Déjà Vu: Multilingual LLM Evaluation through the Lens of Machine Translation Evaluation

Julia Kreutzer[1], Eleftheria Briakou[2], Sweta Agrawal[2], Marzieh Fadaee[1], and Kocmi Tom[3]

[1]Cohere Labs, [2]Google, [3]Cohere

Corresponding authors: Julia Kreutzer (juliakreutzer@cohere.com)

Abstract

Generation capabilities and language coverage of multilingual large language models (mLLMs) are advancing rapidly. However, evaluation practices for generative abilities of mLLMs are still lacking comprehensiveness, scientific rigor, and consistent adoption across research labs, which undermines their potential to meaningfully guide mLLM development. We draw parallels with machine translation (MT) evaluation, a field that faced similar challenges and has, over decades, developed transparent reporting standards and reliable evaluations for multilingual generative models. Through targeted experiments across key stages of the generative evaluation pipeline, we demonstrate how best practices from MT evaluation can deepen the understanding of quality differences between models. Additionally, we identify essential components for robust meta-evaluation of mLLMs, ensuring the evaluation methods themselves are rigorously assessed. We distill these insights into a checklist of actionable recommendations for mLLM research and development.

1 Introduction

Evaluating LLMs in a multilingual context involves testing their capabilities across different languages and tasks, with particular attention to less-studied and lower-resourced non-English languages (Huang et al., 2024). Naturally, it inherits challenges from monolingual LLM evaluation, such as benchmark contamination (Yang et al., 2023; Deng et al., 2024; Dong et al., 2024; Li et al., 2024; Ni et al., 2025), label noise (Vendrow et al., 2025), costs vs coverage trade-offs (Zhang et al., 2024a), standardization, reliability, and diversity issues (McIntosh et al., 2024). These challenges become even more evident when attempting to draw conclusions about *model progress across multiple languages*.

Prior classification benchmarks from cross/multilingual studies that pre-date the decoder-only-LLM era can be re-used to gain performance insights for mLLMs (Hu et al., 2020; Ruder et al., 2021; Liang et al., 2020; Ahuja et al., 2023; Asai et al., 2024). However, many of these benchmarks have reached saturation (Kiela et al., 2021; 2023) and are not separating models sufficiently (Zhang et al., 2024c). They are unreliable predictors of generative abilities of mLLMs (Üstün et al., 2024), as they serve primarily for knowledge testing. Generative abilities are key in real-world applications (Tamkin et al., 2024), and have thus moved into the spotlight of LLM evaluations (Dubois et al., 2023; Chiang et al., 2024; Lin et al., 2024). Multilingual models shine especially in these generative tasks, outperforming monolingual models across the bench (evidence in Appendix D). However, particularly this area of evaluation is still in the early stages.

Current generative evaluation approaches for multilingual models *lack nuances in reporting, reproducibility, standardization, robustness and reliability,* and most notably, *meta-evaluation.* These challenges, albeit new in the mLLM evaluation field, are familiar problems in a sister field, the evaluation of machine translations. In this paper, we thus establish a connection to machine translation (MT) evaluation research, *linking new questions in mLLM evaluation research to known solutions in MT evaluation research.*

MT has had a headstart on navigating these complexities in multilingual generation evaluation.

As one of the core tasks in the NLP field, it has a rich research history of evaluations

with automatic metrics (Papineni et al., 2002; Koehn & Monz, 2006a; Lavie & Agarwal, 2007; Stanojević & Sima'an, 2014; Popović, 2015; Rei et al., 2020) and human judgments (Vilar et al., 2007; Birch & Osborne, 2010; Lopez, 2012; Graham et al., 2013; Freitag et al., 2021; Kocmi et al., 2024b), spurred by venues like the annual Conference on Machine Translation (WMT). The development of evaluation went hand in hand with gradual improvement of model abilities and language coverage: Evaluation metrics that once worked sufficiently for statistical models became ineffective for neural models with superior translation quality (Freitag et al., 2022), or for newly added languages (Bapna et al., 2022). Meta-evaluation (Callison-Burch et al., 2007; 2008; Macháček & Bojar, 2013; Post, 2018; Mathur et al., 2020; Amrhein et al., 2022; Deutsch et al., 2023, inter alia), i.e., the evaluation of evaluations, led to the development of evaluation and transparency standards and built a framework for metric builders.

Elements of this progress have yet to be seen in mLLM evaluation, due to traditionally disjoint research streams, and the rapid speed of mLLM development. To bridge this gap, we first **identify challenges** in generative mLLM evaluation through an assessment of current benchmarks and their adoption in model releases (Section 2). We then highlight **five concrete evaluation principles** that are lacking in mLLM evaluations but established in MT (Section 3). Finally, we establish which **prerequisites are necessary for meta-evaluations** (Section 4).

We distill these findings into an **actionable checklist** for mLLM research (Appendix I),[1] to help steering mLLM development towards more reliable, expressive, and rigorous evaluations.

2   The Status Quo of Multilingual LLM Generation Evaluation

We compile a non-exhaustive list of open multilingual generative benchmarks in Table 1 to survey the mLLM landscape.[2] We summarize trends as follows, and link (▶▶◀) them to proposed strategies from MT evaluation research in Section 3 and meta-evaluation in Section 4.

Multilinguality via translation Most tasks rely on the translation of the original English benchmark for multilingual expansion. Only *XLSum* (Hasan et al., 2021), *Aya human-annotated* (Singh et al., 2024b), *SeaBench* (Zhang et al., 2024b), *MAFAND-MT* (Adelani et al., 2022) are directly curated in the target languages. Automatic prompt translations might not be universally applicable or high-quality (Zhang et al., 2023; Plaza et al., 2024; Agrawal et al., 2024a; Thellmann et al., 2024), which some benchmarks address with post-editing or localization (*e.g. SEA-IFEval* (Ong

---

[1] https://github.com/Cohere-For-AI/multilingual-llm-evaluation-checklist
[2] We exclude classification benchmarks such as MCQA problems, see discussion in Appendix C.

Benchmark	Task	Rank	Size	Judge?	Source	#Langs	Transl.		Benchmark	Task	Rank	Size	Judge?	Source	#Langs	Transl.
FLORES-200	📖	★★★	≈ 1,000	✗	🌐	200	M+		MGSM	▦	★★☆	250	✗	👤	10	H
NTREX-128	📖	☆☆☆	≈ 2,000	✗	🌐	128	H		AfriMGSM	▦	☆☆☆	250	✗	👤	16	H
WMT24++	📖	☆☆☆	≈ 2,000	✗	🌐	55	H		SeaBench	💬	★☆☆	300	✓	🧑‍🔬	3	-
General MT	📖	★★☆	≈ 2,000	✗	🌐	≥ 11	H		Sea-MTBench	💬	★☆☆	58	✓	[🧑‍🔬]	6	H
MAFAND-MT	📖	☆☆☆	1,000	✗	🌐	21	-		MTG	✏	★☆☆	3,000	✓	👤/🌐	5	M+
XLSum	✂	★★⯪	500–11,000	✗	🌐	45	-		OMGEval	✏	☆☆☆	804	✓	[⚙]	5	M
CrossSum-In	✂	☆☆☆	500	✗	[🌐]	29	H		mArenaHard	✏	★☆☆	500	✓	👤	23	M
SEA-IFEval	☰	★☆☆	105	✗	[⚙]	6	H		Dolly translated	✏	★★☆	200	✓	👤	101	M+
MIFEval	☰	☆☆☆	96	✗	[⚙]	10	M		Aya human-ann.	✏	★☆☆	250	✓	👤	7	-
MultiIF	☰	☆☆☆	454–909	✗	[⚙]	7	M		PolyWrite	✏	★☆☆	≈ 155	✓	⚙	240	M
									MultiQ	✏	☆☆☆	200	✓	⚙/👤	137	M

Table 1: Public generative benchmarks for downstream text-based evaluation of mLLMs. They are sourced from the web (🌐), crowds (👤), experts (🧑‍🔬), or machine generated (⚙). Brackets indicate extension of previous benchmarks. Translations of prompts are denoted as H(uman) and M(achine), with M+ indicating human post-edits. We mark LLM judged benchmarks (✓), and rank them by popularity ★ ☆☆☆ in model releases.

& Limkonchotiwat, 2023)). While translation achieves a broad coverage of languages, it limits the cultural representativeness and might propagate Western-centric and Anglo-centric biases (Singh et al., 2024a; Guo et al., 2024) (⏭ Section 3.1).

Small and not so mighty The majority of test sets contain less than 500 prompts per language, with MT benchmarks as outliers with over 1,000 samples. While prompt sourcing is a challenging task, especially with experts, such small sets raise questions of statistical power (⏭ Section 3.2). When included, human evaluations tend to cover even fewer instances (Gehrmann et al., 2023). Most benchmarks provide only a test split, lacking development sets for tuning, which increases the risk of overfitting and diminishes the significance of reported improvements over time (van der Goot, 2021; Ott et al., 2022). Qualitative insights beyond aggregated task metrics are rarely included in evaluation reports (⏭ Section 3.4).

Divergences in benchmark adoption and reporting Only few generative benchmarks are well-established, i.e., multiple labs use them for reporting results in open mLLM releases (Appendix A). Flores-200 (Costa-jussà et al., 2022), MGSM (Shi et al., 2023), and XLSum (Hasan et al., 2021) are the most popularly used benchmarks, as indicated by the rank in table 1. These are closed generative evaluation tasks that have the advantage of having relatively well-defined evaluation paradigms. Open generation tasks like chat and open-ended QA have less standardized evaluations and tend to rely on LLM judges, which introduces more ambiguities. What complicates cross-paper comparisons even when using the same benchmark, is the lack of transparency and standardization in evaluation reporting. This goes from the choice of automatic metric (or LLM judge), over prompting conditions and formulations (⏭ Section 3.5), to the selection and aggregation across languages for comparison (⏭ Section 3.3). For instance, performance on Flores-200 is measured with different metrics (spBLEU (Goyal et al., 2022), ChrF (Popović, 2015), COMET-22 (Rei et al., 2022)), and for MGSM model reports vary the number of shots, or even define new criteria (Barcelona Supercomputing Center, 2024). When metrics are established for a generative task in English, they might not transfer equally well to all evaluated languages (Gehrmann et al., 2023). Sometimes, it is not even stated which languages of a benchmark are chosen for evaluation, and rarely do they cover all of the supported languages of a model (see table 5).

Generative models are becoming the metric The emergence of new generative tasks, such as chat and open-ended generation, do not come with decades of task-specific metrics research. Thus, LLM judges are used to express preferences through pairwise comparisons of model outputs, both in training (Lee et al., 2024), and in evaluation (Zheng et al., 2023a; Gu et al., 2025). However, this in itself is a generative evaluation task measuring how good mLLMs are at judging multilingual generations (Gureja et al., 2024; Doddapaneni et al., 2024) – raising questions about biases in used judges (Ye et al., 2024) (▶▶⊩ Section 4.1).

Evaluation with a rapidly moving target All of today's leading mLLMs and most of the generative benchmarks are less than a year old. Benchmarks quickly "expire", due to score saturation as a consequence of overfitting, evolved capacities or contamination (Ahuja et al., 2024), or they simply lose relevance for LLM user or broader research needs (Zheng et al., 2023b; Tamkin et al., 2024). Many new model releases include testing on newly introduced benchmarks to highlight new strengths, but these benchmarks are rarely adopted by consequent releases of other labs. Open leaderboards attempt to close this gap, tracking progress on a selection of tasks and languages across models (see Appendix B), but they are also prone to expiry, might lack utility (Ethayarajh & Jurafsky, 2020) and heavily rely on aggregations for interpretation, which requires particular care for multilingual models (Hulagadri et al., 2025) (▶▶⊩ Section 3.3). This calls for a larger arc of evaluation, namely the evaluation of evaluations themselves, including automatic (▶▶⊩ Section 4.1) and human evaluation (▶▶⊩ Section 4.2).

3 Adopting Evaluation Practices from MT Evaluation

Based on the challenges outlined in Section 2, we identify five central questions in the mLLM evaluation pipeline and relate them to insights and practices from MT. The guiding question is: *What knowledge would we gain about mLLMs, if we supplemented their evaluations with MT-style evaluation techniques?*

3.1 Where Does the Data Come From? Treating Synthetic Data with Care

Machine translated datasets are commonly used in mLLM training (Dang et al., 2024a) and evaluation (Lai et al., 2023), with the intention to reduce data scarcity across languages (Muennighoff et al., 2023; Holmström & Doostmohammadi, 2023; Üstün et al., 2024). However, synthetic, model-generated data is prone to systematic biases (Ahn et al., 2022; Lukasik et al., 2022; Shimabucoro et al., 2024). In particular, machine-translated prompts may contain translation artifacts affecting evaluation outcomes (Chen et al., 2024; Guo et al., 2024; Agrawal et al., 2024a). In MT research, studies have shown that grammar, structure, or word choice of the source text can systematically influence human and machine translations – a phenomenon known as *translationese* (Gellerstam, 1986; Laviosa, 2011). In evaluations, the presence of translationese in the sources (i.e., because the source text was itself translated from the target language) has been found to decrease the difficulty of the task (Zhang & Toral, 2019), even leading to false claims of human parity (Hassan et al., 2018; Toral et al., 2018; Graham et al., 2020). As a result, maintaining source authenticity has become a critical principle in the creation of test sets (Barrault et al., 2019).

