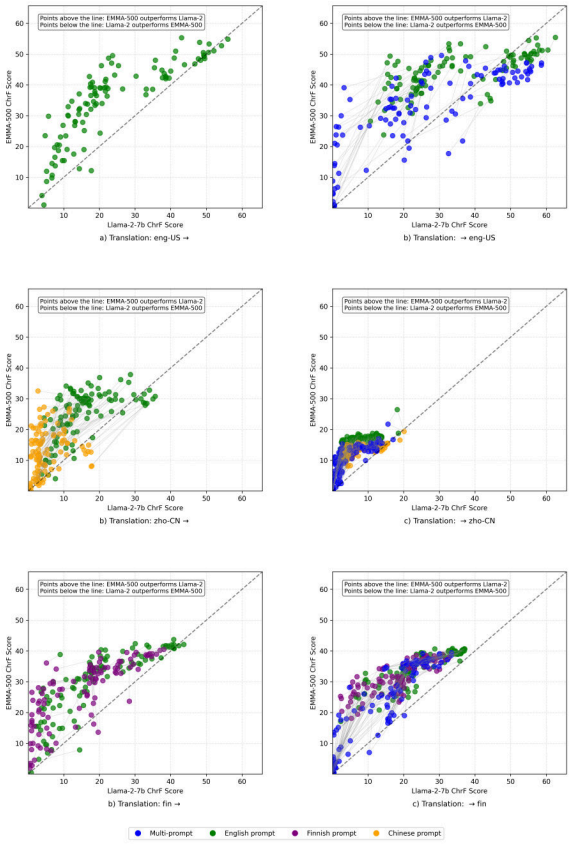


Figure 4: ChrF scores for different translation directions comparing EMMA-500 and Llama-2-7B across four prompting strategies.

The results of our case study (Figure 4) clearly demonstrate EMMA-500's performance compared to Llama-2-7B in multilingual instruction follow-

Prompt Strategy	Tested Translation Language Pair	Prompt Template
`multi` language-specific	`fra → fin` French → Finnish	Traduisez la phrase suivante de **Langue française** en **Langue finnoise** [Langue française] : {source_text_in_finnish} [Langue finnoise] :
`fin_Latn` Finnish	`vie → zho-CN` Vietnamese → Chinese (Simplified)	Käännä seuraava lause **Vietnamin kieli** muotoon **Kiinan kieli (yksinkertaistettu)** [Vietnamin kieli]: {source_text_in_vietnamese} [Kiinan kieli (yksinkertaistettu)]:

Table 2: A demonstration of prompt templates of translation tasks in different prompt strategies.

ing capabilities and non-English-centric translation tasks. Specifically, EMMA-500 shows consistently higher ChrF scores across most language pairs for all six translation directions. This performance advantage is particularly pronounced when using non-English prompting strategies, highlighting EMMA-500's enhanced ability to process and respond to instructions in diverse languages.

The experimental design was implemented using GlotEval, which facilitated the systematic manipulation of variables through simple configuration settings. By simply modifying the prompting strategy parameter and central language settings in the multi-aligned MT benchmark configuration, we are able to comprehensively assess the language models' multilingual capabilities, including both instruction following and non-English-centric multilingual translation.

5 Conclusion and Future Work

In this work, we introduced GlotEval, a lightweight yet comprehensive framework for massively multilingual evaluation of LLMs. By supporting consistent multilingual benchmarking, incorporating language-specific prompt templates, and supporting flexible non-English-centric translation setups, GlotEval enables consistent assessments of LLMs in diverse linguistic contexts—including low-resource settings often neglected by traditional benchmarks. Our case study on multilingual machine translation with two LLMs illustrates the utility of GlotEval in revealing the strengths and weaknesses of multilingual LLMs and in identifying directions for future optimization. Overall, GlotEval aims to encourage more inclusive, transparent, and holistic evaluations of language models across a wide array of languages and tasks, thereby advancing robust multilingual NLP research.

