Multi-lingual Functional Evaluation for Large Language Models

Victor Ojewale[1], **Inioluwa Deborah Raji**[2], **Suresh Venkatasubramanian**[1]

[1]The Center for Tech Responsibility, Brown University, USA
[2]University of California, Berkeley, USA

Abstract

Multi-lingual competence in large language models is often evaluated via static data benchmarks such as Belebele, M-MMLU and M-GSM. However, these evaluations often fail to provide an adequate understanding of the practical performance and robustness of models across multi-lingual settings. In response, we create multi-lingual *functional* benchmarks – Cross-Lingual Grade School Math Symbolic (CL-GSM Symbolic) and Cross-Lingual Instruction-Following Eval (CL-IFEval) – by translating existing functional benchmark templates from English to five additional languages that span the range of resources available for NLP: French, Spanish, Hindi, Arabic and Yoruba.

Our results show that the gap between static and functional evaluations is highly uneven across models and languages: the drop from Belebele to CL-IFEval ranges from 15% to over 35% depending on model and language, while M-GSM and CL-GSMSym scores diverge substantially across models, with some models scoring higher on the functional benchmark and others lower. Critically, functional benchmarks not only expose larger language performance gaps than static ones, but also produce *different model rankings* – suggesting that static benchmark performance is an incomplete proxy for multilingual deployment readiness. We further find that model robustness varies significantly across languages, with failures concentrated in specific instruction categories and problem families rather than distributed uniformly.

1 Introduction

Despite some meaningful progress, models operating in languages other than English have been regularly found to be more biased (Talat et al., 2022), less safe (Yong et al., 2023) and overall meaningfully less performant and robust (Ojo et al., 2023).

Popular multi-lingual LLM evaluations, such as Multilingual MMLU (M-MMLU) and the Multi-lingual Grade School Math (M-GSM) benchmark (Hendrycks et al., 2021; Shi et al., 2022; Cobbe et al., 2021), while useful, often fail to capture more meaningful indications of *functional* multi-lingual model performance – that is, the robust execution of a given prompt across a variety of languages (see Figure 1). In this paper, we extend the scope of two English functional evaluation datasets – IFEval (Zhou et al., 2023) and GSM-Symbolic (Mirzadeh et al., 2024) – by translating their prompt templates into five additional languages: French, Spanish, Hindi, Arabic and Yoruba.

Our experiments show that the relationship between functional and static benchmark performance varies significantly across models, languages, and benchmarks. While models often score higher on static benchmarks than on functional evaluations, the size of this gap varies substantially, with some static benchmarks aligning much more closely with functional performance than others. More importantly, model *rankings* are not stable across evaluation paradigms: a model that appears superior under static evaluation may rank substantially lower under functional evaluation for the same languages. Language performance gaps also vary significantly across settings as their magnitude and direction shift across specific languages and models. Finally, functional benchmarks surface robustness failures that are concentrated in specific instruction categories and problem families, revealing that models can be stable for certain task types yet brittle for others, regardless of language resource level.

2 Related Work

Common multi-lingual data benchmarks such as M-MMLU (Hendrycks et al., 2021; Lai et al., 2023a), FLORES (Goyal et al., 2022), BeleBele (Bandarkar et al., 2024), and XLSum (Hasan et al., 2021) pos-

Findings of the Association for Computational Linguistics: ACL 2026, pages 34672–34691
July 2-7, 2026

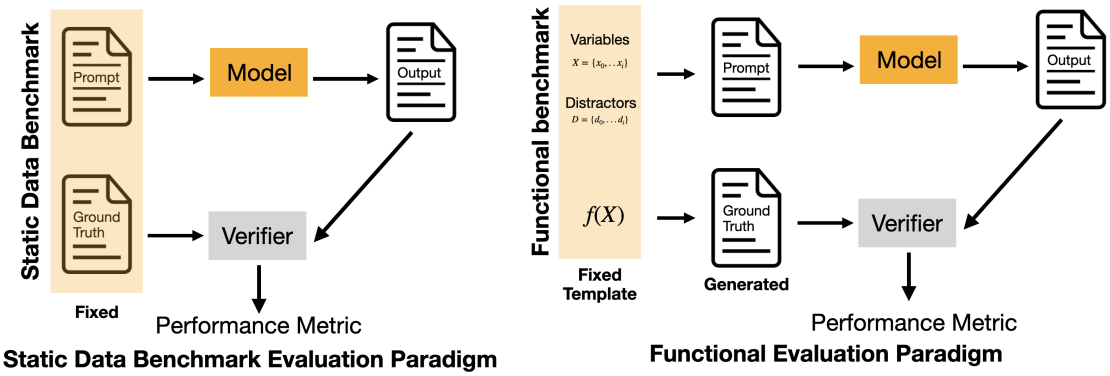

Figure 1: **Description of the functional evaluation paradigm.** Unlike with static data benchmarks, in the functional evaluation paradigm, model input prompts are not fixed but generated through a fixed template and a set of variables X (modifiable prompt attributes meant to impact model outputs) and a set of distractors D (modifiable prompt attributes meant to be ignored). The ground truth in this setting is generated through a fixed functional transformation $f(X)$. For instance, the prompt "Sally bought 2 red apples and 3 green apples. How much fruit did Sally buy?" is generated from the fixed template "$\{name\}$ bought $\{n_1\}$ $\{color_1\}$ apples and $\{n_2\}$ $\{color_2\}$ apples. How much fruit did $\{name\}$ buy?". This template involves the variables $X = \{n_1, n_2\}$ and the distractors $D = \{name, color_1, color_2\}$. The correct fixed output function in this case is $f(X) = n_1 + n_2$.

sess known limitations. The use of direct translation for many of these benchmarks has been critiqued as being devoid of realistic cultural context (Romanou et al., 2024; Singh et al., 2024). Furthermore, research reveals that English language benchmark data contamination might distort reported benchmark performance in English or possibly additional languages (as is the case for MMLU (Dodge et al., 2021) and GSM (Zhang et al., 2024)).

Functional evaluation involves templating a common benchmark with modifiable variables. For example, GSM-Symbolic templates examples of the static data benchmark GSM8k (Cobbe et al., 2021) to generate input permutations. The ground truth for the math problems is then calculated using literal template-based functional mappings from input values to the expected output (see Figure 1). Recent work has attempted to set up similar symbolic annotations for natural language benchmarks (Hennigen et al., 2023), and we can see a similar verifiable, function-based template format with instruction-following benchmarks (Liu et al., 2024; Chang et al., 2023) such as the IFEval dataset Zhou et al. (2023). Although some concurrent work – a proposed Multi-lingual IFEval (M-IFEval) (Dussolle et al., 2025) – has attempted to translate the IFEval template to Spanish, French and Japanese, we have yet to see more systematic analysis of how such functional evaluations can inform better assessments of performance and robustness in multi-lingual deployment settings.

3 Benchmark Datasets and Templates

We construct multi-lingual *functional* benchmarks by translating (a) the **100** English GSM-Symbolic templates (Mirzadeh et al., 2024) and (b) the **541** English IFEval prompts (Zhou et al., 2023) each containing at least one verifiable instruction into French, Spanish, Yoruba, Hindi, and Arabic. Following Bang et al. (2023); Lai et al. (2023b), we refer to languages by ISO 639-1 codes and select them to span resource levels based on CommonCrawl coverage (Table 5).

Initial translations are produced with Google Translate (Wu et al., 2016). We then conduct **native-speaker validation** of all translated GSM-Symbolic templates and all translated IFEval prompts *before* any GSM-Symbolic item instantiation.

Validators and recruitment. For each language, we recruited at least **two native speakers** to validate translations. Validators were recruited primarily from students at a US-based university. For Yoruba, we additionally recruited a native speaker based in Nigeria to ensure dialectal and cultural appropriateness in a low-resource setting.