⚗ To illustrate the effects of prompt translation in multilingual generative evaluation, we conduct an experiment using 250 Aya human annotated prompts for Arabic, Chinese, English, Portuguese and Turkish. We round-trip-translate them automatically via a pivot language to create a comparison
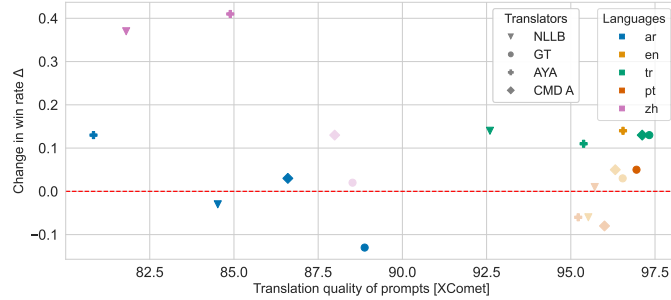
Figure 1: The effect of prompt translation quality on win rates differences between AYA EXPANSE 8B vs GEMMA2 9B: Win rate Δs mostly increase compared to the ones under original prompts ($y = 0$). Transparent points reflect non-significant win rate Δs (at 95% CI).

Average:	XCOMET ↑	Change in Win Rate Δ ↓
NLLB	90.03	0.06
GT	93.65	0.02
AYA	90.57	0.14
CMD A	92.80	0.05

Table 2: Average roundtrip translation quality of translation models tested on *Aya human annotated* prompts across five languages (ar, en, pt, tr, zh), and the change in win rate Δ when comparing AYA EXPANSE 8B and GEMMA 2 9B. Ideally, translation should not affect the win rate.

between original prompts and translated prompts (Chen et al., 2024). For translation, we use a diverse range of translators: Google Translate (GT), and NLLB-200-3.3B (NLLB Team, 2022), AYA EXPANSE 32B (Dang et al., 2024b) and COMMAND A (Cohere et al., 2025). We measure how GPT-4o-as-a-judge win rates change when comparing mLLM generations for original prompts to those for translated prompts. We focus on a comparison of GEMMA2 9B (Gemma Team, 2024) and AYA EXPANSE 8B (Dang et al., 2024b), selected from a wider range of experiments in Appendix G.

🔍 We find that **win-rate differences in pairwise evaluations are affected by translation, with magnitudes that vary across languages and translation models** (fig. 1), depending on translation quality (table 2). We can see that the majority of translations tilt the scale in favor of AYA EXPANSE 8B, increasing the win rate delta over GEMMA2 9B from 0.18 to 0.32 on average across languages (especially for Chinese and Turkish). Why hypothesize that AYA EXPANSE 8B is more robust to translation artifacts in the prompts due to exposure during training (Artetxe et al., 2020a). We also note that mLLMs used for translation (AYA EXPANSE 32B, COMMAND A) appear to have proportionally larger downstream effects. Overall, this simulation demonstrates that win rates computed on translated evaluation prompts might systematically favor particular models that are more robust to translation artifacts, leading to inflated win rates.

💡 **Recommendation 1:** For evaluation, prefer target-language original prompts over translated alternatives (*silver standards*, coined by Holtermann et al. (2024)). If translations are unavoidable, ensure that their quality is optimized without assuming off-the-shelf adequacy for any task (*e.g.* choosing best MT, adding post-edits, localization). Measure and document translation quality on a representative subset for each task.

3.2   What do Score Differences Mean? Measuring Significance, Power & Effect Size

Although platforms like Chatbot Arena report confidence intervals using bootstrapping, significance testing is not yet a standard part of the LLM development pipeline (Vaugrante et al., 2024; Ackerman et al., 2025). To address this, Miller (2024) proposed best LLM evaluation practices, emphasizing the importance of reporting sample size, confidence intervals, and standard errors, particularly for clustered and paired tests. In MT research, such reporting and significance tests have a long
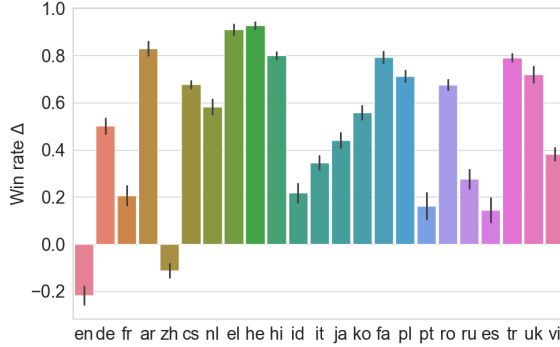
Figure 2: Win rates deltas for AYA EXPANSE 8B vs. QWEN2.5 7B on mArenaHard prompts (23 languages), with GPT-4o as a judge. Error bars denote std. dev. across 5 samples for each prompt.
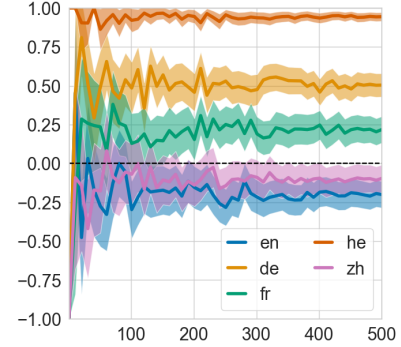

Figure 3: Win rate deltas in relation to sample size. Differences are significant when the 95% confidence interval (shaded) lies above/below zero.

history (Koehn, 2004; Riezler & Maxwell, 2005; Graham et al., 2014b; 2020) and have found moderate adoption (Marie et al., 2021), also enabled by ease of use in tools like `sacrebleu` (Post, 2018) or `comet-compare` (Rei et al., 2020). Statistical power analyses can further help determine the sample size required for reliable evaluations (Card et al., 2020), *e.g.* for human preference evaluation. In MT, for instance, statistical power is usually sufficient (> 0.8) to rank even close models with $\approx$1.5K sources (Graham et al., 2020), yet smaller sample sizes may be insufficient (Wei et al., 2022). Based on test sizes of the benchmarks reviewed in Section 2, it is likely that especially under metrics with high variance such as pairwise LLM judgments, many mLLM evaluations might be underpowered.

There are some pitfalls to be aware of: first, different metrics for the same task may have varying sensitivity (Riezler & Maxwell, 2005), which could lead to differences in one metric being significant but insignificant in another. Second, the more statistical tests are done, the more likely false positives will be encountered. This becomes particularly relevant for testing multiple models on multiple languages and multiple benchmarks. Correction (Zerva et al., 2022; Ulmer et al., 2022) can prevent this inflation, *e.g.* by increasing the threshold of significance for individual tests (Bonferroni correction), implemented at WMT.

It is important to recognize that statistical significance does not necessarily imply that a difference is noticeable or meaningful to humans (Mathur et al., 2020; Agrawal et al., 2024b). With sufficiently large sample sizes, even very small differences in metric scores can become statistically significant, despite being too subtle to notice in practice. This issue is specifically known in MT, where the magnitude of the effect size plays a crucial role in determining whether system improvements are genuinely meaningful (Kocmi et al., 2024c).

⚗ To illustrate the benefits of statistical significance testing, we inspect pairwise comparisons of AYA EXPANSE 8B and QWEN2.5 7B INSTRUCT on the 500 prompts of the mArenaHard benchmark for 23 languages, with GPT4-o as a judge. This comparison was previously reported (Dang et al., 2024b), but without considering significance tests or sample sizes. We compare win rates across languages and compute significance based on 95% confidence intervals, as recommended by Miller (2024). Experimental details are in Appendix E.

🔍 AYA EXPANSE 8B wins with an average win rate delta of 0.49 across languages and five runs.

Models	All			High			Medium			Low		
	Avg	GSM8k	MMLU	Avg	GSM8k	MMLU	Avg	GSM8k	MMLU	Avg	GSM8k	MMLU
Qwen2-7B	1	1	2	1	1	2	1	1	3	1	1	4
Mistral-Nemo-Base-12.2B_2407	3	3	3	3	3	3	2	2	2	2	2	1
Mixtral-8x7B-v0.1	2	2	1	2	2	1	3	3	1	4	4	3
Gemma-7B	5	5	7	5	5	5	4	5	4	3	3	5
Mistral-NeMo-Minitron-8B-Base	4	4	4	4	4	4	5	4	8	5	5	8

Table 3: Effect of different aggregation strategies on model ranking of top-5 pretrained systems as generated by the European Leaderboard on GSM8k and MMLU datasets.

Individual win rate differences vary widely between languages, from -0.21 (loss) for English to 0.93 for Hebrew (Figure 2). Moreover, **the significance of the wins is dependent on sample sizes and languages**: Figure 3 illustrates how these win rate differences behave under different sample sizes. A few Hebrew samples already reveal significant win rate differences, whereas even 300 Chinese samples are insufficient. This analysis highlights what our current setup cannot reliably determine, namely, whether the models qualitatively differ in Chinese text generation capabilities. For smaller test sets, as in most generative benchmarks Section 2, such analysis is essential to avoid overconfidence in low-power evaluations.

♀ **Recommendation 2:** Test the statistical significance of evaluation results rather than relying on metric differences alone, following task-specific recommendations (Dror et al., 2018; Miller, 2024; Ackerman et al., 2025). Estimate statistical power, particularly when working with small sample sizes. Additionally, consider the magnitude of the effect size to determine whether observed differences are also meaningful in practice.

3.3 What Gets Lost In Averages? Aggregating Responsibly

With mLLMs, we are modeling multiple languages and tasks at once. How we aggregate results thus naturally informs the interpretation of model comparisons. The go-to approach is to report uniformly weighted averages across languages and tasks. This is not necessarily a fair evaluation – due to differences in training distributions – nor is it expressive enough – as outliers (*e.g.*, by unseen languages) can disproportionately affect system rankings (Hulagadri et al., 2025). Languages and tasks also differ in their expressive power, as seen in Section 3.2.

In multilingual MT, several aggregation formats were explored beyond reporting plain averages across languages. For instance, grouping by language resourcedness, *e.g.* to study language-specific routing (Zhang et al., 2021); by directionality (Zhang et al., 2020); by unseen/seen languages (Aharoni et al., 2019) to isolate zero-shot generalization. Additionally, WMT offers a constrained track to isolate model improvements from data gains.

⚗ Table 3 shows the ranking of the top 5 systems obtained using the European Leaderboard under different configurations: a) by language and b) by task. We categorize languages based on number of speakers into high ($> 50M+$; en, es, pt, de, fr, it, pl), medium (< 50 and $> 10M$; nl, el, hu, sv, cz, ro) and low ($< 10M$; dk, fi, sk, sl, bg, lt, lv, et) resource.[3]
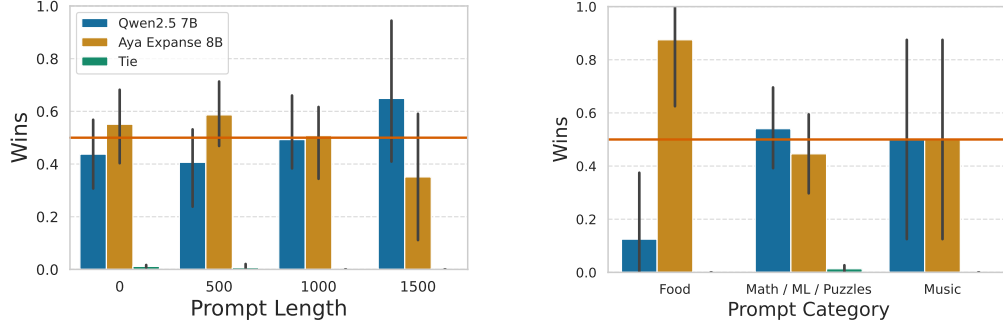

Figure 4: Win-rates for AYA EXPANSE 8B vs. QWEN2.5 7B on mArenaHard prompts bucketed by a) prompt length (left) and b) prompt categories (right).

Q Based on average scores, we would conclude that MIXTRAL-8X7B-V0.1 is the second best system after QWEN2-7B, whereas when looking at task-specific aggregates, we find it consistently outperforms QWEN2-7B on MMLU. For medium and low-resource languages, however, its performance for GSM8k drops, leaving the second rank to others. This shows, that **system rankings can shift based on task and language focus**. Optimal model selection for a specific task and language group can thereby deviate from the average best system.

♀ Recommendation 3: When comparing models across multiple languages, consider differences in language support and aggregate results according to languages being seen by the multilingual models in question. Report task and language-specific scores as a supplement to averages. When you discuss averages, take language coverage into account.

3.4 Where Do Models Differ? Conducting Richer Analyses

Aggregate benchmark metrics do not provide insights into what differentiates the outputs of two models—yet identifying these distinctions is often the first step in human preference evaluation. In MT, specialized tools were developed to facilitate pairwise comparisons on specific examples, such as `MT Compare Eval` (Klejch et al., 2015) and `compare-mt` (Neubig & Hu, 2018). In parallel, there has been a steady effort to create challenge sets and test suites designed to probe particular capabilities and phenomena of MT (Stanovsky et al., 2019; Bawden & Sagot, 2023; Manakhimova et al., 2024). In contrast, LLM evaluations typically rely on user preferences in an arena setting or automatic judges with limited explainability. Before investing in human evaluation, automatic metrics can already offer insights into quality differences between model outputs across languages. Auxiliary metrics such as diversity scores, length statistics, bias detectors (*e.g.*, toxicity), language confusion statistics, and edit distance can highlight key trends. While these do not fully explain model differences as humans might, they can reveal biases, such as verbosity, that could influence both human and automatic pairwise evaluations. As Gehrmann et al. (2023) put it, there is a "systemic difference between selecting the best model and characterizing how good this model really is".