As for future work, we plan to integrate more diverse and comprehensive multilingual benchmarks to better evaluate LLM performance. Plus, we will explore the integration of benchmarks that the synergistic combination of automatic and human evaluation; for example, this could achieved through our pilot development of a lightweight web interface that supports crowd-sourced and expert-driven evaluation to supplement the automatic evaluation.[6]

Ethical Considerations and Broader Impact

Ethical Considerations We strive to uphold the principles outlined in the ACL Code of Ethics. While GlotEval advances multilingual evaluation, several limitations remain. Many benchmarks still lack sufficient or high-quality data for truly low-resource languages, potentially skewing performance assessments. Additionally, as noted by Joshi et al. (2025), existing datasets often inherit cultural and linguistic biases, favoring dominant dialects or standardized language forms over regional or marginalized variants. Computational costs further constrain accessibility: large-scale evaluations are resource-intensive, posing barriers for smaller research teams. More critically, reference-free metrics introduce inherent biases, as they effectively pit one generative model against another (Deutsch et al., 2022). Such metrics struggle to capture fluency, accuracy, or cultural appropriateness, particularly in low-resource contexts where human judgments are essential.

Broader Impact GlotEval promotes equitable progress in NLP by enabling systematic evaluation of large language models (LLMs) across diverse languages. We aim to support researchers and developers in creating language technologies that serve diverse communities more effectively via a more inclusive and holistic evaluation suite.

Acknowledgments

This project is funded by the AI-DOC program hosted by Finnish Center of Artificial Intelligence (decision number VN/3137/2024-OKM-6).

The work has received funding from the European Union's Horizon Europe research and in-

[6]Source code and documentation are available at `https://github.com/MaLA-LM/GlotEval-HumanEval` and `https://gloteval-humaneval.readthedocs.io`

novation programme under grant agreement No 101070350 and from UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10052546], and the Digital Europe Programme under grant agreement No 101195233.

The authors wish to acknowledge CSC - IT Center for Science, Finland, the Leonardo and LUMI supercomputers, owned by the EuroHPC Joint Undertaking, for providing computational resources.

Sawal Devkota, Bhavani Sai Praneeth Varma Mantina, Ananda Sreenidhi, Mengjie Wang, and Samea Yusofi contributed to this project as part of the "Data Analysis Software Project for Natural Language" course at TU Darmstadt, under the guidance of Shaoxiong Ji. This teaching activity was funded by LOEWE Center DYNAMIC as part of the Hessian program for the promotion of cutting-edge research LOEWE under the grant number of LOEWE1/16/519/03/09.001(0009)/98.

References

Idris Abdulmumin, Sthembiso Mkhwanazi, Mahlatse S. Mbooi, Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Neo N. Putini, Miehleketo Mathebula, Matimba Shingange, Tajuddeen Gwadabe, and Vukosi Marivate. 2024. Correcting FLORES evaluation dataset for four African languages. In *Proceedings of the Ninth Conference on Machine Translation*, Miami, USA. Association for Computational Linguistics.

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, et al. 2022. A few thousand translations go a long way! leveraging pre-trained models for african news translation. pages 3053–3070.

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

AI4Bharat, Jay Gala, Pranjal A. Chitale, Raghavan AK, Sumanth Doddapaneni, Varun Gumma, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages.

Felermino Dario Mario Ali, Henrique Lopes Cardoso, and Rui Sousa-Silva. 2024. Expanding FLORES+ benchmark for more low-resource settings: Portuguese-Emakhuwa machine translation evaluation. In *Proceedings of the Ninth Conference on Machine Translation*, Miami, USA. Association for Computational Linguistics.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, et al. 2020. Tico-19: the translation initiative for covid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Laurent Bié, Aleix Cerdà-i Cucó, Hans Degroote, Amando Estela, Mercedes García-Martínez, Manuel Herranz, Alejandro Kohan, Maite Melero, Tony O'Dowd, Sinéad O'Gorman, Mārcis Pinnis, Roberts Rozis, Riccardo Superbo, and Artūrs Vasiļevskis. 2020. Neural translation for the European Union (NTEU) project. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 477–478, Lisboa, Portugal. European Association for Machine Translation.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass.