Quality control, rubric, and prompt filtering. Validation followed a structured rubric (Appendix F). Validators compared each translated prompt or template against the English source to ensure fidelity to meaning, intent, and constraint strength, and to verify that instruction realizations

Benchmark	Avail. Lang	Prompts/Lang	Templates	Samples/Template	Task
Functional Benchmarks					
Cross-Lingual GSM Symbolic (Ours)	5	5000	100	50	Math reasoning
Cross-Lingual IFEval (Ours)	5	371	7	31–163	Instruction following
Static Benchmarks					
Multilingual MMLU (Lai et al., 2023a)	26	13062	N/A	N/A	MC knowledge recall
Multilingual GSM (Shi et al., 2022)	10	250	N/A	N/A	Math reasoning
Belebele (Bandarkar et al., 2024)	122	900	N/A	N/A	MC reading comprehension

Table 1: Descriptive statistics of all analyzed benchmarks.

remained valid in the target language. Validators discussed and, when possible, corrected problematic translations, and also checked for misspellings, duplicated sentences, and template–instance alignment issues.

As part of this validation, validators flagged a subset of IFEval prompts whose instructions are not meaningfully evaluable in scripts without case distinction (e.g., uppercase/lowercase constraints in Arabic and Hindi). We removed these prompts to preserve cross-lingual comparability, yielding a **371**-prompt CL-IFEval set.

After validation, each GSM-Symbolic template was instantiated into **50 variants per language**, yielding 5,000 QA pairs per language, while the validated IFEval prompts constitute CL-IFEval. Further dataset details appear in Table 1.

CL-IFEval instruction categories. CL-IFEval prompts are grouped into seven verifiable instruction categories, defined in Table 2. Categories involving case distinction (`change_case`) were filtered for Arabic and Hindi as noted above.

CL-GSMSym template families. For robustness analysis, we use a representative subset of 10 templates drawn from the full 100-template CL-GSMSym set, selected to cover the main reasoning types present in the GSM-Symbolic dataset (Mirzadeh et al., 2024): simple counting, rate-based reasoning, probabilistic inference, chained operations, comparison, unit conversion, narrative arithmetic, geometry, sequential accumulation, and algebraic setup. These 10 families are defined in Table 3. Each template is instantiated into 50 variants per language by substituting numerical values and distractor attributes, yielding 500 controlled items per language for this analysis.

For comparison with multi-lingual *static* benchmarks, we evaluate on M-MMLU (Lai et al., 2023a; Hendrycks et al., 2021), M-GSM (Shi et al., 2022), and Belebele (Bandarkar et al., 2024). The Cross-Lingual GSM Symbolic and Cross-Lingual IFEval datasets are publicly available. [1] [2]

Further dataset details can be found in Table 1.

4 Model Evaluation

For tasks that benefit from step-by-step reasoning, we use Chain-of-Thought (CoT) prompting in the same language as the query where applicable. Unless otherwise stated, CL-GSMSym follows the native 8-shot GSM8K setting, and M-GSM uses the `native_cot` configuration. All evaluations are orchestrated with the EleutherAI LM Evaluation Harness (`lm-eval`), which standardizes prompts, decoding, and metrics (Gao et al., 2024).

Our evaluations cover six open-source large language models, including instruction-tuned and multilingual variants: Aya 23-35B, Aya Expanse-32B, Gemma-2-9B-it, Qwen3-8B, Mistral-7B-Instruct-v0.3, and Mixtral-8x7B-Instruct-v0.1. Model details are provided in Appendix A. Using CL-IFEval, we compute prompt-level strict/loose and instruction-level strict/loose accuracy as in Zhou et al. (2023). For CL-GSMSym, we instantiate **50 variants per template per language** and evaluate a random sample of 500 items per language.

5 Results

We organize our findings around three questions: (1) How do functional and static benchmark scores compare in absolute terms, and does the gap vary by model or language? (2) How do language performance gaps differ between evaluation paradigms? (3) Which instruction categories and problem families drive cross-lingual failures, and are these failures language-dependent or universal?

[1] https://huggingface.co/datasets/vojewale/Cross-lingualGSMSymbolic

[2] https://huggingface.co/datasets/vojewale/Cross-lingualIFEval

Category	Description
length_constraints	Response must contain a specified number of words, sentences, or paragraphs
keyword	Response must include or exclude specific words or phrases
punctuation	Constraints on punctuation use (e.g., no commas)
startend	Response must begin or end with a specified string
change_case	Constraints on capitalization (e.g., all-caps, capital word frequency); removed for Arabic and Hindi
language	Response must be written in a specified language
detectable_format	Response must follow a structural format (e.g., JSON, markdown, numbered list)

Table 2: **CL-IFEval instruction categories.** Each prompt contains one or more verifiable instructions from these categories. change_case instructions were removed for Arabic and Hindi, which lack upper/lower case distinction.

ID	Family	Description
T1	Percentage & conversion	Unit conversion (inches to feet) followed by percentage calculation
T2	Rate-based	Unit-rate and proportional reasoning
T3	Probabilistic	Comparing compound probabilities over discrete sample spaces
T4	Chained operations	Arithmetic chained across multiple dependent quantities
T5	Sequential accumulation	Multi-step accumulation with changing rates across intervals
T6	Ratio & proportion	Part-to-whole reasoning from a given ratio
T7	Narrative arithmetic	Multi-quantity multiplication with named distractors
T8	Multi-item aggregation	Summing weighted quantities and dividing by a threshold
T9	Fractional chaining	Sequential application of fractions to a whole
T10	Algebraic setup	Single-variable equation setup and solve

Table 3: **CL-GSMSym template families used in robustness analysis.** Ten representative families selected from the full 100-template set to cover the main reasoning types in GSM-Symbolic. Full template examples appear in Appendix H.

5.1 Average Performance and Model Ranking

Full accuracy results across all models and languages are reported in Appendix C (Tables 10 and 11) and Appendix D (Tables 12, 13, and 14). We highlight several patterns below.

Static benchmarks compress model differences. On static natural language benchmarks, the six models cluster closely for high-resource languages. On Belebele (Table 12), English scores span 79.4% - 93.3% and on M-MMLU (Table 13) the English range is 59.0% - 75.3%. The mathematical reasoning benchmark M-GSM (Table 14) shows a wider spread (English: 44.8% - 88.0%), but relative model ordering remains broadly stable across languages.

Functional benchmarks expose a clear two-tier split. Under functional evaluation the picture changes substantially. On CL-IFEval (Table 10), English strict prompt-level accuracy ranges from 56.1% (Mixtral-8x7B) to 87.6% (Qwen3-8B), a 31.5 percentage point spread compared to a 14 point spread on Belebele. On CL-GSMSym (Table 11), English accuracy spans from 48.6% (Mistral-7B) to 86.8% (Qwen3-8B). The models separate into two tiers: a better-performing group (Qwen3-8B, Aya-expanse-32B, Gemma-2-9b-it) and a lower-performing group (Aya-23-35B, Mixtral-8x7B, Mistral-7B), with a gap of roughly 15 - 20 percentage points between tiers even in English.

Model rankings shift between paradigms. Aya-23-35B outranks Gemma-2-9b-it on M-GSM for most languages, yet Gemma-2-9b-it consistently outperforms Aya-23-35B on CL-GSMSym (Tables 14, 11). Similarly, Aya-23-35B outranks Aya-expanse-32B on Belebele for Yoruba, but this advantage disappears on CL-IFEval (Tables 12, 10). A practitioner selecting a model for multilingual deployment based solely on static scores could therefore make systematically wrong choices.