⚗ We illustrate this by comparing the win rates of AYA EXPANSE 8B and QWEN2.5 7B INSTRUCT on a subset of languages (en, de, fr, zh) from the mArenaHard benchmark bucketed by a) the prompt length and b) manually annotated prompt category in Figure 4.
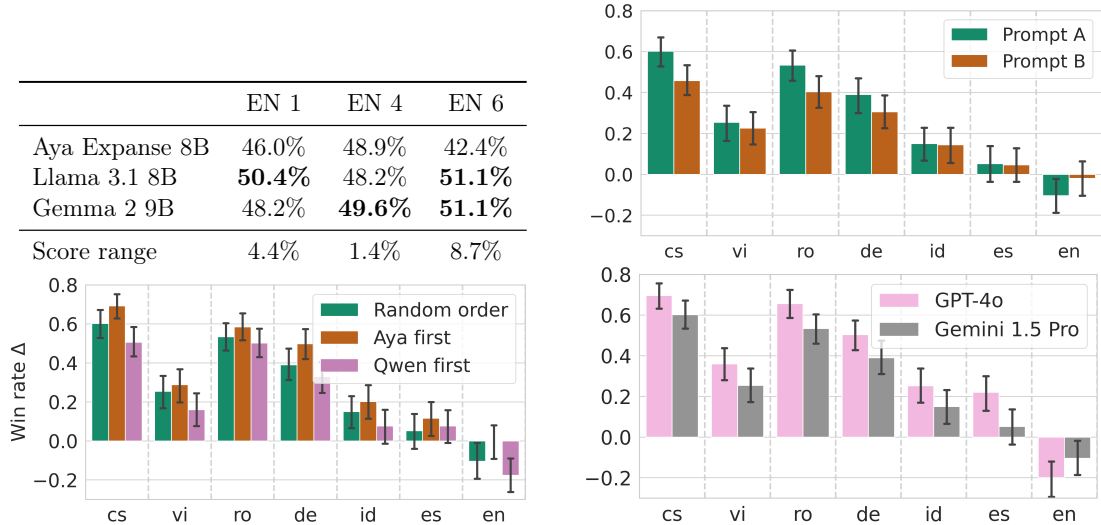
	EN 1	EN 4	EN 6
Aya Expanse 8B	46.0%	48.9%	42.4%
Llama 3.1 8B	**50.4%**	48.2%	**51.1%**
Gemma 2 9B	48.2%	**49.6%**	**51.1%**
Score range	4.4%	1.4%	8.7%

Figure 9: Accuracy on German MCQA (Include 44) with three instruction variants (top-left). Win-rates of Aya Expanse 8B vs. Qwen 2.5 7B on mArenaHard prompts showing prompt (top-right), positional (bottom-left), and judge bias (bottom-right).

Q The first plot shows a clear trend: **QWEN2.5 7B INSTRUCT tends to win on longer prompts, while AYA EXPANSE 8B performs better on shorter prompts**, suggesting that QWEN2.5 7B INSTRUCT can handle detailed and long queries better. On the other hand, from the second plot, AYA EXPANSE 8B emerges victorious in all categories except for "Math / ML / Puzzles" problems, where QWEN2.5 7B INSTRUCT has a clear advantage. These results provide valuable insights: while the average win rates in Figure 2 suggest a general preference for AYA EXPANSE 8B, they obscure QWEN2.5 7B INSTRUCT's clear advantage on specific prompt types. Such findings can guide targeted test set design, inform human evaluation sampling, and steer future model development.

♥ Recommendation 4: Complement automatic metric analyses with qualitative error analysis to better understand systematic patterns. Use visualization and systematic category breakdowns to contextualize metric results, ensuring that observed differences align with meaningful distinctions rather than incidental artifacts.

3.5 What Do We Need to Share? Advancing Reproducibility Through Transparency

Reproducing evaluation results in the LLM era has become increasingly challenging, if not impossible (Vaugrante et al., 2024). Not only are many evaluations stochastic, but they are also dependent on configurations that are rarely fully disclosed, such as preambles or system prompts, task formatting, decoding strategies, temperature, or answer parsing. A similar challenge arose in MT, where even a straightforward metric like BLEU was implemented differently across frameworks, leading to discrepancies in reported scores. The introduction of SacreBLEU (Post, 2018) marked a turning point by standardizing the evaluation pipeline into a single toolkit, with each evaluation assigned a unique signature containing all relevant parameters, ensuring comparability across papers. Efforts like `simple-evals`[4] and the LM Evaluation Harness (Gao et al., 2024) aim to standardize
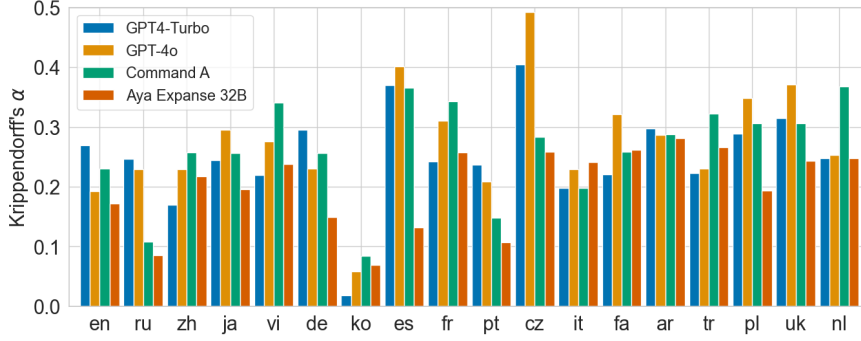
---

[4] https://github.com/openai/simple-evals

Figure 10: LLM-as-a-judge agreement with humans on arena.

	α
GPT-4-Turbo	0.25
GPT-4o	**0.28**
Command A	0.26
Aya Expanse 32B	0.20

	# Best
GPT-4-Turbo	5
GPT-4o	**6**
Command A	**6**
Aya Expanse 32B	1

Table 4: Summary across languages.

task formulations and output parsing for LLM evaluations. However, full transparency requires *open evaluation releases* that contain publicly available code with exact versioning (*e.g.*, commit hashes), full release of all prompts (including instruction text, exact wording, punctuation, and formatting), and disclosure of task formulations in each language.[5] For example, Briakou et al. (2024) demonstrate that minor variations in prompt wording for translation tasks can lead to drastically different outcomes, such as models refusing to translate or producing overly verbose responses. Such findings are enabled by the practice of releasing model outputs, championed in the annual WMT shared task competitions (Koehn & Monz, 2006b), and has kindled metrics and meta-evaluation research by allowing retroactive comparisons and enabling longitudinal studies (Graham et al., 2014a).

⚗ We illustrate configuration's impact on accuracy results for German MCQA (INCLUDE 44 (Romanou et al., 2024b)) (3 prompts) and mArenaHard LLM-as-a-judge win rates varying a) the prompt, b) the compared systems' order, and c) the judge (GPT-4o vs Gemini 1.5Pro).

🔍 Figure 9 shows that system accuracy on MMLU-like evaluations changes significantly with different instruction wordings, undermining the robustness of benchmarking (Alzahrani et al., 2024). The use of LLMs as judges further complicates reproducibility. Variability in model choice, decoding strategies (Appendix E.1), various biases (Ye et al., 2024; Shimabucoro et al., 2024), and prompt phrasing adds layers of complexity. Figure 9 illustrates how evaluations can be manipulated through positional biases (system presentation order) and prompt formulation differences, yielding significantly divergent outcomes. Finally, LLM version obsolescence prevents reliable comparisons of results across papers and over time. Evaluations are often non-transitive (Xu et al., 2025), making optimization a moving target.

💡 **Recommendation 5:** Use standardized pipelines, publish the exact prompt wording, and release the evaluation code, model outputs, and evaluation scores with versioning.

---

[5]As an example, we release the pairwise evaluation artifacts from this paper: `https://huggingface.co/datasets/CohereForAI/deja-vu-pairwise-evals`.

4 Evaluating mLLM Evaluation: Towards Meta Evaluation

The field of MT has been consistently involved in meta-evaluation of machine translation evaluation methods for the last twenty years (Callison-Burch et al., 2007). This process, whether implicit or explicit, has driven progress in MT systems by redefining which metrics best correlate with human judgments throughout various milestones of model improvements. Given this progress in MT evaluation, a natural question arises: why has not similar progress been observed in multilingual evaluation? To better understand this disparity, we revisit the prerequisites of meta-evaluation. Meta-evaluation fundamentally requires three components: system outputs, human judgments (of those outputs), and automatic evaluations of those same outputs. In the following sections, we identify the missing components in the multilingual setting and outline the steps necessary to overcome these challenges.

4.1 Need for (More) Metrics: Beyond One-size-fits All

Lesson from MT While LLMs-as-a-judge offers the convenience of using a single model for multilingual assessments, MT meta-evaluation shows there is no universally best metric (Marie et al., 2021; Anugraha et al., 2024). Central to this issue was the widespread use of pretrained language models as backbones for developing learned metrics (Lo, 2020), with varying degrees of language representation coverage. Although learned metrics often outperform string-based ones – like BLEU (Papineni et al., 2002) – the choice between a string-based and a learned metric is heavily language and domain dependent. This has led to the informal convention of employing different types of metrics for targeted evaluations and reporting multiple metrics.

Application to multilingual evaluation Zheng et al. (2023b) used human evaluations from Chatbot Arena (Chiang et al., 2024) to study the reliability of LLM judges. We extend their predominantly English analysis to non-English "battles" (pairwise comparisons) from the released sample of battles (`lmarena-ai/arena-human-preference-100k`), focusing on the 18 languages with >200 prompts. We score a subset of 200 generation pairs for each language with two open and two closed mLLMs as judges. We measure agreement with human preferences with Krippendorff's α (interval measurements), shown in Figure 10. Overall agreement, and the choice of the best judge varies across languages. Just like in MT, the optimal choice of an LLM judge (i.e., metric) for multilingual evaluation is language dependent (Table 4).

4.2 Need for Nuanced Human Evaluation: Towards Richer Assessments

Lesson from MT Human evaluation of translation quality is a multifaceted challenge, rooted in the fundamental questions of *what* to measure and *how* to elicit accurate, consistent human assessments. These questions have long been a central focus in machine translation. Below, we highlight key areas and insights from this extensive body of work.

Detailed work has explored different **evaluation protocols** (Vilar et al., 2007; Graham et al., 2013; 2014a) and **quality dimensions**, including fluency versus adequacy (Koehn & Monz, 2006a; Bojar et al., 2016). The trend then went from monolithic assessment scores of pair-wise assessments towards more **fine-grained protocols**, *e.g.* highlighting and annotating errors using established taxonomies, such as the Multidimensional Quality Metrics (MQM) framework (Burchardt, 2013; Freitag et al., 2021). These taxonomies are getting refined to balance cognitive load and annotation

effectiveness (Ge et al., 2024), and adapted for non-professional annotators (Graham et al., 2015; Castilho et al., 2017; Wang et al., 2024). Annotation efforts have also expanded to target specific use cases, such as **critical error** detection (Specia et al., 2021; Zerva et al., 2022), detection of errors grounded in high-stake scenarios (Mehandru et al., 2023), and to contextualize evaluations within **user-centric frameworks** (Briakou et al., 2023; Savoldi et al., 2025).

Application to multilingual evaluation Chatbot Arena, where users compare two models in a chat and choose a winner, are the primary source of public human mLLM evaluations. The maintainers periodically releases data,[6] forming the largest public collections of multilingual human preferences. In the data from 2024, 27–43% of battles per language end in a tie (analysis in Appendix H), suggesting that human evaluation in the pairwise arena format lacks sensitivity to fine-grained differences or inherently includes significant uncertainty. The scarcity of publicly available multilingual preference data limits research in this area. Emerging efforts such as MM-EVAL (Son et al., 2024b) introduce multilingual meta-evaluation benchmarks, highlighting that LLMs-as-a-judge often lack fairness and consistency across languages.

4.3   Need for meta-evaluation research: Closing the Loop on Evaluation

Lesson from MT Meta-evaluation research has been formally conducted within the WMT Metrics shared task since 2007 (Callison-Burch et al., 2007). This research aims to improve MT evaluation by identifying best performing metrics, addressing weaknesses in correlation-based metrics, *e.g.*, proper handling of ties (Deutsch et al., 2023), meta-evaluation techniques (Kocmi et al., 2021; Thompson et al., 2024), and challenges of conducting reliable human evaluations, such as ensuring replicable human evaluations (Riley et al., 2024) and studying inter-annotator agreement (Popović, 2021; Popović & Belz, 2022).

Application to multilingual evaluation Evaluation solely based on correlation with human pairwise preferences on prompt-level comparisons is not sustainable, as human agreement decreases with shrinking quality differences between the contrasted systems (Zheng et al., 2023b) – a finding that was also observed in mLLM development (Üstün et al., 2024). We get a glimpse of this loss of signal in Chatbot arena battles: from 2023 to 2024, the ratio of ties has grown significantly, from an average of 29% to 40% across the six most dominant languages (Appendix H). For long-term progress in mLLM modeling and evaluation, we need to iterate the meta-evaluation loop. As a first step, we need to answer the questions which differences between models matter to humans and how to capture these. Then, we can adapt automatic metrics accordingly. Finally, we can measure modeling progress automatically and reliably, which results in models with enhanced qualities, bringing us back to the first step.

5   Conclusion

MT has long grappled with the complexities of multilingual generative evaluations, from constructing datasets to benchmarking evaluation metrics. We demonstrated that established practices from MT can enhance the understanding and reliability of comparisons among mLLMs, and outlined which elements are necessary for establishing meta-evaluations. Our recommendations are distilled into a practical checklist (Appendix I).

---

[6]https://github.com/lm-sys/FastChat/blob/main/docs/dataset_release.md

Limitations Our experiments are focused on open-ended generative tasks, which could be still considered more ambiguous than typical MT tasks, since MT evaluation criteria are more defined with respect to a reference generation. Therefore, not all approaches for MT evaluation might transfer equally well. However, with more advances in quality, mLLM evaluations are trending towards more tightly defined benchmarks that require in-depth expert knowledge (e.g. coding, math), which brings these tasks closer to the conditions of MT evaluations. Furthermore, recommendations and best practices from other sub-fields of NLP that are now sharing multilingual benchmarks should also be considered, see for example the recommendations by Gehrmann et al. (2023) for evaluating text generation, or those by Iskender et al. (2021) for human evaluation of text summarization.