Ona de Gibert, Robert Pugh, Ali Marashian, Raul Vazquez, Abteen Ebrahimi, Pavel Denisov, Enora Rice, Edward Gow-Smith, Juan C. Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno Veliz, Ángel Lino Campos, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E. Ortega, Luis Chiruzzo, Arturo Oncevay, Shruti Rijhwani, Katharina von der Wense, and Manuel Mager. 2025. Findings of the AmericasNLP 2025 Shared Tasks on Machine Translation, Creation of Educational Material, and Translation Metrics for Indigenous Languages of the Americas. In *Proceedings of the 5th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2025)*, Albuquerque, New Mexico. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. On the limitations of reference-free evaluations of generated text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10960–10977, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Moussa Doumbouya, Baba Mamadi Diané, Solo Farabado Cissé, Djibrila Diané, Abdoulaye Sow, Séré Moussa Doumbouya, Daouda Bangoura, Fodé Moriba Bayo, Ibrahima Sory 2. Condé, Kalo Mory Diané, Chris Piech, and Christopher Manning. 2023. Machine translation for nko: Tools, corpora, and baseline results. In *Proceedings of the Eighth Conference on Machine Translation*, pages 312–343, Singapore. Association for Computational Linguistics.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Clémentine Fourrier, Nathan Habib, Hynek Kydlíček, Thomas Wolf, and Lewis Tunstall. 2023. Lighteval: A lightweight framework for llm evaluation.

Jay P Gala, Pranjal A Chitale, AK Raghavan, Varun Gumma, Sumanth Doddapaneni, Kumar M Aswanth, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, et al. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*, 2023.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, et al. 2023. A framework for few-shot language model evaluation. Zenodo.

Isai Gordeev, Sergey Kuldin, and David Dale. 2024. Flores+ translation and machine translation evaluation for the Erzya language. In *Proceedings of the Ninth Conference on Machine Translation*, Miami, USA. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Chaoqun He, Renjie Luo, Shengding Hu, Yuanqian Zhao, Jie Zhou, Hanghao Wu, Jiajie Zhang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2024. Ultraeval: A lightweight platform for flexible and comprehensive evaluation for llms. *Preprint*, arXiv:2404.07584.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Xu Huang, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, and Fei Yuan. 2025. Benchmax: A comprehensive multilingual evaluation suite for large language models. *Preprint*, arXiv:2502.07346.

Shaoxiong Ji, Zihao Li, Indraneil Paul, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O'Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, and Barry Haddow. 2024. EMMA-500: Enhancing massively multilingual adaptation of large language models. *arXiv preprint arXiv:2409.17892*.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2025. Natural language processing for dialects of a language: A survey. *ACM Comput. Surv.*, 57(6).

Amir Hossein Kargaran, François Yvon, and Hinrich Schütze. 2024. GlotScript: A resource and tool for low resource writing system identification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7774–7784, Torino, Italia. ELRA and ICCL.

Ali Kuzhuget, Airana Mongush, and Nachyn-Enkhedorzhu Oorzhak. 2024. Enhancing Tuvan language resources through the FLORES dataset. In *Proceedings of the Ninth Conference on Machine Translation*, Miami, USA. Association for Computational Linguistics.

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

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Chunlan Ma, Ayyoob ImaniGooghari, Haotian Ye, Renhao Pei, Ehsaneddin Asgari, and Hinrich Schütze. 2024. Taxi1500: A multilingual dataset for text classification in 1500 languages. *arXiv preprint arXiv:2305.08487*.

Mukhammadsaid Mamasaidov and Abror Shopulatov. 2024. Open Language Data Initiative: Advancing low-resource machine translation for Karakalpak. In *Proceedings of the Ninth Conference on Machine Translation*, Miami, USA. Association for Computational Linguistics.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163, Reykjavik, Iceland. European Language Resources Association.

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, Loïc Barrault, Gabriel M. Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik R. 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. *arXiv preprint arXiv:2207.04672*.

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. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Juan Antonio Perez-Ortiz, Felipe Sánchez-Martínez, Víctor M. Sánchez-Cartagena, Miquel Esplà-Gomis, Aaron Galiano Jimenez, Antoni Oliver, Claudi Aventín-Boya, Alejandro Pardos, Cristina Valdés, Jusép Loís Sans Socasau, and Juan Pablo Martínez. 2024. Expanding the flores+ multilingual benchmark with translations for Aragonese, Aranese, Asturian, and Valencian. In *Proceedings of the Ninth Conference on Machine Translation*, Miami, USA. Association for Computational Linguistics.