A medium-resource language outperforms high-resource ones. For Aya-23-35B on CL-GSMSym, Arabic accuracy (61.2%) exceeds both English (56.0%) and French (55.0%) (Table 11). This is the only model–language combination where a medium-resource language outperforms

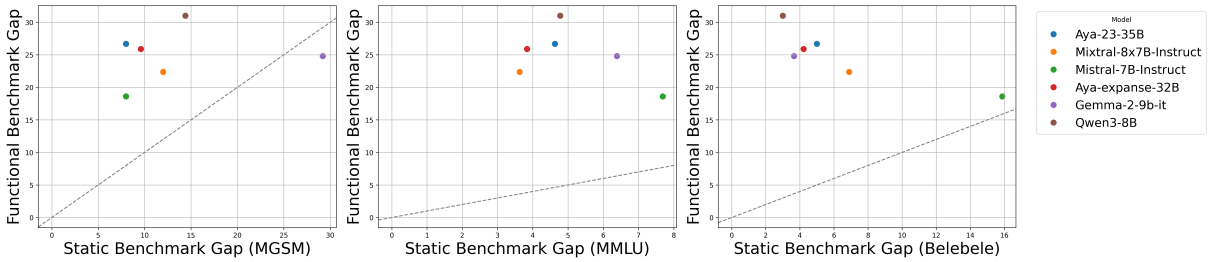

Figure 2: Correlation plots of performance gap between M-GSM, M-MMLU, and Belebele (left to right) versus CL-IFEval for high-resource languages only (en, fr, es). The language performance gap (difference between highest- and lowest-performing language) is consistently larger in functional evaluations than in static data benchmarks across all models.

all high-resource languages on a functional benchmark, and it persists across multiple template families rather than being driven by a single problem type. The static M-GSM benchmark, which does not include Arabic, does not surface this signal at all thereby illustrating how static benchmark coverage gaps can mask meaningful cross-lingual capability differences.

Low-resource language performance is uniformly poor and model-invariant. Yoruba presents a qualitatively different failure pattern. On CL-IFEval (Table 10), the full range across all six models is 12.4% - 22.9%, a spread of only 10.5 percentage points; in English the comparable spread is 31.5 points. When model choice barely affects performance, the bottleneck is data scarcity upstream of model design rather than architectural capability. This is qualitatively distinct from Arabic and Hindi, where model choice continues to matter substantially (CL-IFEval Arabic range: 25.3% - 50.1%; Hindi: 21.8%–49.9%).

5.2 Language Performance Gaps

We define the *language performance gap* as the difference in accuracy between a model's best- and worst-performing language within a given set. Table 4 reports these gaps for high-resource languages across all benchmarks.

Functional benchmarks reveal larger within-model language gaps than static benchmarks. Table 4 and Figure 2 show that across high-resource languages, functional benchmarks consistently expose larger language performance gaps than static ones. Aya-expanse-32B, a model explicitly marketed for multilingual capability, shows a gap of 4.22 points on Belebele but 25.88 points on CL-IFEval for the same three languages. Qwen3-8B shows a 3.00-point gap on Belebele and a 4.78-point gap on M-MMLU, yet a 31.00-point gap on CL-IFEval.

Model	CL-IFEval	CL-GSMSym	M-GSM	M-MMLU	Belebele
Qwen3-8B	31.00	8.6	14.4	4.78	3.00
Aya-exp-32B	25.88	6.2	9.6	3.84	4.22
Gemma-2-9b	24.80	27.4	29.2	6.39	3.66
Aya-23-35B	26.68	4.4	8.0	4.63	5.00
Mixtral-8x7B	22.36	7.0	12.0	3.63	6.89
Mistral-7B	18.60	14.0	8.0	7.69	15.88

Table 4: **Language performance gap (%) for high-resource languages (en, fr, es).** Gap = accuracy on best-performing language minus accuracy on worst-performing language. Extended gaps including medium- and low-resource languages appear in Appendix B.

The math gap is an exception to this pattern. For mathematical reasoning, the direction partially reverses: CL-GSMSym sometimes yields *smaller* language performance gaps than M-GSM. For Qwen3-8B, the CL-GSMSym high-resource gap is 8.6 points versus 14.4 points on M-GSM. The templated structure of CL-GSMSym, where number substitution is the primary variation, may partially decouple language resource level from performance on arithmetic tasks.

Gaps expand substantially when low-resource languages are included. Extending the analysis to include Yoruba (Table 8 in Appendix B) produces gaps that substantially exceed those in high-resource settings. Qwen3-8B's language performance gap reaches 69.81 points on CL-IFEval and 73.00 points on CL-GSMSym when Yoruba is included. Notably, Aya-expanse-32B shows the largest absolute gap of any model (56.87 points on CL-IFEval, 71.60 points on CL-GSMSym) not because its Yoruba performance is especially weak, but because its English performance is especially strong which is a reminder that gap metrics reflect both ends of the distribution.

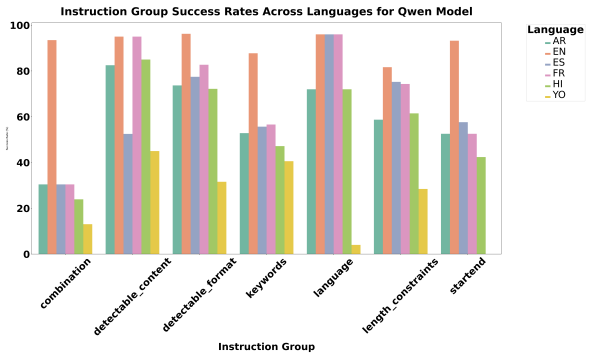

Figure 3: **CL-IFEval instruction-category success rates for Qwen3-8B across all languages.** Yoruba shows complete failure on `startend` across all models. `length_constraints` and `keyword` degrade more gradually, suggesting they generalize more robustly across languages. Equivalent plots for Aya-23-35B and Gemma-2-9b-it appear in Appendix B (Figures 5-6).

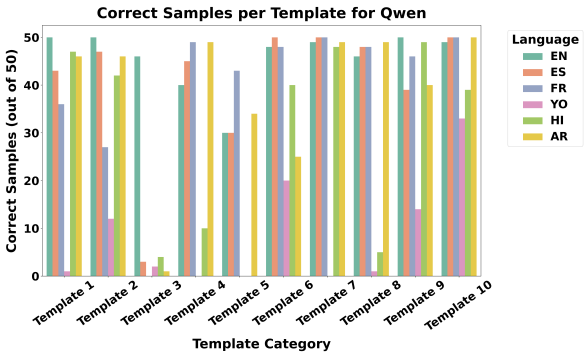

Figure 4: **CL-GSMSym per-template accuracy for Qwen3-8B across all languages.** Template 3 (probabilistic inference) is the weakest family across all languages including English. Templates 2, 4, and 10 show competitive performance in Arabic and Hindi relative to French and Spanish. Equivalent plots for Aya-23-35B and Gemma-2-9b-it appear in Appendix B (Figures 8-9).

5.3 Instruction-Level and Template-Level Robustness

Cross-lingual instruction-following breaks down unevenly across categories. Figures 5-7 show per-category success rates for all three models; Figure 3 (Qwen3-8B) is reproduced here for reference. Across all models, Yoruba shows complete failure on `startend`: no model produces a passing response in Yoruba for this category across any evaluation run. This failure is categorical rather than graded, suggesting models cannot reliably parse or apply positional string constraints in a language this far from their training distribution. By contrast, `length_constraints` and `keyword` degrade more gradually across languages, indicating these instruction types function more as language-agnostic operations that generalize even under low-resource conditions.

Probabilistic reasoning is a universal failure regardless of language. Figure 4 shows per-template accuracy for Qwen3-8B on CL-GSMSym. Template 3 (probabilistic inference, T3 in Table 3) produces the lowest accuracy of any template family for all three models and all six languages. Critically, this is not a low-resource phenomenon: even in English, all three models perform substantially worse on T3 than on any other template. Probabilistic reasoning is a domain-level weakness that functional benchmarking can isolate precisely because it controls for surface variation across template instances, a signal that static benchmarks, which do not template problem variants, cannot reliably surface.

Medium-resource languages are competitive on certain problem families. Against the general pattern of performance declining from high to low resource languages, Templates 2, 4, and 10 show Arabic and Hindi performance comparable to or exceeding French and Spanish for Aya-23-35B (Figure 8 in Appendix). For well-structured arithmetic templates with limited linguistic ambiguity, medium-resource languages do not incur the penalty they show on more linguistically demanding tasks. This finding has direct implications for model selection in Arabic and Hindi deployment contexts, where aggregate benchmark scores would underestimate actual task performance.