Outlook Since mLLMs are now also competing with non-LLM MT models (Kocmi et al., 2023; 2024a; Zhu et al., 2024b), and MT benchmarks have become established evaluation tasks for mLLMs (Zhu et al., 2024a), the sharing of knowledge and insights across both disciplines becomes even more important to drive meaningful progress. Our checklist with practical recommendations for mLLM evaluations is the first step towards the aim of bringing research communities closer.

Acknowledgments

References

Hugo Abonizio, Thales Sales Almeida, Thiago Laitz, Roseval Malaquias Junior, Giovana Kerche Bonás, Rodrigo Nogueira, and Ramon Pires. Sabiá-3 technical report, 2025. URL https://arxiv.org/abs/2410.12049.

Samuel Ackerman, Eitan Farchi, Orna Raz, and Assaf Toledo. Statistical multi-metric evaluation and visualization of llm system predictive performance, 2025. URL https://arxiv.org/abs/2501.18243.

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3053–3070, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.223. URL https://aclanthology.org/2022.naacl-main.223.

Ashish Agrawal, Barah Fazili, and Preethi Jyothi. Translation errors significantly impact low-resource languages in cross-lingual learning. In Yvette Graham and Matthew Purver (eds.),

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 319–329, St. Julian's, Malta, March 2024a. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-short.28.

Sweta Agrawal, António Farinhas, Ricardo Rei, and Andre Martins. Can automatic metrics assess high-quality translations? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14491–14502, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.802. URL https://aclanthology.org/2024.emnlp-main.802.

Roee Aharoni, Melvin Johnson, and Orhan Firat. Massively multilingual neural machine translation. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3874–3884, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1388. URL https://aclanthology.org/N19-1388.

Jaimeen Ahn, Hwaran Lee, Jinhwa Kim, and Alice Oh. Why knowledge distillation amplifies gender bias and how to mitigate from the perspective of DistilBERT. In Christian Hardmeier, Christine Basta, Marta R. Costa-jussà, Gabriel Stanovsky, and Hila Gonen (eds.), *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pp. 266–272, Seattle, Washington, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.gebnlp-1.27. URL https://aclanthology.org/2022.gebnlp-1.27.

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. MEGA: Multilingual evaluation of generative AI. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4232–4267, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.258. URL https://aclanthology.org/2023.emnlp-main.258.

Sanchit Ahuja, Varun Gumma, and Sunayana Sitaram. Contamination report for multilingual benchmarks, 2024. URL https://arxiv.org/abs/2410.16186.

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13787–13805, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.744. URL https://aclanthology.org/2024.acl-long.744.

Chantal Amrhein, Nikita Moghe, and Liane Guillou. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri (eds.), *Proceedings of the Seventh Conference on*

Machine Translation (WMT), pp. 479–513, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.44.

David Anugraha, Garry Kuwanto, Lucky Susanto, Derry Tanti Wijaya, and Genta Winata. MetaMetrics-MT: Tuning meta-metrics for machine translation via human preference calibration. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, pp. 459–469, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.32. URL https://aclanthology.org/2024.wmt-1.32.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Translation artifacts in cross-lingual transfer learning. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7674–7684, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.618. URL https://aclanthology.org/2020.emnlp-main.618.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4623–4637, Online, July 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.421. URL https://aclanthology.org/2020.acl-main.421.

Akari Asai, Sneha Kudugunta, Xinyan Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. BUFFET: Benchmarking large language models for few-shot cross-lingual transfer. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1771–1800, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.100. URL https://aclanthology.org/2024.naacl-long.100.

Nishant Balepur, Rachel Rudinger, and Jordan Lee Boyd-Graber. Which of these best describes multiple choice evaluation with llms? a) forced b) flawed c) fixable d) all of the above, 2025. URL https://arxiv.org/abs/2502.14127.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. Building machine translation systems for the next thousand languages, 2022. URL https://arxiv.org/abs/2205.03983.

Barcelona Supercomputing Center. Salamandra. https://huggingface.co/BSC-LT/salamandra-7b-instruct, 2024.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2019 conference on machine translation (WMT19). In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi,

and Karin Verspoor (eds.), *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pp. 1–61, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5301. URL https://aclanthology.org/W19-5301.

Rachel Bawden and Benoît Sagot. RoCS-MT: Robustness challenge set for machine translation. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 198–216, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.21. URL https://aclanthology.org/2023.wmt-1.21.

Alexandra Birch and Miles Osborne. LRscore for evaluating lexical and reordering quality in MT. In Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, and Omar Zaidan (eds.), *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pp. 327–332, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://aclanthology.org/W10-1749.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. Findings of the 2016 conference on machine translation. In Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann, Liane Guillou, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Aurélie Névéol, Mariana Neves, Pavel Pecina, Martin Popel, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Lucia Specia, Karin Verspoor, Jörg Tiedemann, and Marco Turchi (eds.), *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pp. 131–198, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2301. URL https://aclanthology.org/W16-2301.

Eleftheria Briakou, Navita Goyal, and Marine Carpuat. Explaining with contrastive phrasal highlighting: A case study in assisting humans to detect translation differences. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11220–11237, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.690. URL https://aclanthology.org/2023.emnlp-main.690.

Eleftheria Briakou, Zhongtao Liu, Colin Cherry, and Markus Freitag. On the implications of verbose llm outputs: A case study in translation evaluation. *ArXiv*, abs/2410.00863, 2024. URL https://api.semanticscholar.org/CorpusID:273023109.

Aljoscha Burchardt. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK, November 28-29 2013. Aslib. URL https://aclanthology.org/2013.tc-1.6.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. (meta-) evaluation of machine translation. In Chris Callison-Burch, Philipp Koehn, Cameron Shaw Fordyce, and Christof Monz (eds.), *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 136–158, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/W07-0718.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. Further meta-evaluation of machine translation. In Chris Callison-Burch, Philipp Koehn, Christof Monz,

Josh Schroeder, and Cameron Shaw Fordyce (eds.), *Proceedings of the Third Workshop on Statistical Machine Translation*, pp. 70–106, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL https://aclanthology.org/W08-0309.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9263–9274, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.745. URL https://aclanthology.org/2020.emnlp-main.745.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Panayota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Valerio Miceli-Barone, and Maria Gialama. A comparative quality evaluation of PBSMT and NMT using professional translators. In Sadao Kurohashi and Pascale Fung (eds.), *Proceedings of Machine Translation Summit XVI: Research Track*, pp. 116–131, Nagoya Japan, September 18 – September 22 2017. URL https://aclanthology.org/2017.mtsummit-papers.10.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. When is multilinguality a curse? language modeling for 250 high- and low-resource languages, 2023. URL https://arxiv.org/abs/2311.09205.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. Goldfish: Monolingual language models for 350 languages, 2024. URL https://arxiv.org/abs/2408.10441.

Iaroslav Chelombitko and Aleksey Komissarov. Specialized monolingual BPE tokenizers for Uralic languages representation in large language models. In Mika Hämäläinen, Flammie Pirinen, Melany Macias, and Mario Crespo Avila (eds.), *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*, pp. 89–95, Helsinki, Finland, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.iwclul-1.11.

Pinzhen Chen, Simon Yu, Zhicheng Guo, and Barry Haddow. Is it good data for multilingual instruction tuning or just bad multilingual evaluation for large language models? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 9706–9726, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.542. URL https://aclanthology.org/2024.emnlp-main.542.

Yiran Chen, Zhenqiao Song, Xianze Wu, Danqing Wang, Jingjing Xu, Jiaze Chen, Hao Zhou, and Lei Li. MTG: A benchmark suite for multilingual text generation. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 2508–2527, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.192. URL https://aclanthology.org/2022.findings-naacl.192.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024. URL https://arxiv.org/abs/2403.04132.

Team Cohere, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, Zahara Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, Neeral Beladia, Walter Beller-Morales, Alexandre Bérard, Andrew Berneshawi, Anna Bialas, Phil Blunsom, Matt Bobkin, Adi Bongale, Sam Braun, Maxime Brunet, Samuel Cahyawijaya, David Cairuz, Jon Ander Campos, Cassie Cao, Kris Cao, Roman Castagné, Julián Cendrero, Leila Chan Currie, Yash Chandak, Diane Chang, Giannis Chatziveroglou, Hongyu Chen, Claire Cheng, Alexis Chevalier, Justin T. Chiu, Eugene Cho, Eugene Choi, Eujeong Choi, Tim Chung, Volkan Cirik, Ana Cismaru, Pierre Clavier, Henry Conklin, Lucas Crawhall-Stein, Devon Crouse, Andres Felipe Cruz-Salinas, Ben Cyrus, Daniel D'souza, Hugo Dalla-Torre, John Dang, William Darling, Omar Darwiche Domingues, Saurabh Dash, Antoine Debugne, Théo Dehaze, Shaan Desai, Joan Devassy, Rishit Dholakia, Kyle Duffy, Ali Edalati, Ace Eldeib, Abdullah Elkady, Sarah Elsharkawy, Irem Ergün, Beyza Ermis, Marzieh Fadaee, Boyu Fan, Lucas Fayoux, Yannis Flet-Berliac, Nick Frosst, Matthias Gallé, Wojciech Galuba, Utsav Garg, Matthieu Geist, Mohammad Gheshlaghi Azar, Seraphina Goldfarb-Tarrant, Tomas Goldsack, Aidan Gomez, Victor Machado Gonzaga, Nithya Govindarajan, Manoj Govindassamy, Nathan Grinsztajn, Nikolas Gritsch, Patrick Gu, Shangmin Guo, Kilian Haefeli, Rod Hajjar, Tim Hawes, Jingyi He, Sebastian Hofstätter, Sungjin Hong, Sara Hooker, Tom Hosking, Stephanie Howe, Eric Hu, Renjie Huang, Hemant Jain, Ritika Jain, Nick Jakobi, Madeline Jenkins, JJ Jordan, Dhruti Joshi, Jason Jung, Trushant Kalyanpur, Siddhartha Rao Kamalakara, Julia Kedrzycki, Gokce Keskin, Edward Kim, Joon Kim, Wei-Yin Ko, Tom Kocmi, Michael Kozakov, Wojciech Kryściński, Arnav Kumar Jain, Komal Kumar Teru, Sander Land, Michael Lasby, Olivia Lasche, Justin Lee, Patrick Lewis, Jeffrey Li, Jonathan Li, Hangyu Lin, Acyr Locatelli, Kevin Luong, Raymond Ma, Lukas Mach, Marina Machado, Joanne Magbitang, Brenda Malacara Lopez, Aryan Mann, Kelly Marchisio, Olivia Markham, Alexandre Matton, Alex McKinney, Dominic McLoughlin, Jozef Mokry, Adrien Morisot, Autumn Moulder, Harry Moynehan, Maximilian Mozes, Vivek Muppalla, Lidiya Murakhovska, Hemangani Nagarajan, Alekhya Nandula, Hisham Nasir, Shauna Nehra, Josh Netto-Rosen, Daniel Ohashi, James Owers-Bardsley, Jason Ozuzu, Dennis Padilla, Gloria Park, Sam Passaglia, Jeremy Pekmez, Laura Penstone, Aleksandra Piktus, Case Ploeg, Andrew Poulton, Youran Qi, Shubha Raghvendra, Miguel Ramos, Ekagra Ranjan, Pierre Richemond, Cécile Robert-Michon, Aurélien Rodriguez, Sudip Roy, Laura Ruis, Louise Rust, Anubhav Sachan, Alejandro Salamanca, Kailash Karthik Saravanakumar, Isha Satyakam, Alice Schoenauer Sebag, Priyanka Sen, Sholeh Sepehri, Preethi Seshadri, Ye Shen, Tom Sherborne, Sylvie Chang Shi, Sanal Shivaprasad, Vladyslav Shmyhlo, Anirudh Shrinivason, Inna Shteinbuk, Amir Shukayev, Mathieu Simard, Ella Snyder, Ava Spataru, Victoria Spooner, Trisha Starostina, Florian Strub, Yixuan Su, Jimin Sun, Dwarak Talupuru, Eugene Tarassov, Elena Tommasone, Jennifer Tracey, Billy Trend, Evren Tumer, Ahmet Üstün, Bharat Venkitesh, David Venuto, Pat Verga, Maxime Voisin, Alex Wang, Donglu Wang, Shijian Wang, Edmond Wen, Naomi White, Jesse Willman, Marysia Winkels, Chen Xia, Jessica Xie, Minjie Xu, Bowen Yang, Tan Yi-Chern, Ivan Zhang, Zhenyu Zhao, and Zhoujie Zhao. Command a: An enterprise-ready large language model, 2025.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging cross-lingual structure in pretrained language models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6022–6034, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.536. URL https://aclanthology.org/2020.acl-main.536.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan,

Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

John Dang, Arash Ahmadian, Kelly Marchisio, Julia Kreutzer, Ahmet Üstün, and Sara Hooker. RLHF can speak many languages: Unlocking multilingual preference optimization for LLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 13134–13156, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main. 729. URL https://aclanthology.org/2024.emnlp-main.729.

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. Aya expanse: Combining research breakthroughs for a new multilingual frontier, 2024b. URL https://arxiv.org/abs/2412.04261.

Leandro Rodrigues de Souza, Rodrigo Nogueira, and Roberto Lotufo. On the ability of monolingual models to learn language-agnostic representations, 2021. URL https://arxiv.org/abs/2109.01942.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. Investigating data contamination in modern benchmarks for large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8706–8719, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.482. URL https://aclanthology.org/2024.naacl-long.482.