Wessel Poelman and Miryam de Lhoneux. 2024. The roles of english in evaluating multilingual language models. *Preprint*, arXiv:2412.08392.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, and *et al.* 2024a. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *arXiv preprint arXiv:2412.03304*.

Shivalika Singh, Freddie Vargas, 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, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 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.

Xiaoqing Ellen Tan, Prangthip Hansanti, Carleigh Wood, Bokai Yu, Christophe Ropers, and Marta R. Costa-jussà. 2024. Towards massive multilingual holistic bias. *Preprint*, arXiv:2407.00486.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Daniel Varab and Natalie Schluter. 2021. MassiveSumm: a very large-scale, very multilingual, news summarisation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10150–10161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Hongjian Yu, Yiming Shi, Zherui Zhou, and Christopher Haberland. 2024. Machine translation evaluation benchmark for Wu. In *Proceedings of the Ninth Conference on Machine Translation*, Miami, USA. Association for Computational Linguistics.

Shimao Zhang, Changjiang Gao, Wenhao Zhu, Jiajun Chen, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2024. Getting more from less: Large language models are good spontaneous multilingual learners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8037–8051. Association for Computational Linguistics.

A Benchmark Settings

A.1 Intrinsic Evaluation

Given the input $X = (x_0, x_1, \ldots, x_{n_t})$, the negative log-likelihood (NLL) is defined as:

$$\text{NLL} = -\sum_{i=1}^{n_t} \log p_\theta(x_i|x_{<i}) \tag{1}$$

while perplexity (PPL) is computed as:

$$\text{PPL} = \exp\{-\frac{1}{n_t}\sum_{i=1}^{n_t} \log p_\theta(x_i|x_{<i})\} \tag{2}$$

Intuitively, PPL evaluates a model's ability to predict tokens in a given corpus, with lower values indicating better performance. In contrast, NLL measures the overall likelihood of the corpus under the model. Notably, due to its length normalization, PPL is directly influenced by the tokenization scheme, whereas NLL remains unaffected. Therefore, we use NLL for model comparisons to ensure consistency across models with different tokenization methods.

We compute NLL by concatenating the input sentences and applying a strided sliding window of size 1024.

A.2 Machine Translation

FLORES+ This work builds upon previous efforts on multilingual machine translation and evaluation datasets (NLLB Team et al., 2024; Goyal et al., 2022; Guzmán et al., 2019; Doumbouya et al., 2023; AI4Bharat et al., 2023; Perez-Ortiz et al., 2024; Abdulmumin et al., 2024; Ali et al., 2024; Kuzhuget et al., 2024; Yu et al., 2024; Mamasaidov and Shopulatov, 2024; Gordeev et al., 2024).

AmericasNLP Only the development set is used, as the test set is not disclosed. Note that this dataset is aligned with Spanish, but not English.

Tatoeba (v2023-09-26) We keep only test sets with over 1,000 sentences.

BLEU In our experiments, BLEU scores are computed via `SacreBLEU` (Post, 2018) with the `flores200` tokenizer to quantify translation quality. The BLEU signature is:

```
nrefs:1 | case:mixed | eff:no | tok:flores200 | smooth:exp | version:2.4.2
```

COMET Users can specify the customized model in the configuration file. The default model is `Unbabel/wmt22-comet-da`.

ChrF with Gender ChrF with gender is an evaluation metric that calculates the standard chrF score separately for sentences marked with different grammatical genders (masculine and feminine). By comparing these scores, one can assess whether a translation system favors one gender form over the other, thereby revealing potential gender bias in its outputs.

A.3 Text Classification

In classification tasks, the model predicts by ranking logits for each category; candidate labels are tokenized, and the label corresponding to the token with the highest probability is selected.