Summary. Functional benchmarks expose model weaknesses that static benchmarks obscure. The gap between paradigms is largest for instruction-following, most severe for low-resource languages, and concentrated in specific instruction categories (`startend`) and problem families (probabilistic reasoning, Template 3). Model rankings are not stable across paradigms and strong static benchmark performance does not imply robust multilingual task execution, and in at least one case (Aya-23-35B on Arabic), functional evaluation reveals a medium-resource language *outperforming* high-resource ones, a finding invisible to static benchmarks.

6 Conclusion

We introduce CL-IFEval and CL-GSMSym, two multilingual functional benchmarks covering six

languages that span the full range of NLP resource levels. Our experiments demonstrate that static benchmark performance is an unreliable proxy for multilingual deployment readiness: functional benchmarks expose larger language performance gaps, produce different model rankings, and surface failures concentrated in specific instruction categories and problem families. The `startend` instruction category shows complete failure in Yoruba across all models; probabilistic reasoning (Template 3) is universally the hardest problem family regardless of language; and Aya-23-35B on Arabic outperforms all high-resource languages on CL-GSMSym, a signal entirely invisible to existing static benchmarks. Both datasets are released publicly, providing the community with validated functional resources for multilingual evaluation including the extremely low-resource Yoruba.

Limitations. We use automated translation tools like Google Translate in constructing CL-IFEval and CL-GSMSym. While these tools offer broad language coverage and facilitate large-scale data generation, they introduce potential inaccuracies, particularly for lower resourced languages like Yoruba, and when dealing with conversions across metric and imperial measurement systems.

Another limitation is the emphasis in our analysis on open-weight models. Our primary focus remains on open-weight models such as Mixtral-8x7B, Mistral-7B, Gemma2-9B-it, Qwen3-8B and AYA models. This creates a possible inherent bias in the scope of our comparisons, as proprietary models may perform significantly better than their open-weight counter-parts. On the other hand, the consistency and transparency of open-weight models make them the preferable object of study – the incorporation of proprietary models can make results hard to reproduce reliably.

Future Work. Future directions include systematically curating higher-quality translations, expanding into multi-lingual code or multi-modal instruction evaluation, and further investigating the robustness and error patterns in both functional and static benchmarks. Also, as functional evaluations involve automatic verification, there is some possibility of extrapolating this framework in the training and evaluation of multi-lingual reasoning models (Yong et al., 2025).

Acknowledgments

The authors would like to thank Zheng-Xin Yong for initial feedback on the work, and Meredith Mendola and the language validators(full list on the respective dataset repositories) who contributed to the translation validation process. This work was supported in part by the MacArthur Foundation, the Mozilla Foundation, and the Heising-Simons Foundation.

References

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open Weight Releases to Further Multilingual Progress. *arXiv preprint*. ArXiv:2405.15032 [cs].

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

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

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. 2023. A Survey on Evaluation of Large Language Models. *arXiv preprint*. ArXiv:2307.03109 [cs].

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy,

Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*.

Antoine Dussolle, Andrea Cardeña Díaz, Shota Sato, and Peter Devine. 2025. M-ifeval: Multilingual instruction-following evaluation. *arXiv preprint arXiv:2502.04688*.

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

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*, 10:522–538. Place: Cambridge, MA Publisher: MIT Press.

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.

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)*.

Lucas Torroba Hennigen, Shannon Shen, Aniruddha Nrusimha, Bernhard Gapp, David Sontag, and Yoon Kim. 2023. Towards verifiable text generation with symbolic references. *arXiv preprint arXiv:2311.09188*.

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. *Preprint*, arXiv:2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023a. 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.

Viet Dac Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023b. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024. Mitigating hallucination in large multi-modal models via robust instruction tuning. *Preprint*, arXiv:2306.14565.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2024. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models.

Jessica Ojo, Kelechi Ogueji, Pontus Stenetorp, and David I Adelani. 2023. How good are large language models on african languages? *arXiv preprint arXiv:2311.07978*.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A Haggag, Alfonso Amayuelas, et al. 2024. Include: Evaluating multilingual language understanding with regional knowledge. *arXiv preprint arXiv:2411.19799*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung,

Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language Models are Multilingual Chain-of-Thought Reasoners. *arXiv preprint*. ArXiv:2210.03057 [cs].

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, et al. 2024. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *arXiv preprint arXiv:2412.03304*.

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, et al. 2022. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zheng-Xin Yong, M. Farid Adilazuarda, Jonibek Mansurov, Ruochen Zhang, Niklas Muennighoff, Carsten Eickhoff, Genta Indra Winata, Julia Kreutzer, Stephen H. Bach, and Alham Fikri Aji. 2025. Crosslingual reasoning through test-time scaling. *Preprint*, arXiv:2505.05408.

Zheng-Xin Yong, Cristina Menghini, and Stephen H Bach. 2023. Low-resource languages jailbreak gpt-4. *arXiv preprint arXiv:2310.02446*.

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Dylan Slack, Qin Lyu, et al. 2024. A careful examination of large language model performance on grade school arithmetic. *arXiv preprint arXiv:2405.00332*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-Following Evaluation for Large Language Models. *arXiv preprint*. ArXiv:2311.07911 [cs].

A Additional Experimental Details

Language	CC-MAIN-2025-05 (%)	Category
English (en)	43.37	High-resource (HRL)
French (fr)	4.52	High-resource (HRL)
Spanish (es)	4.64	High-resource (HRL)
Arabic (ar)	0.66	Medium-resource (MRL)
Hindi (hi)	0.20	Medium-resource (MRL)
Yoruba (yo)	0.0009	Extremely low-resource (X-LRL)

Table 5: Classification of languages based on resource availability in CommonCrawl (CC-MAIN-2025-05).

A.1 Models

Our evaluations spanned six open-source large language models:

- **Aya 23-35B**: A multilingual instruction-tuned model based on Cohere's Command framework (Aryabumi et al., 2024).
- **Aya Expanse-32B**: Part of the Aya Expanse series, designed to enhance multilingual performance capabilities (Dang et al., 2024).
- **Gemma-2-9B-it**: An instruction-tuned model trained on 8 trillion tokens from diverse sources including web documents, code, and mathematical text (Team et al., 2024).
- **Qwen3-8B**: A dense and mixture-of-experts (MoE) model supporting 100+ languages, with specialized modes for logical reasoning, code generation, and agent-based tasks (Yang et al., 2025).
- **Mistral-7B-Instruct-v0.3**: An open-source instruction-tuned variant of Mistral-7B (Jiang et al., 2023).
- **Mixtral-8x7B-Instruct-v0.1**: A sparse mixture-of-experts model with 12.9B active parameters per token, fine-tuned for instruction tasks (Jiang et al., 2024).

B Full Performance Gap Results

B.1 High-Resource Language Performance Gaps

Model	CL-IFEval	CL-GSMSym	M-GSM	M-MMLU	Belebele
Aya-23-35B	26.68	4.4	8.0	4.63	5.00
Mixtral-8x7B-Instruct	22.36	7.0	12.0	3.63	6.89
Mistral-7B-Instruct	18.60	14.0	8.0	7.69	15.88
Aya-expanse-32B	25.88	6.2	9.6	3.84	4.22
Gemma-2-9b-it	24.80	27.4	29.2	6.39	3.66
Qwen3-8B	31.00	8.6	14.4	4.78	3.00

Table 6: **Performance Gap on High-Resource Languages**, as measured by the benchmark accuracy difference between the results on the best performant language and worst performant language across the set of analyzed languages (en, fr, es).

B.2 High to Medium-Resource Language Performance Gaps

Model	CL-IFEval	CL-GSMSym	M-MMLU	Belebele
Aya-23-35B	26.68	17.6	18.08	22.45
Mixtral-8x7B-Instruct	31.41	33.2	30.65	36.78
Mistral-7B-Instruct	37.74	33.8	27.92	38.44
Aya-expanse-32B	28.03	11.0	15.29	15.45
Gemma-2-9b-it	28.04	27.4	18.20	19.55
Qwen3-8B	37.73	27.4	20.03	23.45

Table 7: **Performance Gap on High to Medium-Resource Languages**, as measured by the benchmark accuracy difference between the results on the best performant language and worst performant language across the set of analyzed languages (en, fr, es, ar, hi).