Daniel Deutsch, George Foster, and Markus Freitag. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12914–12929, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.798. URL https://aclanthology.org/2023.emnlp-main.798.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Sumanth Doddapaneni, Mohammed Safi Ur Rahman Khan, Dilip Venkatesh, Raj Dabre, Anoop Kunchukuttan, and Mitesh M. Khapra. Cross-lingual auto evaluation for assessing multilingual llms, 2024. URL https://arxiv.org/abs/2410.13394.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational*

Linguistics: ACL 2024, pp. 12039–12050, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.716. URL https://aclanthology.org/2024.findings-acl.716.

Longxu Dou, Qian Liu, Fan Zhou, Changyu Chen, Zili Wang, Ziqi Jin, Zichen Liu, Tongyao Zhu, Cunxiao Du, Penghui Yang, Haonan Wang, Jiaheng Liu, Yongchi Zhao, Xiachong Feng, Xin Mao, Man Tsung Yeung, Kunat Pipatanakul, Fajri Koto, Min Si Thu, Hynek Kydlíček, Zeyi Liu, Qunshu Lin, Sittipong Sripaisarnmongkol, Kridtaphad Sae-Khow, Nirattisai Thongchim, Taechawat Konkaew, Narong Borijindargoon, Anh Dao, Matichon Maneegard, Phakphum Artkaew, Zheng-Xin Yong, Quan Nguyen, Wannaphong Phatthiyaphaibun, Hoang H. Tran, Mike Zhang, Shiqi Chen, Tianyu Pang, Chao Du, Xinyi Wan, Wei Lu, and Min Lin. Sailor2: Sailing in south-east asia with inclusive multilingual llm. *arXiv preprint arXiv:2502.12982*, 2025.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1383–1392, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1128. URL https://aclanthology.org/P18-1128.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=4hturzLcKX.

Kawin Ethayarajh and Dan Jurafsky. Utility is in the eye of the user: A critique of NLP leaderboards. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4846–4853, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.393. URL https://aclanthology.org/2020.emnlp-main.393.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474, 2021. doi: 10.1162/tacl_a_00437. URL https://aclanthology.org/2021.tacl-1.87.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri (eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 46–68, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.2.

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. A framework for few-shot language model evaluation, 07 2024. URL https://zenodo.org/records/12608602.

Jinchao Ge, Zeyu Zhang, Minh Hieu Phan, Bowen Zhang, Akide Liu, and Yang Zhao. Esa: Annotation-efficient active learning for semantic segmentation. *ArXiv*, abs/2408.13491, 2024. URL https://api.semanticscholar.org/CorpusID:271957367.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166, 2023.

Martin Gellerstam. Translationese in swedish novels translated from english. *Translation studies in Scandinavia*, 1:88–95, 1986.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile van Krieken, and Pasquale Minervini. Are we done with mmlu?, 2024. URL https://arxiv.org/abs/2406.04127.

Gemma Team. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022. doi: 10.1162/tacl_a_00474. URL https://aclanthology.org/2022.tacl-1.30.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. Continuous measurement scales in human evaluation of machine translation. In Antonio Pareja-Lora, Maria Liakata, and Stefanie Dipper (eds.), *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pp. 33–41, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-2305.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. Is machine translation getting better over time? In Shuly Wintner, Sharon Goldwater, and Stefan Riezler (eds.), *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 443–451, Gothenburg, Sweden, April 2014a. Association for Computational Linguistics. doi: 10.3115/v1/E14-1047. URL https://aclanthology.org/E14-1047.

Yvette Graham, Nitika Mathur, and Timothy Baldwin. Randomized significance tests in machine translation. In Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, and Lucia Specia (eds.), *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 266–274, Baltimore, Maryland, USA, June 2014b. Association for Computational Linguistics. doi: 10.3115/v1/W14-3333. URL https://aclanthology.org/W14-3333.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23:3 – 30, 2015. URL https://api.semanticscholar.org/CorpusID:41892872.

Yvette Graham, Barry Haddow, and Philipp Koehn. Statistical power and translationese in machine translation evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 72–81, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020 .emnlp-main.6. URL https://aclanthology.org/2020.emnlp-main.6.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025. URL https://arxiv.org/abs/2411.15594.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995, 2024. doi: 10.1162/tacl_a_00683. URL https://aclanthology.org/2024.tacl-1.54.

Yanzhu Guo, Simone Conia, Zelin Zhou, Min Li, Saloni Potdar, and Henry Xiao. Do large language models have an english accent? evaluating and improving the naturalness of multilingual llms, 2024. URL https://arxiv.org/abs/2410.15956.

Srishti Gureja, Lester James V. Miranda, Shayekh Bin Islam, Rishabh Maheshwary, Drishti Sharma, Gusti Winata, Nathan Lambert, Sebastian Ruder, Sara Hooker, and Marzieh Fadaee. M-rewardbench: Evaluating reward models in multilingual settings, 2024. URL https://arxiv.org/abs/2410.15522.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4693–4703, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.finding s-acl.413. URL https://aclanthology.org/2021.findings-acl.413.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. Achieving human parity on automatic chinese to english news translation, 2018. URL https://arxiv.org/abs/1803.05567.

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

Oskar Holmström and Ehsan Doostmohammadi. Making instruction finetuning accessible to non-English languages: A case study on Swedish models. In Tanel Alumäe and Mark Fishel (eds.), *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pp. 634–642, Tórshavn, Faroe Islands, May 2023. University of Tartu Library. URL https://aclanthology.o rg/2023.nodalida-1.62.

Carolin Holtermann, Paul Röttger, Timm Dill, and Anne Lauscher. Evaluating the elementary multilingual capabilities of large language models with MultiQ. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4476–4494, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.265. URL https://aclanthology.org/2024.findings-acl.265.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International conference on machine learning*, pp. 4411–4421. PMLR, 2020.

Kaiyu Huang, Fengran Mo, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, Jian-Yun Nie, and Yang Liu. A survey on large language models with multilingualism: Recent advances and new frontiers, 2024. URL https://arxiv.org/abs/2405.10936.

Adithya Venkatadri Hulagadri, Julia Kreutzer, Jian Gang Ngui, and Xian Bin Yong. Towards fair and comprehensive multilingual llm benchmarking, 2025. URL https://cohere.com/blog/towards-fair-and-comprehensive-multilingual-and-multicultural-llm-benchmarking.

Neslihan Iskender, Tim Polzehl, and Sebastian Möller. Reliability of human evaluation for text summarization: Lessons learned and challenges ahead. In Anya Belz, Shubham Agarwal, Yvette Graham, Ehud Reiter, and Anastasia Shimorina (eds.), *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pp. 86–96, Online, April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.humeval-1.10.

Shaoxiong Ji, Zihao Li, Indraneil Paul, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O'Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, and Barry Haddow. Emma-500: Enhancing massively multilingual adaptation of large language models, 2024. URL https://arxiv.org/abs/2409.17892.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4110–4124, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.324. URL https://aclanthology.org/2021.naacl-main.324.

Douwe Kiela, Tristan Thrush, Kawin Ethayarajh, and Amanpreet Singh. Plotting progress in ai. *Contextual AI Blog*, 2023. https://contextual.ai/blog/plotting-progress.

Ondrej Klejch, Eleftherios Avramidis, Aljoscha Burchardt, and Martin Popel. Mt-compareval: Graphical evaluation interface for machine translation development. *The Prague Bulletin of Mathematical Linguistics*, 104:63 – 74, 2015. URL https://api.semanticscholar.org/CorpusID:11488228.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz (eds.), *Proceedings of the Sixth Conference on Machine Translation*, pp. 478–494, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.57.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 1–42, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.1. URL https://aclanthology.org/2023.wmt-1.1.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, pp. 1–46, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.1. URL https://aclanthology.org/2024.wmt-1.1.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. Error span annotation: A balanced approach for human evaluation of machine translation. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, pp. 1440–1453, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.131. URL https://aclanthology.org/2024.wmt-1.131.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1999–2014, Bangkok, Thailand, August 2024c. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.110. URL https://aclanthology.org/2024.acl-long.110.

Philipp Koehn. Statistical significance tests for machine translation evaluation. In Dekang Lin and Dekai Wu (eds.), *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 388–395, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-3250.

Philipp Koehn and Christof Monz. Manual and automatic evaluation of machine translation between European languages. In Philipp Koehn and Christof Monz (eds.), *Proceedings on the Workshop on Statistical Machine Translation*, pp. 102–121, New York City, June 2006a. Association for Computational Linguistics. URL https://aclanthology.org/W06-3114.

Philipp Koehn and Christof Monz (eds.). *Proceedings on the Workshop on Statistical Machine Translation*, New York City, June 2006b. Association for Computational Linguistics. URL https://aclanthology.org/W06-3100.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In Yansong Feng and Els Lefever (eds.), *Proceedings of the 2023*

Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 318–327, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v 1/2023.emnlp-demo.28. URL https://aclanthology.org/2023.emnlp-demo.28.

Alon Lavie and Abhaya Agarwal. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In Chris Callison-Burch, Philipp Koehn, Cameron Shaw Fordyce, and Christof Monz (eds.), *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 228–231, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/W07-0734.

Sara Laviosa. Corpus linguistics and translation studies. In *Perspectives on corpus linguistics*, pp. 131–154. John Benjamins Publishing Company, 2011.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. RLAIF: Scaling reinforcement learning from human feedback with AI feedback, 2024. URL https://openreview.net/forum?i d=AAxIs3D2ZZ.

Yucheng Li, Yunhao Guo, Frank Guerin, and Chenghua Lin. An open-source data contamination report for large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 528–541, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024 .findings-emnlp.30. URL https://aclanthology.org/2024.findings-emnlp.30.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6008–6018, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.484. URL https://aclanthology.org/2020.emnlp-main.484.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. Wildbench: Benchmarking llms with challenging tasks from real users in the wild, 2024. URL https://arxiv.org/abs/2406.04770.

Llama Team. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Chi-kiu Lo. Extended study on using pretrained language models and YiSi-1 for machine translation evaluation. In Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, and Matteo Negri (eds.), *Proceedings of the Fifth Conference on Machine Translation*, pp. 895–902, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.wmt-1.99.

Adam Lopez. Putting human assessments of machine translation systems in order. In Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia (eds.), *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pp. 1–9, Montréal, Canada, June 2012. Association for Computational Linguistics. URL https://aclanthology.o rg/W12-3101.

Michal Lukasik, Srinadh Bhojanapalli, Aditya Krishna Menon, and Sanjiv Kumar. Teacher's pet: understanding and mitigating biases in distillation. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=ph3AYXpwEb.

Matouš Macháček and Ondřej Bojar. Results of the WMT13 metrics shared task. In Ondrej Bojar, Christian Buck, Chris Callison-Burch, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Herve Saint-Amand, Radu Soricut, and Lucia Specia (eds.), *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pp. 45–51, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-2202.

Shushen Manakhimova, Vivien Macketanz, Eleftherios Avramidis, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. Investigating the linguistic performance of large language models in machine translation. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, pp. 355–371, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.28. URL https://aclanthology.org/2024.wmt-1.28.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7297–7306, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.566. URL https://aclanthology.org/2021.acl-long.566.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. Eurollm: Multilingual language models for europe, 2024. URL https://arxiv.org/abs/2409.16235.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4984–4997, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.448. URL https://aclanthology.org/2020.acl-main.448.

Timothy R. McIntosh, Teo Susnjak, Nalin Arachchilage, Tong Liu, Paul Watters, and Malka N. Halgamuge. Inadequacies of large language model benchmarks in the era of generative artificial intelligence, 2024. URL https://arxiv.org/abs/2402.09880.

Nikita Mehandru, Sweta Agrawal, Yimin Xiao, Ge Gao, Elaine Khoong, Marine Carpuat, and Niloufar Salehi. Physician detection of clinical harm in machine translation: Quality estimation aids in reliance and backtranslation identifies critical errors. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11633–11647, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.712. URL https://aclanthology.org/2023.emnlp-main.712.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations, 2024. URL https://arxiv.org/abs/2411.00640.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15991–16111, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.891. URL https://aclanthology.org/2023.acl-long.891.

Graham Neubig and Junjie Hu. Rapid adaptation of neural machine translation to new languages. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 875–880, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1103. URL https://aclanthology.org/D18-1103.

Shiwen Ni, Xiangtao Kong, Chengming Li, Xiping Hu, Ruifeng Xu, Jia Zhu, and Min Yang. Training on the benchmark is not all you need, 2025. URL https://arxiv.org/abs/2409.01790.

NLLB Team. No language left behind: Scaling human-centered machine translation, 2022. URL https://arxiv.org/abs/2207.04672.

David Ong and Peerat Limkonchotiwat. SEA-LION (Southeast Asian languages in one network): A family of Southeast Asian language models. In Liling Tan, Dmitrijs Milajevs, Geeticka Chauhan, Jeremy Gwinnup, and Elijah Rippeth (eds.), *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pp. 245–245, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.nlposs-1.26. URL https://aclanthology.org/2023.nlposs-1.26.

Simon Ott, Adriano Barbosa-Silva, Kathrin Blagec, Jan Brauner, and Matthias Samwald. Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nature Communications*, 13(1), November 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-34591-0. URL http://dx.doi.org/10.1038/s41467-022-34591-0.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

Irene Plaza, Nina Melero, Cristina del Pozo, Javier Conde, Pedro Reviriego, Marina Mayor-Rocher, and María Grandury. Spanish and llm benchmarks: is mmlu lost in translation?, 2024. URL https://arxiv.org/abs/2406.17789.

Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina (eds.), *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049. URL https://aclanthology.org/W15-3049.