B Throughput Statistics

GlotEval provides a uniform pipeline for measuring both decoding-heavy and classification-style tasks across different languages, scripts, and hardware setups. According to efficiency analysis conducted in section 4.1, table 3 and 4 show throughput results on both NVIDIA A100 40GB and AMD MI250X 64GB GPU environments.

Language	FLORES-200(Eng-X) (3-shot)	Aya (0-shot)	XLSum (0-shot)	SIB-200 (3-shot)	Global-MMLU (0-shot)	WikiANN (3-shot)
French (fra_Latn)	854 / 0.88 = 969.55	447 / 0.77 = 583.55	67 / 0.09 = 720.32	10 / 0.53 = 18.88	10 / 0.27 = 36.17	70 / 2.36 = 29.60
Swahili (swa_Latn)	1174 / 0.92 = 1274.74	812 / 0.80 = 1020.78	150 / 0.56 = 268.32	10 / 0.56 = 17.71	10 / 0.31 = 32.65	61 / 1.97 = 30.91
Vietnamese (vie_Latn)	1206 / 0.92 = 1304.01	443 / 0.76 = 581.40	172 / 0.74 = 233.62	10 / 0.48 = 20.99	10 / 0.26 = 37.87	74 / 2.41 = 30.66
Indonesian (ind_Latn)	776 / 0.87 = 893.11	259 / 0.75 = 346.56	308 / 0.75 = 411.16	10 / 0.53 = 18.91	10 / 0.28 = 35.65	54 / 2.02 = 26.79
Latin Scri.	4010 / 3.59 = 1116.99	1961 / 3.08 = 636.69	697 / 2.14 = 325.70	40 / 2.10 = 19.05	40 / 1.12 = 35.71	259 / 8.76 = 29.57
Kyrgyz (kir_Cyrl)	1174 / 0.93 = 1259.10	436 / 0.76 = 573.19	324 / 0.75 = 429.72	10 / 0.72 = 13.95	10 / 0.35 = 28.98	72 / 4.20 = 17.16
Russian(rus_Cyrl)	1280 / 1.86 = 688.45	551 / 0.77 = 712.23	339 / 0.67 = 507.16	10 / 0.53 = 18.92	10 / 0.28 = 35.25	71 / 3.51 = 20.20
Serbian (srp_Cyrl)	1118 / 0.92 = 1207.56	475 / 0.76 = 621.45	342 / 0.76 = 452.07	10 / 0.62 = 16.25	10 / 0.30 = 33.14	48 / 1.94 = 24.76
Ukrainian (ukr_Cyrl)	1083 / 0.91 = 1191.05	404 / 0.76 = 532.91	43 / 0.09 = 470.72	10 / 0.68 = 14.65	10 / 0.31 = 31.68	132 / 7.43 = 17.78
Cyrillic Scri.	4655 / 4.62 = 1007.58	1866 / 3.05 = 611.80	1048 / 2.27 = 461.67	40 / 2.55 = 15.69	40 / 1.24 = 32.26	323 / 17.08 = 18.91
Arabic (arb_Arab)	852 / 0.87 = 974.46	74 / 0.41 = 181.59	228 / 1.62 = 140.36	10 / 0.53 = 18.85	10 / 0.28 = 36.32	76 / 2.75 = 27.65
Persian (fas_Arab)	852 / 0.89 = 958.99	264 / 0.75 = 353.62	333 / 0.76 = 440.70	10 / 0.68 = 14.63	10 / 0.31 = 31.74	54 / 12.48 = 4.32
Arabic Scri.	1704 / 1.76 = 968.18	338 / 1.16 = 291.38	561 / 2.38 = 235.71	20 / 1.21 = 16.53	20 / 0.59 = 33.90	130 / 15.23 = 8.54
Bengali (ben_Beng)	1143 / 0.96 = 1190.74	973 / 0.81 = 1195.97	260 / 0.71 = 366.53	10 / 1.26 = 7.91	10 / 0.45 = 21.99	39 / 2.13 = 18.32
Hindi (hin_Deva)	1167 / 0.96 = 1210.17	960 / 0.81 = 1182.68	223 / 0.75 = 296.00	10 / 1.10 = 9.07	10 / 0.39 = 25.62	52 / 2.50 = 20.79
Nepali (npi_Deva)	1250 / 1.01 = 1247.45	803 / 0.80 = 1009.63	231 / 0.60 = 384.25	10 / 1.02 = 9.78	10 / 0.41 = 24.57	69 / 3.98 = 17.32
Devanagari	3560 / 2.93 = 1215.02	2736 / 2.42 = 1130.58	714 / 2.06 = 346.60	30 / 3.38 = 8.88	30 / 1.25 = 24.00	160 / 8.61 = 18.58
Sinhala (sin_Sinh)	1280 / 1.04 = 1226.21	1280 / 0.86 = 1485.77	103 / 0.17 = 601.67	10 / 1.57 = 6.38	10 / 0.52 =19.38	69 / 5.43 = 12.70
Telugu (tel_Telu)	1208 / 1.02 = 1188.70	559 / 0.80 = 697.54	74 / 0.14 = 537.25	10 / 1.57 = 6.38	10 / 0.55 = 18.21	71 / 8.01 = 8.86
Amharic (amh_Ethi)	1280 / 1.00 = 1278.40	1280 / 0.85 = 1498.47	65 / 0.09 = 700.75	10 / 1.00 = 9.95	10 / 7.37 = 1.36	53 / 10.31 = 5.14
Japanese (jpn_Jpan)	714 / 0.87 = 820.25	152 / 0.21 = 707.20	274 / 0.75 = 365.11	10 / 0.48 = 21.01	10 / 0.28 = 35.99	389 / 33.70 = 11.54
Korean (kor_Hang)	1016 / 0.90 = 1129.38	284 / 0.76 = 374.84	59 / 0.12 = 493.29	10 / 0.54 = 18.58	10 / 0.27 = 36.78	91 / 5.20 = 17.50
Chinese (zho_Hans)	676 / 0.87 = 780.69	403 / 0.62 = 651.94	59 / 0.12 = 491.30	10 / 0.41 = 24.11	10 / 0.26 =37.60	419 / 42.26 = 9.91