B.3 High to Extremely Low-Resource Language Gaps

Model	CL-IFEval	CL-GSMSym	Belebele
Aya-23-35B	51.21	57.4	54.12
Mixtral-8x7B-Instruct	43.66	56.8	57.11
Mistral-7B-Instruct	40.98	44.2	50.11
Aya-expanse-32B	56.87	71.6	60.45
Gemma-2-9b-it	52.29	64.4	52.44
Qwen3-8B	69.81	73.0	62.89

Table 8: **Performance Gap on High to X Low-Resource Languages**, as measured by the benchmark accuracy difference between the results on the best performant language and worst performant language across the set of analyzed languages (en, fr, es, ar, hi, yo).

C Full Functional Benchmark Results

Model	English	French	Spanish	Arabic	Hindi	Yoruba
AYA Exp-32B	74.93	55.52	49.05	46.90	49.86	18.06
AYA 23-35B	67.11	48.78	40.43	43.93	40.70	15.90
Mistral-7B-Instruct	59.57	41.78	40.97	25.33	21.83	18.59
Mixtral-8x7B-Instruct	56.06	36.39	33.70	31.53	24.65	12.40
Gemma-2-9B-it	75.20	53.36	50.40	47.16	48.24	22.91
Qwen3-8B	87.60	61.46	56.60	50.13	49.87	17.79

Table 9: Cross-Lingual-IFEval Prompt-level Strict performance across different models and languages.

Language	Model	Strict PL (%)	Strict IL (%)	Loose PL (%)	Loose IL (%)
English	AYA 23-35B	67.11	73.75	70.89	77.41
	Gemma2-9B-it	75.20	81.08	79.24	84.17
	AYA Exp-32B	74.93	79.92	78.97	83.39
	Mistral-7B	59.57	68.15	64.42	72.39
	Mixtral-8x7B	56.06	64.48	61.19	69.50
	Qwen3-8B	87.60	90.70	91.37	93.63
French	AYA 23-35B	48.78	58.30	53.90	62.94
	Gemma2-9B-it	53.36	61.38	56.06	63.89
	AYA Exp-32B	55.52	63.12	58.22	65.83
	Mistral-7B	41.78	51.35	46.90	56.18
	Mixtral-8x7B	36.39	45.37	39.62	49.03
	Qwen3-8B	61.46	69.11	63.07	70.46
Spanish	AYA 23-35B	40.43	50.77	44.20	54.24
	Gemma2-9B-it	50.40	59.07	51.75	61.00
	AYA Exp-32B	49.05	58.49	51.21	60.62
	Mistral-7B	40.97	50.96	43.66	53.86
	Mixtral-8x7B	33.70	43.47	37.20	47.00
	Qwen3-8B	56.60	65.06	57.41	65.83
Yoruba	AYA 23-35B	15.90	22.77	16.17	23.74
	Gemma2-9B-it	22.91	31.08	25.88	34.75
	AYA Exp-32B	18.06	28.19	29.48	30.31
	Mistral-7B	18.59	28.95	20.21	31.27
	Mixtral-8x7B	12.40	20.27	15.00	23.36
	Qwen3-8B	17.79	27.22	20.21	30.50
Arabic	AYA 23-35B	43.93	53.47	46.90	56.37
	Gemma2-9B-it	47.16	57.14	48.79	58.88
	AYA Exp-32B	46.90	56.76	49.06	58.69
	Mistral-7B	25.33	36.10	28.57	39.38
	Mixtral-8x7B	31.53	40.73	35.85	45.95
	Qwen3-8B	50.13	60.62	52.30	62.16
Hindi	AYA 23-35B	40.70	50.19	44.47	54.44
	Gemma2-9B-it	48.24	56.94	50.94	59.65
	AYA Exp-32B	49.86	58.69	52.56	61.78
	Mistral-7B	21.83	32.24	25.06	36.67
	Mixtral-8x7B	24.65	33.24	32.83	41.31
	Qwen3-8B	49.87	58.11	52.02	60.62

Table 10: Comprehensive instruction-following results across multiple models and languages. **PL** = Prompt-Level Accuracy; **IL** = Instruction-Level Accuracy.

Model	en	fr	es	ar	hi	yo
Aya-23-35B	56.00	55.00	59.40	61.20	43.60	3.80
Mixtral-8x7B-Instruct	56.80	55.00	62.00	32.40	28.80	5.20
Mistral-7B-Instruct	48.60	39.40	34.60	18.40	14.80	4.40
Aya-expanse-32B	82.60	76.40	82.20	76.80	71.60	11.00
Gemma-2-9b-it	77.00	49.60	54.80	55.20	57.00	12.60
Qwen3-8B	86.80	78.20	81.80	78.80	59.40	13.80

Table 11: CL-GSMSym (8-shot) accuracy (%) across models and languages.

D Full Static Data Benchmark Results

Model	en	fr	es	ar	hi	yo
Aya-23-35B	0.856	0.813	0.806	0.767	0.631	0.314
Mixtral-8x7B-Instruct	0.852	0.810	0.783	0.634	0.484	0.281
Mistral-7B-Instruct	0.794	0.671	0.636	0.512	0.410	0.293
Aya-expanse-32B	0.899	0.884	0.857	0.837	0.744	0.294
Gemma-2-9b-it	0.933	0.920	0.897	0.887	0.738	0.409
Qwen3-8B	0.926	0.909	0.896	0.870	0.691	0.297

Table 12: Belebele accuracy across models and languages.

Model	en	fr	es	ar	hi
Aya-23-35B	0.662	0.617	0.615	0.539	0.481
Mixtral-8x7B-Instruct	0.685	0.651	0.649	0.414	0.379
Mistral-7B-Instruct	0.590	0.513	0.518	0.333	0.311
Aya-expanse-32B	0.740	0.701	0.703	0.627	0.587
Gemma-2-9b-it	0.715	0.664	0.654	0.557	0.533
Qwen3-8B	0.753	0.705	0.709	0.607	0.553

Table 13: M-MMLU (5-shot) accuracy across models and languages. Yoruba not available in this benchmark.

Model	en	fr	es
Aya-23-35B	0.668	0.588	0.600
Mixtral-8x7B-Instruct	0.632	0.512	0.536
Mistral-7B-Instruct	0.448	0.444	0.368
Aya-expanse-32B	0.856	0.760	0.840
Gemma-2-9b-it	0.664	0.372	0.512
Qwen3-8B	0.880	0.736	0.852
GPT-4o-mini	0.656	0.556	0.432
Claude 3.5 Sonnet	0.900	0.716	0.812

Table 14: M-GSM (5-shot) accuracy (native CoT, strict match). Hindi, Arabic, and Yoruba not available in this benchmark.