Maja Popović. Agree to disagree: Analysis of inter-annotator disagreements in human evaluation of machine translation output. In Arianna Bisazza and Omri Abend (eds.), *Proceedings of the*

25th Conference on Computational Natural Language Learning, pp. 234–243, Online, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.conll-1.18. URL https://aclanthology.org/2021.conll-1.18.

Maja Popović and Anya Belz. On reporting scores and agreement for error annotation tasks. In Antoine Bosselut, Khyathi Chandu, Kaustubh Dhole, Varun Gangal, Sebastian Gehrmann, Yacine Jernite, Jekaterina Novikova, and Laura Perez-Beltrachini (eds.), *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pp. 306–315, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.gem-1.26. URL https://aclanthology.org/2022.gem-1.26.

Matt Post. A call for clarity in reporting BLEU scores. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor (eds.), *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL https://aclanthology.org/W18-6319.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2685–2702, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.213. URL https://aclanthology.org/2020.emnlp-main.213.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri (eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 578–585, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.52.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 841–848, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.73. URL https://aclanthology.org/2023.wmt-1.73.

Stefan Riezler and John T. Maxwell. On some pitfalls in automatic evaluation and significance testing for MT. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss (eds.), *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 57–64, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0908.

Parker Riley, Daniel Deutsch, George Foster, Viresh Ratnakar, Ali Dabirmoghaddam, and Markus Freitag. Finding replicable human evaluations via stable ranking probability. In Kevin Duh, Helena

Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4908–4919, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.275. URL https://aclanthology.org/2024.naacl-long.275.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A. Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Viraat Aryabumi, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Diress, Sharad Duwal, Daniil Dzenhaliou, Daniel Fernando Erazo Florez, Fabian Farestam, Joseph Marvin Imperial, Shayekh Bin Islam, Perttu Isotalo, Maral Jabbarishiviari, Börje F. Karlsson, Eldar Khalilov, Christopher Klamm, Fajri Koto, Dominik Krzemiński, Gabriel Adriano de Melo, Syrielle Montariol, Yiyang Nan, Joel Niklaus, Jekaterina Novikova, Johan Samir Obando Ceron, Debjit Paul, Esther Ploeger, Jebish Purbey, Swati Rajwal, Selvan Sunitha Ravi, Sara Rydell, Roshan Santhosh, Drishti Sharma, Marjana Prifti Skenduli, Arshia Soltani Moakhar, Bardia Soltani Moakhar, Ran Tamir, Ayush Kumar Tarun, Azmine Toushik Wasi, Thenuka Ovin Weerasinghe, Serhan Yilmaz, Mike Zhang, Imanol Schlag, Marzieh Fadaee, Sara Hooker, and Antoine Bosselut. Include: Evaluating multilingual language understanding with regional knowledge, 2024a. URL https://arxiv.org/abs/2411.19799.

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

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10215–10245, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.802. URL https://aclanthology.org/2021.emnlp-main.802.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. How good is your tokenizer? on the monolingual performance of multilingual language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3118–3135, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.243. URL https://aclanthology.org/2021.acl-long.243.

Beatrice Savoldi, Alan Ramponi, Matteo Negri, and Luisa Bentivogli. Translation in the hands of many: Centering lay users in machine translation interactions. *arXiv preprint arXiv:2502.13780*, 2025.

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. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=fR3wGCk-IXp.

Luísa Shimabucoro, Sebastian Ruder, Julia Kreutzer, Marzieh Fadaee, and Sara Hooker. LLM see, LLM do: Leveraging active inheritance to target non-differentiable objectives. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 9243–9267, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.521. URL https://aclanthology.org/2024.emnlp-main.521.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation, 2024a. URL https://arxiv.org/abs/2412.03304.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, 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. Aya dataset: An open-access collection for multilingual instruction tuning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11521–11567, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.620. URL https://aclanthology.org/2024.acl-long.620.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. Kmmlu: Measuring massive multitask language understanding in korean, 2024a. URL https://arxiv.org/abs/2402.11548.

Guijin Son, Dongkeun Yoon, Juyoung Suk, Javier Aula-Blasco, Mano Aslan, Vu Trong Kim, Shayekh Bin Islam, Jaume Prats-Cristià, Lucía Tormo-Bañuelos, and Seungone Kim. Mm-eval: A multilingual meta-evaluation benchmark for llm-as-a-judge and reward models. *arXiv preprint arXiv:2410.17578*, 2024b.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. Findings of the WMT 2021 shared task on quality estimation. In Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz (eds.), *Proceedings of the Sixth Conference on Machine Translation*, pp. 684–725, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.71.

Miloš Stanojević and Khalil Sima'an. BEER: BEtter evaluation as ranking. In Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, and Lucia Specia (eds.), *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 414–419, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-3354. URL https://aclanthology.org/W14-3354.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. Evaluating gender bias in machine translation. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1679–1684, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1164. URL https://aclanthology.org/P19-1164.

Hui Su, Xiao Zhou, Houjin Yu, Xiaoyu Shen, Yuwen Chen, Zilin Zhu, Yang Yu, and Jie Zhou. Welm: A well-read pre-trained language model for chinese, 2023. URL https://arxiv.org/abs/2209.10372.

Haoran Sun, Renren Jin, Shaoyang Xu, Leiyu Pan, Supryadi, Menglong Cui, Jiangcun Du, Yikun Lei, Lei Yang, Ling Shi, Juesi Xiao, Shaolin Zhu, and Deyi Xiong. FuxiTranyu: A multilingual large language model trained with balanced data. In Franck Dernoncourt, Daniel Preoţiuc-Pietro, and Anastasia Shimorina (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1499–1522, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-industry.110. URL https://aclanthology.org/2024.emnlp-industry.110.

Alex Tamkin, Miles McCain, Kunal Handa, Esin Durmus, Liane Lovitt, Ankur Rathi, Saffron Huang, Alfred Mountfield, Jerry Hong, Stuart Ritchie, Michael Stern, Brian Clarke, Landon Goldberg, Theodore R. Sumers, Jared Mueller, William McEachen, Wes Mitchell, Shan Carter, Jack Clark, Jared Kaplan, and Deep Ganguli. Clio: Privacy-preserving insights into real-world ai use, 2024. URL https://arxiv.org/abs/2412.13678.

Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, and Mehdi Ali. Towards multilingual llm evaluation for european languages, 2024. URL https://arxiv.org/abs/2410.08928.

Brian Thompson, Nitika Mathur, Daniel Deutsch, and Huda Khayrallah. Improving statistical significance in human evaluation of automatic metrics via soft pairwise accuracy. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, pp. 1222–1234, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.118. URL https://aclanthology.org/2024.wmt-1.118.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor (eds.), *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 113–123, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6312. URL https://aclanthology.org/W18-6312.

Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. deep-significance: Easy and meaningful signifcance testing in the age of neural networks. In *ML Evaluation Standards Workshop at the Tenth International Conference on Learning Representations*, 2022.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne

Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. Aya model: An instruction finetuned open-access multilingual language model. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15894–15939, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.845. URL https://aclanthology.org/2024.acl-long.845.

Rob van der Goot. We need to talk about train-dev-test splits. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4485–4494, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.368. URL https://aclanthology.org/2021.emnlp-main.368.

Laurène Vaugrante, Mathias Niepert, and Thilo Hagendorff. A looming replication crisis in evaluating behavior in language models? evidence and solutions, 2024. URL https://arxiv.org/abs/2409.20303.

Joshua Vendrow, Edward Vendrow, Sara Beery, and Aleksander Madry. Do large language model benchmarks test reliability?, 2025. URL https://arxiv.org/abs/2502.03461.

David Vilar, Gregor Leusch, Hermann Ney, and Rafael E. Banchs. Human evaluation of machine translation through binary system comparisons. In Chris Callison-Burch, Philipp Koehn, Cameron Shaw Fordyce, and Christof Monz (eds.), *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 96–103, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/W07-0713.

Jiayi Wang, David Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Tosin Adewumi, Hamam Mokayed, Christine Mwase, Wangui Kimotho, Foutse Yuehgoh, Anuoluwapo Aremu, Jessica Ojo, Shamsuddeen Muhammad, Salomey Osei, Abdul-Hakeem Omotayo, Chiamaka Chukwuneke, Perez Ogayo, Oumaima Hourrane, Salma El Anigri, Lolwethu Ndolela, Thabiso Mangwana, Shafie Mohamed, Hassan Ayinde, Oluwabusayo Awoyomi, Lama Alkhaled, Sana Alazzawi, Naome Etori, Millicent Ochieng, Clemencia Siro, Njoroge Kiragu, Eric Muchiri, Wangari Kimotho, Toadoum Sari Sakayo, Lyse Naomi Wamba, Daud Abolade, Simbiat Ajao, Iyanuoluwa Shode, Ricky Macharm, Ruqayya Iro, Saheed Abdullahi, Stephen Moore, Bernard Opoku, Zainab Akinjobi, Abeeb Afolabi, Nnaemeka Obiefuna, Onyekachi Ogbu, Sam Ochieng', Verrah Otiende, Chinedu Mbonu, Yao Lu, and Pontus Stenetorp. AfriMTE and AfriCOMET: Enhancing COMET to embrace under-resourced African languages. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5997–6023, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.334. URL https://aclanthology.org/2024.naacl-long.334.

Fangyun Wei, Xi Chen, and Lin Luo. Rethinking generative large language model evaluation for semantic comprehension. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=3Cp042s1Nc.

Johnny Wei, Tom Kocmi, and Christian Federmann. Searching for a higher power in the human evaluation of MT. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus

Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri (eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 129–139, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.7.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, Tianxiang Hu, Shangjie Li, Binyuan Hui, Bowen Yu, Dayiheng Liu, Baosong Yang, Fei Huang, and Jun Xie. Polylm: An open source polyglot large language model, 2023. URL https://arxiv.org/abs/2307.06018.

Yi Xu, Laura Ruis, Tim Rocktäschel, and Robert Kirk. Investigating non-transitivity in llm-as-a-judge, 2025. URL https://arxiv.org/abs/2502.14074.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Yun Xing, Junjue Wang, Huitao Li, Xin Li, Kunyu Yu, Nan Liu, Qingyu Chen, Douglas Teodoro, Edison Marrese-Taylor, Shijian Lu, Yusuke Iwasawa, Yutaka Matsuo, and Irene Li. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation, 2025. URL https://arxiv.org/abs/2503.10497.

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. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E. Gonzalez, and Ion Stoica. Rethinking benchmark and contamination for language models with rephrased samples, 2023. URL https://arxiv.org/abs/2311.04850.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, et al. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*, 2024.

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. Pangea: A fully open multilingual multimodal llm for 39 languages. *arXiv preprint arXiv:2410.16153*, 2024. URL https://arxiv.org/abs/2410.16153.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. Findings of the WMT 2022 shared task on quality estimation. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri

(eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 69–99, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.3.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1628–1639, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.148. URL https://aclanthology.org/2020.acl-main.148.

Biao Zhang, Ankur Bapna, Rico Sennrich, and Orhan Firat. Share or not? learning to schedule language-specific capacity for multilingual translation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Wj4ODo0uyCF.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024a. URL https://arxiv.org/abs/2407.12772.

Mike Zhang and Antonio Toral. The effect of translationese in machine translation test sets. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor (eds.), *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pp. 73–81, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5208. URL https://aclanthology.org/W19-5208.

Wenxuan Zhang, Hou Pong Chan, Yiran Zhao, Mahani Aljunied, Jianyu Wang, Chaoqun Liu, Yue Deng, Zhiqiang Hu, Weiwen Xu, Yew Ken Chia, Xin Li, and Lidong Bing. Seallms 3: Open foundation and chat multilingual large language models for southeast asian languages, 2024b. URL https://arxiv.org/abs/2407.19672.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7915–7927, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.491. URL https://aclanthology.org/2023.emnlp-main.491.

Yidan Zhang, Boyi Deng, Yu Wan, Baosong Yang, Haoran Wei, Fei Huang, Bowen Yu, Junyang Lin, Fei Huang, and Jingren Zhou. P-mmeval: A parallel multilingual multitask benchmark for consistent evaluation of llms, 2024c. URL https://arxiv.org/abs/2411.09116.

Yiran Zhao, Chaoqun Liu, Yue Deng, Jiahao Ying, Mahani Aljunied, Zhaodonghui Li, Lidong Bing, Hou Pong Chan, Yu Rong, Deli Zhao, and Wenxuan Zhang. Babel: Open multilingual large language models serving over 90 URL https://arxiv.org/abs/2503.00865.

Zhixue Zhao and Nikolaos Aletras. Comparing explanation faithfulness between multilingual and monolingual fine-tuned language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3226–3244, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.178. URL https://aclanthology.org/2024.naacl-long.178.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023a. URL https://proceedings.neurip s.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_ and_Benchmarks.pdf.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023b. URL https: //openreview.net/forum?id=uccHPGDlao.

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

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. Multilingual machine translation with large language models: Empirical results and analysis. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2765–2781, Mexico City, Mexico, June 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.176. URL https://aclanthology.org/2024.findings-naacl.176.