Table 3: Throughput with NVIDIA A100 40GB GPU. Each cell contains: $\frac{\#generated\ tokens}{wall\ time\ (seconds)} = average\ tokens/s$.

Language	FLORES-200(Eng-X) (3-shot)	Aya (0-shot)	XLSum (0-shot)	SIB-200 (3-shot)	Global-MMLU (0-shot)	WikiANN (3-shot)
French (fra_Latn)	800 / 1.53 = 524.33	409 / 1.34 = 304.24	164 / 1.01 = 161.69	10 / 29.00 = 0.34	10 / 39.30 = 0.25	70 / 38.18 = 1.83
Swahili (swa_Latn)	1039 / 1.55 = 670.79	136 / 0.43 = 317.94	226 / 0.93 = 244.26	10 / 26.71 = 0.37	10 / 38.61 = 0.26	61 / 38.21 = 1.60
Vietnamese (vie_Latn)	932 / 1.53 = 608.26	675 / 1.39 = 485.18	58 / 0.15 = 379.43	10 / 31.58 = 0.32	10 / 39.46 = 0.25	74 / 38.13 = 1.94
Indonesian (ind_Latn)	1076 / 1.52 = 706.44	779 / 1.40 = 555.64	262 / 1.29 = 203.48	10 / 29.33 = 0.34	10 / 39.14 = 0.26	54 / 37.16 = 1.45
Latin Scri.	3847 / 6.13 = 627.57	1999 / 4.56 = 438.38	710 / 3.38 = 210.06	40 / 29.20 = 1.37	40 / 39.22 = 1.02	259 / 37.98 = 6.82
Kyrgyz (kir_Cyrl)	1051 / 1.57 = 669.63	344 / 1.32 = 261.10	444 / 1.36 = 325.96	10 / 17.18 = 0.58	10 / 37.68 = 0.27	72 / 23.48 = 3.07
Russian(rus_Cyrl)	1280 / 1.86 = 686.47	442 / 1.37 = 322.04	243 / 1.03 = 234.98	10 / 29.37 = 0.34	10 / 39.04 = 0.26	71 / 30.55 = 2.32
Serbian (srp_Cyrl)	1210 / 1.58 = 767.56	560 / 1.38 = 406.04	261 / 1.17 = 222.39	10 / 19.57 = 0.51	10 / 38.61 = 0.26	48 / 36.64 = 1.31
Ukrainian (ukr_Cyrl)	939 / 1.55 = 607.67	378 / 1.33 = 284.88	103 / 0.42 = 244.48	10 / 17.34 = 0.58	10 / 38.31 = 0.26	132 / 23.54 = 5.61
Cyrillic Scri.	4480 / 6.56 = 682.93	1724 / 5.40 = 319.26	1051 / 3.98 = 264.07	40 / 19.90 = 2.01	40 / 38.10 = 1.05	323 / 26.24 = 12.31
Arabic (arb_Arab)	919 / 1.54 = 595.36	160 / 1.24 = 129.25	83 / 0.26 = 318.11	10 / 29.06 = 0.34	10 / 39.23 = 0.25	76 / 37.56 = 2.02