E Robustness Plots

E.1 CL-IFEval Plots

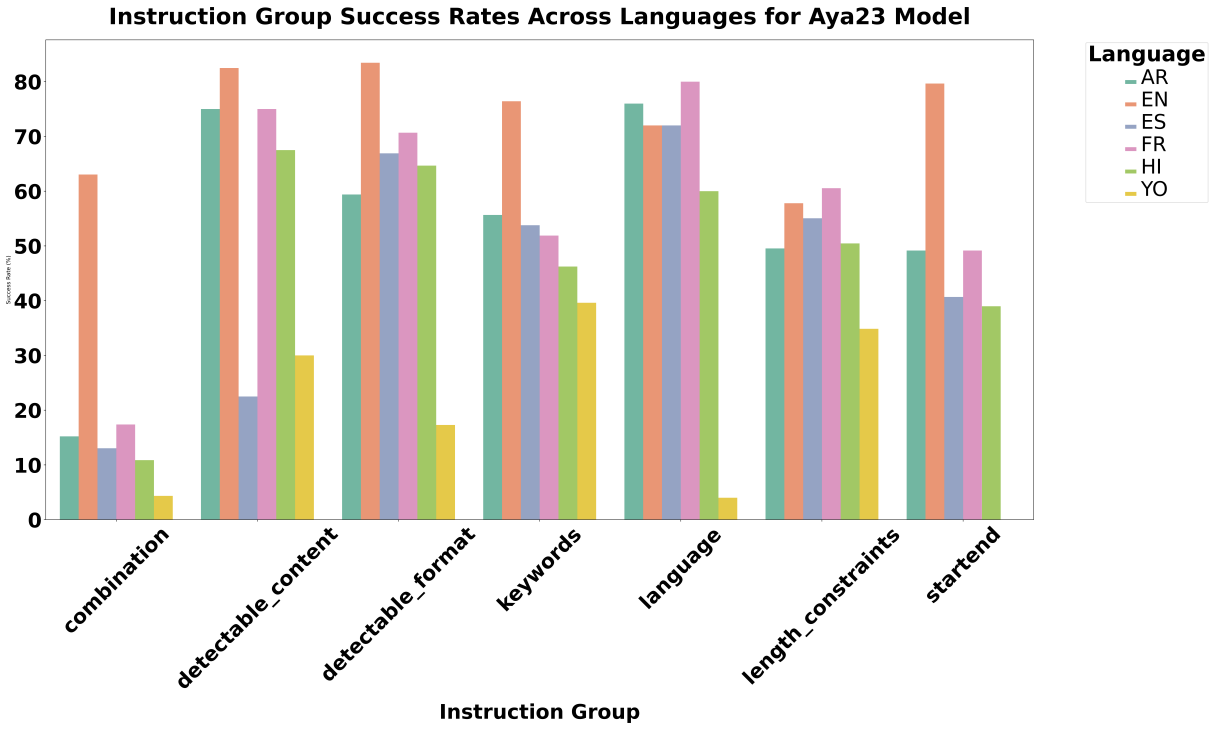

Figure 5: CL-IFEval instruction-category success rates for Aya-23-35B across all languages.

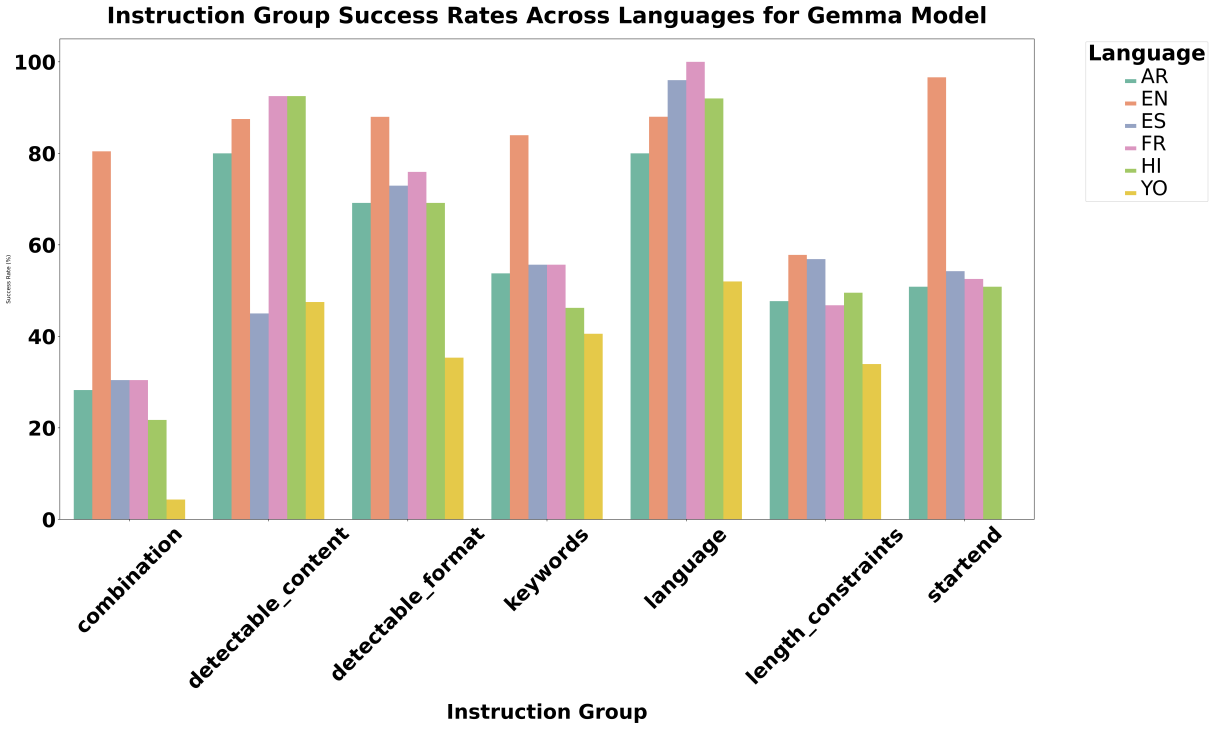

Figure 6: CL-IFEval instruction-category success rates for Gemma-2-9b-it across all languages.

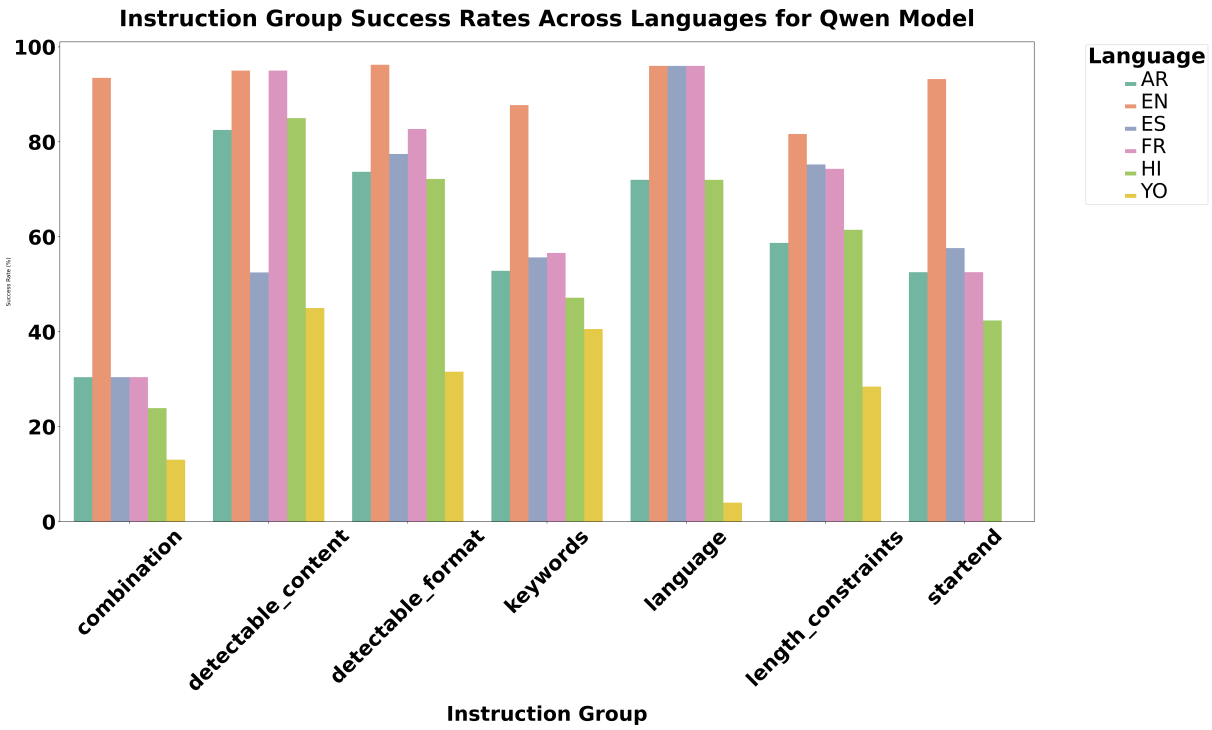

Figure 7: CL-IFEval instruction-category success rates for Qwen3-8b across all languages.