A   Model Release Benchmarks

Rank	Benchmark	Paper	Model Releases (Benchmarked/Supported Languages)
1	Flores-200	(Costa-jussà et al., 2022)	Aya101 (99/101), Aya Expanse (22/23), Qwen2 (?/≈30), EMMA (199/546), EuroLLM (34/35), PangeaLLM (11/39), SeaLLM (12/12), SEA-LION (4/13), Salamandra (3/35), Babel (25/25), Sailor2 (15/15)
2	MGSM	(Shi et al., 2023)	Aya Expanse (7/23), Llama3 (7/8), Qwen2 (10?/≈30), EMMA (10/546), PangeaLLM (10?/39), SeaLLM (6/12), Salamandra (5/35), Babel (10?/25)
3	XLSum	(Hasan et al., 2021)	Aya101 (45/101), EMMA (44/546), FuxiTranyu (15/43), SEA-LION (4/13), Salamandra (2/35)
4	WMT	(Kocmi et al., 2024a, inter alia)	EuroLLM (16/35), FuyiTranyu (3/43), PolyLM (4/8+)
	Dolly translated	(Singh et al., 2024b)	Aya101 (3/101), Aya Expanse (23/23), EMMA (119/546)
5	mArenaHard	(Dang et al., 2024b)	Aya Expanse (23/23)
	Aya human-translated	(Singh et al., 2024b)	Aya101 (5/101)
	PolyWrite	(Ji et al., 2024)	EMMA (240/546)
	SeaBench	(Zhang et al., 2024b)	SeaLLM (3/12)
	SeaMTBench	(Ong & Limkonchotiwat, 2023)	SEA-LION (6/13)
	SEA-IFEval	(Ong & Limkonchotiwat, 2023)	SEA-LION (6/13)
	MTG	(Chen et al., 2022)	PolyLM (5/8)

Table 5: We rank open benchmarks from Table 1 on their popularity in model release reports. For each model we indicate in how many of its supported languages the model is evaluated. For WMT General Benchmarks, we report the union of all subsets.

Table 5 indicates which of these benchmarks were included in recent open (explicitly) multilingual model releases,[7] including Aya-101 (Üstün et al., 2024), Aya Expanse (Dang et al., 2024b), Llama3 (Llama Team, 2024), Qwen2 (Yang et al., 2024), EMMA-500 (Ji et al., 2024) (base model), EuroLLM (Martins et al., 2024), PangeaLLM (Yue et al., 2024) (multi-modal), FuxiTranyu (Sun et al., 2024), PolyLM (Wei et al., 2023), SeaLLMs (Zhang et al., 2024b), SEA-LION (Ong & Limkonchotiwat, 2023), Salamandra (Barcelona Supercomputing Center, 2024), Babel (Zhao et al., 2025), Sailor2 (Dou et al., 2025).[8]

B   Multilingual Leaderboards

Table 6 lists open, non-English leaderboards and the models from our overview in Section 2 that they evaluate. We also report whether as of the current state, multilingual models or specialized target language models are in the lead in the size of up to 14B parameters.

---

[7]Other models such as Gemma2 (Gemma Team, 2024) might have multilingual capabilities but are not explicitly stating that they do.

[8]We excluded models that do not report any generative evaluations, such as Ministral and Mixtral, or those that are only in-house (Qwen). Models might additionally have been benchmarked by external parties or competing model releases (*e.g.* EuroLLM benchmarks Gemma on translation tasks).

	# Languages	Language Focus	Evaluated Open mLLMs	Focus Language(s) LLM win?
European LLM Leaderboard	21	European	Llama3, EuroLLM, Qwen2, Aya23	no
African Languages LLM Eval Leaderboard	18	African	Llama3, Aya101	no
SEA HELM	4	South-East Asian	Llama3, Qwen2, SeaLLMs, SEA-LION, Aya Expanse, Aya23	yes
Indic LLM Leaderboard	7	Indic	Llama3	no
Open Japanese LLM Leaderboard	1	Japanese	Llama3, Qwen2, Aya Expanse	yes
Open Ko-LLM Leaderboard	1	Korean	Qwen2, SeaLLM	yes
Open Persian Leaderboard	1	Persian	Qwen2, Aya Expanse, Llama3	no
Open Portuguese Leaderboard	1	Portuguese	Qwen2, Llama3, Aya Expanse, Aya23, SeaLLM	yes
Open Chinese Leaderboard	1	Chinese	Qwen2, Llama3, SeaLLM	yes
Open Arabic Leaderboard	1	Arabic	Qwen2, Llama3, Aya Expanse, SeaLLM, Aya23, EuroLLM	yes
CzechBench Leaderboard	1	Czech	Llama3	no
Hebrew LLM Leaderboard	1	Hebrew	SeaLLM, Aya Expanse, Qwen2, Aya23, Llama3	yes
Open PL LLM Leaderboard	1	Polish	Llama3, Qwen2, Aya Expanse, EuroLLM, Aya 23	yes
OpenLLM Turkish Leaderboard	1	Turkish	Aya Expanse, Aya 23, EuroLLM, Llama3, Qwen2	yes
Open LLM French Leaderboard	1	French	Qwen2, Llama3, EuroLLM	no

Table 6: Non-English leaderboards evaluating multilingual models with their focus languages and evaluated open mLLMs from Table 5. Based on the average ranking on the respective leaderboards, we measure if LLMs for the respective focus languages win over the more massively multilingual ones, restricted to models below 13B parameters. For leaderboards that involve multiple languages, we aggregate wins via majority votes. This table reflects the state of 10 February 2025, 7 March 2025 for the French leaderboard.

C   Generative Evaluation In Disguise: MMLU

Even though MMLU (Hendrycks et al., 2021) is by design a discriminative task (MCQA), it deserves to be discussed here as the most popular benchmark for multilingual models to-date because of the seemingly ease of evaluation (one of four options is correct). The original MMLU data has been translated automatically and with humans in various efforts with various translation tools (X-MMLU in Okapi work, MMMLU (openAI), LLama3 report uses Google Translate for translation, GlobalMMLU (Singh et al., 2024a)), replicated in other languages (Son et al., 2024a; Xuan et al., 2025), analyzed and corrected (MMLU-Redux (Gema et al., 2024)), sub-categorized (GlobalMMLU), extended (INCLUDE (Romanou et al., 2024a)), and critizices (Balepur et al., 2025). However, since LLMs are generators by design, evaluation is not straightforward, therefore multiple approaches exist,[9] such as based on likelihood rankings of answers, or exact string matching. These details are rarely specified in multilingual MMLU evaluations, but may make the difference for system ranking (Wei et al., 2024). MCQA tasks (the majority of them) can also be turned into a generative benchmark by stripping the answer options from the prompt and having a LLM judge decide whether the model's generation matches the correct answer option, possibly in comparison with another model's generation (Wei et al., 2024). The generative form of evaluation has so far not been explored multilingually.

D   Multilingual vs Monolingual Models and Benchmarks

Individual language benchmarks and models often receive little recognition in multilingual LLM development. While extensive work on English monolingual LLMs is widely respected and adapted for other languages, monolingual or less massively multilingual models tend to be overlooked. One challenge in evaluation arises when moving beyond monolingual to multilingual settings, as the language coverage of individual benchmarks and models often does not fully align. This mismatch can lead to benchmarks being deemed incomplete for certain models or, being overly extensive, unfairly penalizing models for languages they do not support.

---

[9]https://huggingface.co/blog/open-llm-leaderboard-mmlu

One argument against specialization is the potential for *sharing of information* in multilingual settings where knowledge learned from one language can benefit others (Conneau et al., 2020; Artetxe et al., 2020b; de Souza et al., 2021), or general reasoning abilities transfer (Chang et al., 2024), especially with increased model sizes (Chang et al., 2023). Building monolingual models, however, provides the opportunity to specialize the model (Chang et al., 2024), including optimizing tokenization strategies (Chelombitko & Komissarov, 2024; Zhao & Aletras, 2024) and tailoring the training data to the specific linguistic characteristics of the target language (Su et al., 2023; Abonizio et al., 2025).

D.1   Experiment: Monolingual vs Multilingual Model Performance

We compare a multilingual model, AYA EXPANSE 8B, with individual monolingual models on two open-ended generation tasks: general knowledge (Singh et al., 2024b), and a more challenging set of math, code, and reasoning questions (Dang et al., 2024b). We cover a diverse set of monolingual models, including those pretrained from scratch exclusively on a single language, as well as models obtained by specialized finetuning of a multilingual model. Our selection also spans models with language-specific tokenizers versus general tokenizers, and models specialized for domains such as code or math in contrast to general-purpose language models. The languages include French[10], Hebrew[11], Chinese[12], Arabic[13], and Japanese[14]. GPT4-o is used as a judge to evaluate the quality of generations in a pairwise comparison setting.

Figure 11 shows that, in almost all cases, the multilingual model outperforms the monolingual counterparts in both evaluation sets. Previously on a smaller scale Rust et al. (2021) compared mBERT (Devlin et al., 2019) and pretrained monolingual BERT models across selected languages and showed that languages adequately represented in the multilingual model's vocabulary exhibit little to no performance degradation compared to their monolingual counterparts. Now at a larger scale, ranging from 3B to 9B parameters, we observe an even stronger pattern of multilingual models outperforming their monolingual counterparts.

Aside from the challenges of properly configuring an entirely new model to generate coherent text in each new language, multilingual models also appear more powerful for open-ended generation tasks, consistently producing stronger outputs across languages. However, this advantage may also stem from factors such as a higher number of experimental iterations, broader (rather than specialized) evaluation objectives, and the continuous updating and maintenance of multilingual models — benefits that monolingual models, often developed in a more "one-and-done" style might lack.

---

[10] https://huggingface.co/jpacifico/Chocolatine-3B-Instruct-DPO-v1.2
[11] https://huggingface.co/dicta-il/dictalm2.0-instruct
[12] https://huggingface.co/01-ai/Yi-1.5-9B-Chat
[13] https://huggingface.co/CohereForAI/c4ai-command-r7b-arabic-02-2025
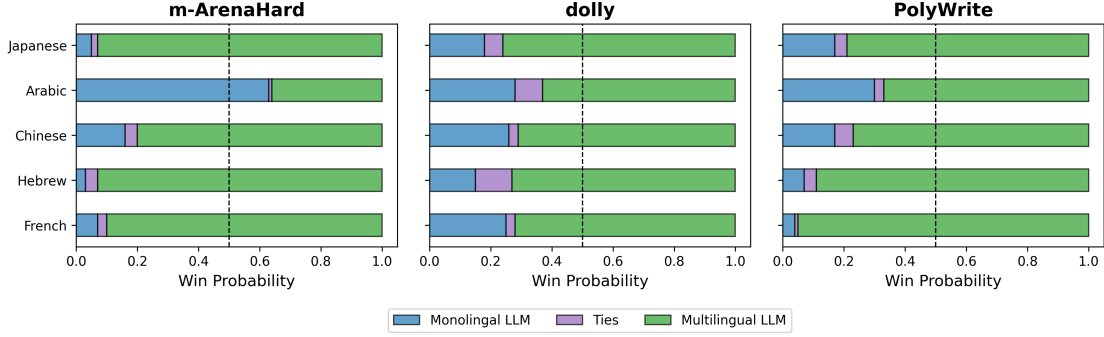[14] https://huggingface.co/llm-jp/llm-jp-3-7.2b-instruct3

Figure 11: Comparing performance of a multilingual model (Aya Expanse 8B) with language-expert Monolingual models on general open-ended questions (dolly-translated-200, (Singh et al., 2024b) and creative writing prompts from PolyWrite (Ji et al., 2024)) and a more challenging set of math, code, and reasoning questions (m-ArenaHard, (Dang et al., 2024b))

E   Win Rate Comparisons

E.1   Sampling

For our win rate comparisons in section 3.2, we sample 5 generations from each model with ancestral sampling (temperature=1.0) as implemented in vLLM. We chose this setup because we did not want to tune temperatures individually for each model, nor was there any guide from either model provider how to set it in the best way. In hindsight, we noticed that a lower temperature would have been beneficial for Qwen2.5, which explains why our Aya Expanse 8B winrates are more inflated than those in the Aya Expanse tech report (Dang et al., 2024b), especially for languages that had already lower quality. Upon communication with the authors, we found out that their evaluations were run with temperature=0.75, and we were able to confirm with spot checks of a few languages (from varying win rate buckets) that Qwen2 generations were of higher quality under that setup, see Table 7. Win rates differences under different temperatures vary heavily, up to around 50 points in the most extreme case. For Japanese and Portuguese, even the directionality of the wins change: under temperatures 0.0 or 0.75, Qwen 2.5 wins overall, while under temperature=1.0, Aya Expanse wins overall. Generally, this highlights how essential the documentation of decoding parameters is for replication.

E.2   LLM-as-a-Judge Prompting

For LLM-as-a-judge evaluations, we use the prompts listed in table 8 and randomize the order of model generations to prevent position bias. When using GPT4o as a judge, we use version 2024-11-20.

E.3   Statistical Significance

Preliminary experiments with GPT4o-mini (2024-07-18) for the setup described in section 3.2 revealed that standard errors for win-rates were much higher than for GPT4o, so that even 500 examples are not enough for the differences to be significant in Chinese.

Language	Temperature	WR Aya Expanse	WR Qwen2.5	WR Δ
hi	0.0	74.4	25.0	49.4
	0.75	76.8	22.2	54.6
	1.0	89.2	10.2	**79.0**
fa	0.0	59.0	40.6	18.4
	0.75	71.0	28.2	42.8
	1.0	90.4	9.2	**81.2**
ja	0.0	44.8	54.8	-10.0
	0.75	47.8	51.1	-3.3*
	1.0	70.8	28.2	**42.6**
pt	0.0	43.6	55.6	-12.0
	0.75	41.8	57.0	-15.2
	1.0	58.8	39.8	**19.0**

Table 7: The effect of temperature settings on win rates (WR, in %) on mArenaHard for pairwise comparisons between Aya Expanse 8B and Qwen2.5 7B Instruct. Win rate differences are notably higher under temperature=1.0. Non-significant differences (95% confidence interval) are marked with asterisk.

F   Instruction wording can change ranking

In this experiment, we use the German questions from Include 44 (Romanou et al., 2024b) test set containing localized multiple choice questions. MCQA test sets are usually evaluated with log-likelihood probability, however, when that is not possible, especially when comparing against models behind API, researchers reformulate the questions into instruction following.