Persian (fas_Arab)	929 / 1.55 = 600.20	16 / 0.12 = 131.16	184 / 1.21 = 152.61	10 / 17.43 = 0.57	10 / 38.25 = 0.26	54 / 13.24 = 4.07
Arabic Scri.	1848 / 3.09 = 598.06	176 / 1.36 = 129.41	267 / 1.47 = 181.63	20 / 21.98 = 0.91	20 / 39.22 = 0.51	130 / 21.35 = 6.09
Bengali (ben_Beng)	1130 / 1.62 = 698.59	1026 / 1.40 = 734.29	178 / 1.21 = 147.66	10 / 11.17 = 0.90	10 / 27.49 = 0.36	39 / 28.34 = 1.38
Hindi (hin_Deva)	1160 / 1.62 = 714.58	650 / 1.41 = 462.11	186 / 1.21 = 154.24	10 / 12.17 = 0.82	10 / 34.96 = 0.29	52 / 31.38 = 1.66
Nepali (npi_Deva)	1280 / 1.66 = 768.85	1126 / 1.40 = 805.59	275 / 1.10 = 250.46	10 / 13.00 = 0.77	10 / 34.68 = 0.29	69 / 2.77 = 24.87
Devanagari	3570 / 4.90 = 728.57	2802 / 4.21 = 665.56	639 / 3.52 = 181.53	30 / 12.05 = 2.49	30 / 31.91 = 0.94	160 / 5.73 = 27.91
Sinhala (sin_Sinh)	1280 / 1.76 = 727.47	1223 / 1.42 = 858.97	140 / 0.93 = 151.08	10 / 9.15 = 1.09	10 / 25.60 = 0.39	69 / 15.70 = 4.40
Telugu (tel_Telu)	1280 / 1.73 = 737.77	507 / 1.45 = 348.78	198 / 0.92 = 214.92	10 / 9.14 = 1.09	10 / 24.87 = 0.40	71 / 11.99 = 5.92
Amharic (amh_Ethi)	1280 / 1.66 = 772.22	1153 / 1.40 = 821.85	211 / 1.12 = 189.18	10 / 13.04 = 0.77	10 / 34.30 = 0.29	53 / 32.18 = 1.65
Japanese (jpn_Jpan)	690 / 1.51 = 458.09	266 / 1.27 = 209.63	250 / 1.02 = 244.87	10 / 31.85 = 0.31	10 / 39.23 = 0.25	389 / 14.98 = 25.96
Korean (kor_Hang)	973 / 1.53 = 633.96	468 / 1.38 = 340.21	204 / 1.07 = 191.09	10 / 28.74 = 0.35	10 / 39.33 = 0.25	91 / 25.01 = 3.64
Chinese (zho_Hans)	823 / 1.52 = 540.58	248 / 1.00 = 248.61	109 / 0.39 = 276.83	10 / 35.54 = 0.28	10 / 39.37 = 0.25	419 / 13.88 = 30.20

Table 4: Throughput with AMD MI250X 64GB GPU. Each cell contains: $\frac{\#generated\ tokens}{wall\ time\ (seconds)} = average\ tokens/s$.