E.2 CL-GSMSym Plots

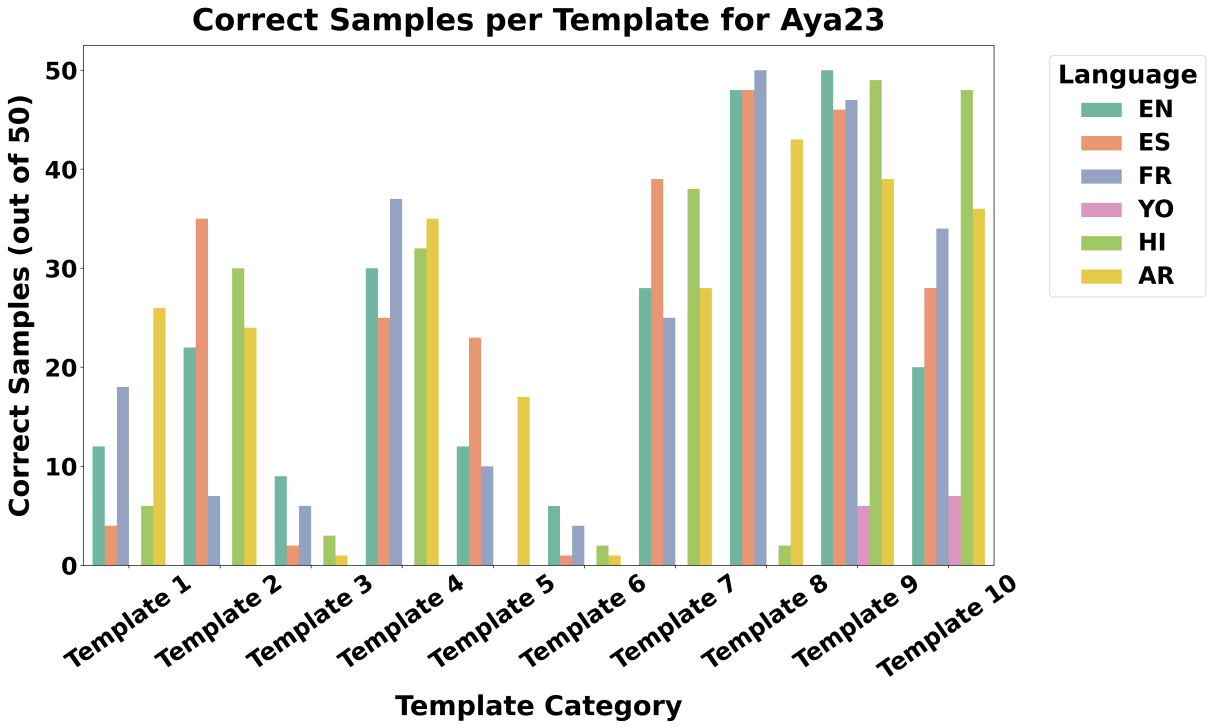

Figure 8: CL-GSMSym per-template accuracy for Aya-23-35B across all languages (50 samples, 10 templates).

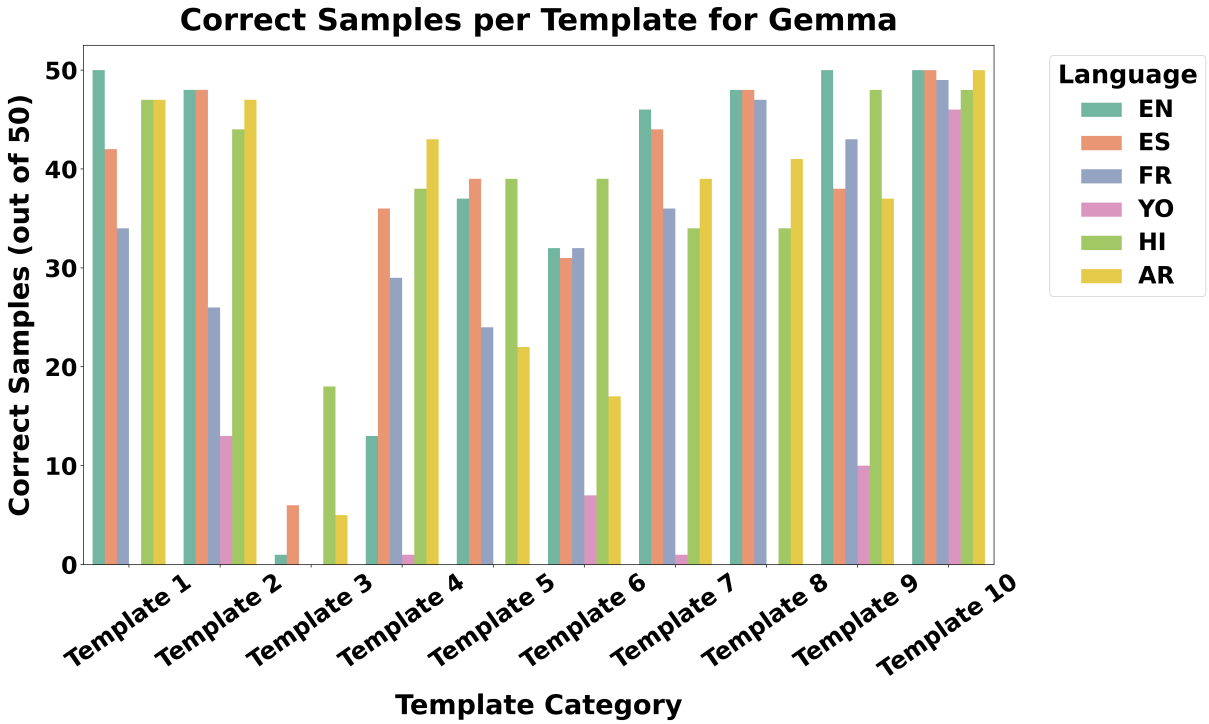

Figure 9: CL-GSMSym per-template accuracy for Gemma-2-9b-it across all languages (50 samples, 10 templates).

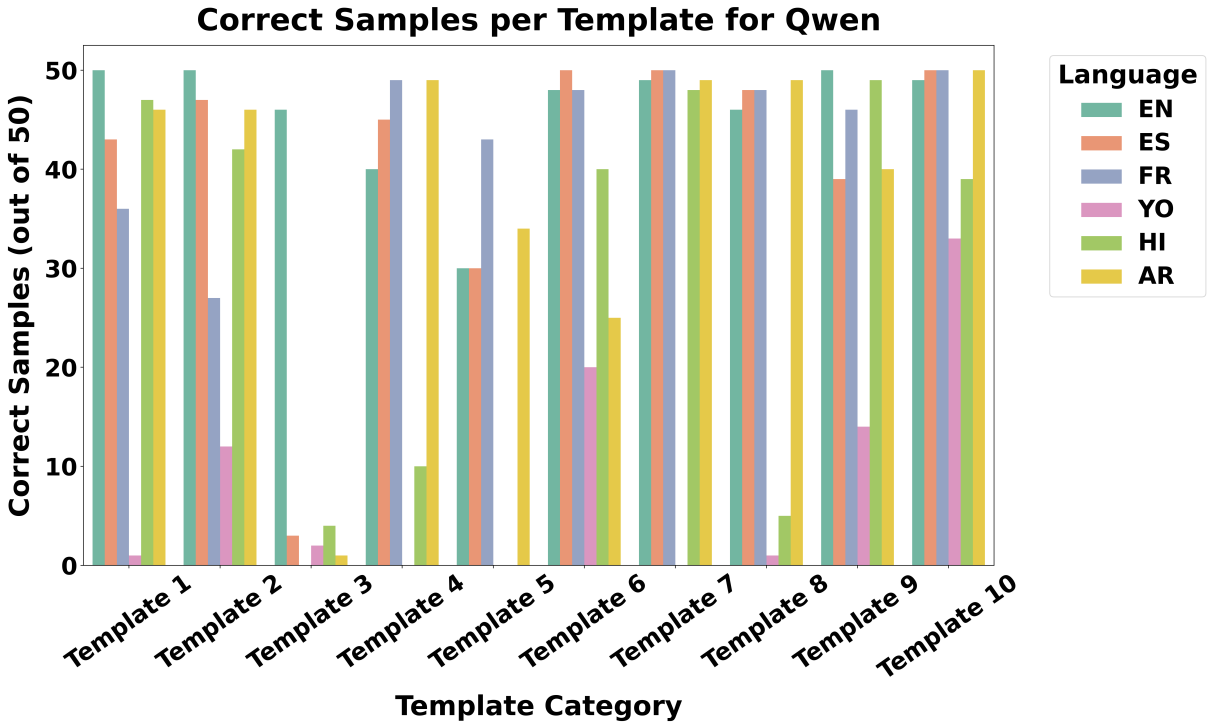

Figure 10: CL-GSMSym per-template accuracy for Qwen3-8b across all languages (50 samples, 10 templates).