In this experiment, we show how much the instruction can change final system ranking, thus opening a room for metric hacking. We design six different instructions and also translate them into German, all listed in appendix F. The model outputs is then automatically parsed with regular expressions to select the proper answer. The final prompt contains the instruction followed with the question and list of all four answers and only the instruction is changed between experiments.

appendix F shows how different wording changes the final model accuracy for three different models.

details of the experiment: Include 44 dataset, only german, different preprompt before each question. The answer is parsed automatically.

The numbering is aligned, thus EN1 = DE1 only translated.

G   Translation Effects

G.1   Experimental Setup

The decoding setup is the same as for the win rate experiments described in Appendix E. We choose the set of languages because they are in the common set of supported languages for our models of interest. We translate the original prompts via a pivot language back into the original language to

System	You are a helpful assistant whose goal is to select the preferred (least wrong) response for a given instruction in `language_name`.
Judge	Which of the following responses is the best one for the given instruction in `language_name`? A good response should follow these rules: 1) It should be in `language_name`, 2) It should complete the request in the instruction, 3) It should be factually correct and semantically comprehensible, 4) It should be grammatically correct and fluent.
	Instruction: `instruction`
	Response (A): `completion_a`
	Response (B): `completion_b`
	FIRST provide a concise comparison of the two responses. If one Response is better, explain which you prefer and why. If both responses are identical or equally good or bad, explain why.
	SECOND, on a new line, state exactly one of 'Response (A)' or 'Response (B)' or 'TIE' to indicate your choice of preferred response.
	Your response should use the format: Comparison: <concise comparison and explanation> Preferred: <'Response (A)' or 'Response (B)' or 'TIE'>

Table 8: Prompts for LLM-as-a-judge evaluations

simulate translation effects on the prompts. The pivot language is English for all languages except English, and Portuguese for English. We translate with Google Translate, NLLB-3.3, Aya Expanse 32B and Command A to have a diverse mix of MT and mLLM translators. For NLLB, we split the prompt into individual sentences with the `sentence_splitter` library,[15] before translation, and concatenate the translations. We do not post-process the translations in any way, but we notice that the translations contain <unk>s, which can throw off generation models.

The translation template for Aya Expanse 32B and Command A is the following: "You are a professional translator. Translate from `src_language` into `target_language`. Return nothing but the translation.
n" We did not do extensive prompt tuning, but noticed that Aya Expanse 32B often answered prompts rather than translating them if we did not include an explicit instruction to only return the translation.

We evaluate outputs from Llama3.1 8B Instruct (Llama Team, 2024), Gemma2 9B (Gemma Team, 2024), Aya Expanse 8B, and Qwen2.5 7B Instruct (Yang et al., 2024) models.

G.2 Translation Quality

Table 10 compares the corpus ChrF (Popović, 2015) and XCOMET-XL (Guerreiro et al., 2024) scores of roundtrip translations, and reference-free `wmt23-cometkiwi-da-xl` (Rei et al., 2023) scores for translation models on `aya-human-annotated` prompts.[16] According to the roundtrip evaluation against the original prompt, Google Translate delivers the highest quality translations with a small

---

[15]https://github.com/mediacloud/sentence-splitter
[16]Sacrebleu signature: `nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.5.1`.

	Aya Expanse 8B	Llama 3.1 8B	Gemma 2 9B
EN 1	46.0%	50.4%	48.2%
EN 2	44.6%	48.2%	47.5%
EN 3	46.8%	48.2%	47.5%
EN 4	48.9%	48.2%	49.6%
EN 5	48.9%	50.4%	48.9%
EN 6	42.4%	51.1%	51.1%
DE 1	45.3%	51.8%	50.4%
DE 2	48.2%	46.8%	51.1%
DE 3	46.8%	48.9%	48.9%
DE 4	46.8%	48.2%	48.2%
DE 5	45.3%	51.8%	48.9%
DE 6	43.9%	47.5%	48.2%

Table 9: Comparison of performance on German MCQA testset from Include 44 when using different instructions.

margin over COMMAND A, followed by NLLB 3.3B and AYA EXPANSE 32B. NLLB translates notably better into Arabic and English than AYA EXPANSE 32B, while AYA translates better into Turkish and Chinese.

G.3   Changes in Generation

We want to measure how translation affects the generations. For that purpose, we compute Spearman correlation between translation quality and the generation quality, relative to the untranslated version. Both quantities are computed with sentence-level ChrF. In Table 11 we report these correlations for various mLLMs and translation models. Overall, we find that correlations are always positive, meaning the better the prompt is translated, the closer the generation is to the generation for the untranslated prompt. For translations from NLLB and AYA this correlation is stronger than for Google Translate, as they also have lower quality and thereby cause more changes to the prompts. Across languages and translations, QWEN is the most susceptible to changes in the prompt.

G.4   Changes in Win Rate

table 12 lists the win rates for all languages in the comparison of AYA EXPANSE 8B vs GEMMA 2 9B under GPT4o as a judge (2024-05-13). We can see that the translation of prompts affects win rates across the bench, with a magnitude depending on the language and translation model.

H   Chatbot Arena Analysis

H.1   Multilinguality

The recently released 100k conversations and preferences collected between June and August 2024 (`lmarena-ai/arena-human-preference-100k`) are more multilingual (54%) than the previously released 33k data (`lmsys/chatbot_arena_conversations`) from April to June 2023 that was 88% English. The set of common languages with more than 200 prompts in each is: English, German,

```
En 1: "Here's a multiple-choice question with answer options. Please respond with
↪  only the letter of the correct choice. Do not include any additional
↪  information in your answer.",
En 2: "Please examine the following multiple-choice question carefully and reply
↪  with just the letter corresponding to the correct answer. No additional text
↪  should be included in your response.",
En 3: "Select the correct option and reply with its corresponding letter only.
↪  Nothing else.",
En 4: "Here's a question for you! Just type the letter of the correct answer - do
↪  not provide any explanations or extra words.",
En 5: "Respond using only the letter of the correct answer. Do not add anything
↪  else.",
En 6: "Whatever you do, don't send anything besides the letter of the correct
↪  answer. No explanations, no extra words!",

De 1: "Hier ist eine Multiple-Choice-Frage mit Antwortoptionen. Bitte antworte nur
↪  mit dem Buchstaben der richtigen Auswahl. Füge deiner Antwort keine
↪  zusätzlichen Informationen hinzu.",
De 2: "Bitte überprüfe die folgende Multiple-Choice-Frage sorgfältig und antworte
↪  nur mit dem Buchstaben der richtigen Antwort. Füge deiner Antwort keinen
↪  zusätzlichen Text hinzu.",
De 3: "Wähle die richtige Option aus und antworte nur mit dem entsprechenden
↪  Buchstaben. Sonst nichts.",
De 4: "Hier ist eine Frage für dich! Gib nur den Buchstaben der richtigen Antwort
↪  ein - keine Erklärungen oder zusätzlichen Wörter.",
De 5: "Antworte nur mit dem Buchstaben der richtigen Antwort. Füge nichts Weiteres
↪  hinzu.",
De 6: "Was auch immer du tust, sende nichts außer dem Buchstaben der richtigen
↪  Antwort. Keine Erklärungen, keine überflüssigen Worte!"
```

Figure 12: English and German instructions for multiple choice question answering.

Spanish, French, Portuguese, Russian.

H.2   Ties

Figure 13 shows the ratio of ties in human pairwise ratings for the five most prominent non-English languages from a total of 100k Chatbot Arena battles that were collected between June and August 2024 (`lmarena-ai/arena-human-preference-100k`), and fig. 14 the same stats for the released battles from April to June 2023 (`lmsys/chatbot_arena_conversations`).

Model	Language	ChrF	XComet	CometKiwi
Google Translate	ar	69.49	88.88	70.76
	en	86.52	96.54	80.87
	pt	81.77	96.95	80.83
	tr	75.94	97.33	76.09
	zh	32.15	88.52	70.68
	Avg	*69.17*	*93.65*	*75.84*
NLLB 3.3B	ar	61.05	84.52	68.83
	en	79.88	95.52	81.12
	pt	75.29	95.71	78.77
	tr	60.66	92.60	71.01
	zh	18.29	81.80	65.80
	Avg	*59.03*	*90.03*	*73.11*
Aya Expanse 32B	ar	35.11	80.83	66.01
	en	77.11	96.55	81.15
	pt	75.70	95.22	79.08
	tr	62.88	95.38	74.99
	zh	22.26	84.89	68.64
	Avg	*54.61*	*90.57*	*73.97*
Command A	ar	62.80	86.60	69.47
	en	82.72	96.32	78.25
	pt	67.06	96.00	79.63
	tr	81.90	97.12	75.88
	zh	37.12	87.99	67.09
	Avg	*66.31*	*92.80*	*74.07*

Table 10: Translation quality of prompt roundtrip translations of the Aya human annotated benchmark. ChrF and XCOMET are reference-based metrics computed for translations from pivot language to target language, and quality is estimated without references for the translation into the pivot language with COMET-KIWI.

Model	NLLB 3.3B	Aya Expanse 32B	Command A	Google Translate	*Avg*
Qwen 2.5 7B Instruct	0.31	0.58	0.23	0.57	***0.34***
Gemma2 9B it	0.41	0.43	0.34	0.28	*0.29*
Aya Expanse 8B	0.37	0.35	0.28	0.26	*0.25*
Llama3.1 7B Instruct	0.23	0.21	0.20	0.15	*0.16*
Avg	***0.34***	*0.33*	*0.23*	*0.30*	*0.24*

Table 11: Pearson correlation between translation quality and generation quality, for several translation and generation models, averaged across languages. NLLB translations correlate the strongest with changes in generations, and Qwen seems most susceptible to transalation artifacts in prompts.

	OG	Translator			
		NLLB	GT	AYA	CMD A
ar	0.71	0.68	0.58	**0.84**	0.74
en	0.01*	-0.05*	0.04*	**0.15**	0.06*
pt	0.08*	0.09*	**0.13**	0.02*	0.00*
tr	0.29	**0.43**	0.42	0.40	0.42
zh	-0.21*	0.16	-0.19*	**0.20**	-0.08*
Avg	*0.18*	*0.26*	*0.20*	***0.32***	*0.23*

Table 12: Win-rate differences (AYA EXPANSE 8B-GEMMA2 9B, *non-significant) on original prompts (OG) vs translated prompts from various translation models. Positive values mean that AYA EXPANSE wins, negative values mean that GEMMA wins.
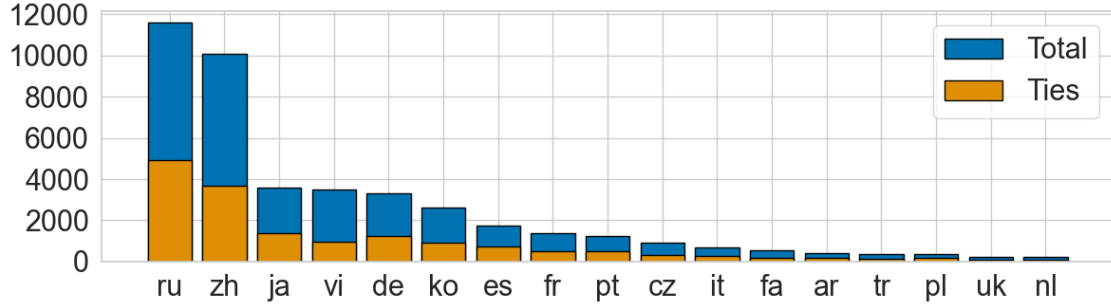

Figure 13: Number of total and tied Chatbot Arena battles (total 100k) for non-English languages with more than 200 prompts from 2024.
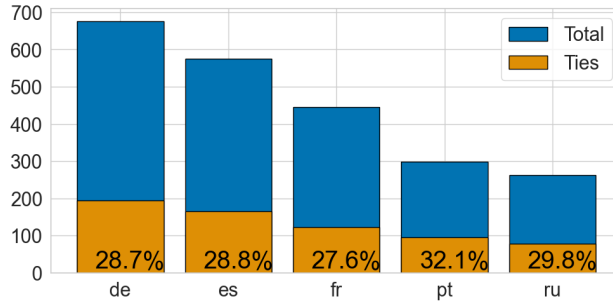

Figure 14: Number of total and tied Chatbot Arena battles (total 33k) for non-English languages with more than 200 prompts from 2023.

I Checklist for Multilingual LLM Evaluation

I.1 Evaluation Prompts

☐ Are evaluation prompts representative samples of all languages included in the evaluation?

☐ Are evaluation prompts human-curated, localized or edited?

☐ If using model generated prompts: Have you analyzed the data for potential biases?

☐ If using translations: Have you estimated, reported, and attempted to optimize translation quality on this particular set of prompts?

I.2 Choice of Metrics

☐ Are metrics adequate for all evaluated languages?

I.3 Statistical Testing

☐ Does the evaluation include adequate statistical significance tests for all included languages?

☐ Does the evaluation include an estimate of statistical power?

☐ If using stochastic decoding, does it include estimates of sampling induced variance?

I.4 Aggregating Results Across Languages

☐ Is the aggregation of results disproportionately influenced by any outliers?

☐ Are language support differences taken into consideration when aggregating results across languages?

☐ Are language support differences documented with the results?

☐ Are task- and language-specific scores reported?

I.5 Qualitative Insights

☐ Are quantitative metrics accompanied by qualitative error analyses?

☐ Are differences in metrics due to meaningful distinctions rather than incidental artifacts?

I.6 Reproducibility

☐ Are the results calculated with standardized pipelines?

☐ Is the evaluation code released?

☐ Are exact evaluation prompts and format published?

☐ Are model outputs released?

☐ Are prompt-level evaluation scores released?

☐ Are metric hyperparameters documented?

☐ Are model versions documented?

I.7 Enabling Meta-Evaluation

☐ Are any human evaluations (consensually) released?