F Translation Validation Rubrics

F.1 CL-IFEval Translation Validation (541 Prompts)

1. Objective

The goal of this validation task is to ensure that every translated prompt accurately represents the intent, structure, and constraints of the original English version while remaining clear and natural in the target language.
Annotators should:

- Confirm that the translation faithfully conveys the same meaning and instructions as the English source.
- Verify that every element in the instruction_id_list is valid and meaningful in the target language.
- Flag any instruction IDs that cannot be reasonably expressed or understood in that language.
- Correct minor translation errors or omissions directly in the prompt, but only when needed for consistency with the English version.
- Record every correction or flag for traceability.

2. Dataset Structure

Each entry contains:

- `key` – unique identifier for the instance
- `prompt` – translated text (the only part that differs from English)
- `instruction_id_list` – the list of instructions applied to the English prompt
- `kwargs` – parameters or values associated with those instructions

Example (simplified):

```
{
"key": 1001,
"prompt": "Translated prompt text...",
"instruction_id_list": ["punctuation:no_comma",
                        "length_constraints:number_words"],
"kwargs": [{}, {"num_words": 300}]
}
```

3. Validation Criteria

- **A. Translation Fidelity**: The translation must convey the same meaning, intent, and constraints as the English prompt. If missing, ambiguous, or distorting, revise to align with the English intent. Record all fixes.
- **B. Instruction–Prompt Alignment**: For every instruction in the instruction_id_list, mark as Realized or Not Applicable (e.g., case sensitivity in scripts without upper/lower case). Keep IDs in metadata; flag separately.
- **C. Linguistic and Logical Consistency**: Grammatically correct, culturally neutral, logically coherent. Prohibitions and quantitative phrases must retain their restrictive force exactly.
- **D. Formatting and Style**: Preserve examples, symbols, and URLs. Maintain Markdown. Do not introduce additional formatting.

4. Correction and Flagging Process

For each entry: compare English and translated version; flag non-applicable instruction IDs; correct small inconsistencies directly; record every modification; do not introduce new instructions or stylistic changes.

5. Final Review and Consistency Check

Ensure all translations are accurate and natural; verify script-dependent IDs are consistently flagged; confirm all modifications are logged; export validated dataset with annotation log.

F.2 CL-GSMSym Translation Validation Guidelines (100 Math Templates)

1. Objective

Ensure every translated mathematical template accurately preserves the structure, logic, and functionality of the original English version. Tasks: preserve all template markup; ensure math logic and units match the original; correct formatting issues; flag every fix.

2. Validation Criteria

- **A. Template Markup Preservation**: Every variable token must remain identical to English (`{var}`, `{var,label}`, etc.). Meta-sections (`#init:`, `#conditions:`, `#answer:`) must remain in English and unaltered. Flag: "Restored placeholder token".
- **B. Math and Unit Consistency**: All numeric quantities, operations, and measurement units must match the English template logic. If units were converted (e.g., inches to cm), revert. Flag: "Unit drift fixed".
- **C. Angle Brackets and Final Answer**: Maintain «. . . » spans exactly. Final answer line must end with `#### <number>` or `#### {expression}`.
- **D. Meta Sections**: `#init:`, `#conditions:`, `#answer:` must remain in English, same order, with unchanged indentation and syntax. Flag: "Meta block reverted to EN".

3. Final Review Checklist

Verify: all placeholders {. . . } and spans «. . . » intact; all templates end with correct final answer format; meta sections in English, properly indented; all numbers and units aligned with original logic; all modifications recorded.

G Failure Case Examples for CL-IFEval

Failure Example: `length_constraints:number_words`

Model: Qwen3-8B **Language:** English

Prompt: *Create a riddle about the name Sheldon using only 10 words. Make sure to only use capital letters in your entire response.*

Model Response: SHELDON, A NAME WITH A FUNKY RING.

Instructions:

- `change_case:english_capital` ✔
- `length_constraints:number_words` ✖

Failure: Response contained only 8 words, violating the 10-word constraint.

Failure Example: `change_case:capital_word_frequency`

Model: Qwen3-8B **Language:** French

Prompt: *Rédigez un CV pour Matthias Algiers. Utilisez des mots en majuscules pour mettre en avant les compétences clés, mais assurez-vous que les mots entièrement en majuscules apparaissent moins de 10 fois. Enveloppez toute la réponse entre guillemets.*

Model Response: "MATTHIAS ALGIERS — RÉSUMÉ PROFESSIONNEL — Professionnel dévoué ... GESTION DE PROJET, PLANIFICATION STRATÉGIQUE ..."

Instructions:

- `change_case:capital_word_frequency` ✖
- `startend:quotation` ✔

Failure: More than 10 fully capitalized words, violating the capital word frequency constraint.

H Template Examples for CL-GSMSym

To illustrate the structure and reasoning complexity of items in the CL-GSMSym benchmark, we present representative template-based examples

Template 3 (French) – Probabilistic inference

Question: *Luis lance un dé à quatre faces. Quelle est la probabilité (exprimée en pourcentage) qu'il obtienne un nombre supérieur à 2 plutôt que deux nombres impairs consécutifs ?*
Answer: Il y a 2 nombres supérieurs à 2 sur le dé, donc les chances d'en obtenir un sont de $\frac{2}{4} = 50\%$. La probabilité d'obtenir un nombre impair est de 50%, donc la probabilité d'en obtenir deux d'affilée est $0.5 \times 0.5 = 25\%$. La différence est $50\% - 25\% = 25\%$. **#### 25**

Template 3 (English) – Probabilistic inference

Question: *Faisal is rolling a four-sided die. How much more likely is it (expressed as a percentage) that he rolls a number greater than 1 than that he rolls two even numbers in a row?*
Answer: There are 3 numbers greater than 1 on the die, so the chance is $\frac{3}{4} = 75\%$. The chance of rolling two even numbers in a row is $0.5 \times 0.5 = 25\%$. The difference is $75\% - 25\% = 50\%$. **#### 50**

Template 10 (Spanish) – Algebraic setup

Question: *Cuando Emma observa a su primo, saca una variedad de juguetes para él. La bolsa de bloques tiene 74 bloques. El contenedor de peluches tiene 37 animales. La torre tiene 30 anillos. Emma compró recientemente un tubo de pelotas saltarinas, elevando el total a 215. ¿Cuántas pelotas vinieron en el tubo?*
Answer: $74 + 37 + 30 + T = 215 \Rightarrow T = 215 - 141 = 74$. **#### 74**

Template 10 (English) – Algebraic setup

Question: *When Winnie watches her nephew, she gets out a variety of toys. The bag of building blocks has 72 blocks. The bin of stuffed animals has 47 animals. The tower of stacking rings has 29 rings. Winnie recently bought a tube of bouncy balls, bringing her total to 215. How many bouncy balls came in the tube?*
Answer: $72 + 47 + 29 + T = 215 \Rightarrow T = 215 - 148 = 67$. **#### 67**