ChiKhaPo: A Large-Scale Multilingual Benchmark for Evaluating Lexical Comprehension and Generation in Large Language Models

Emily Chang[1] **and Niyati Bafna**[2]
[1] Toyota Technological Institute at Chicago;
[2] Johns Hopkins University, Center for Language and Speech Processing

Abstract

Existing benchmarks for large language models (LLMs) are largely restricted to high- or mid-resource languages, and often evaluate performance on higher-order tasks in reasoning and generation. However, plenty of evidence points to the fact that LLMs lack basic linguistic competence in the vast majority of the world's 3800+ written languages. We introduce ChiKhaPo, consisting of eight subtasks of varying difficulty designed to evaluate the lexical comprehension and generation abilities of generative models. ChiKhaPo draws on existing lexicons, monolingual data, and bitext, and provides coverage for 2700+ languages for two word-translation-based subtasks, surpassing any existing benchmark in terms of language coverage. We further show that six SOTA models struggle on our benchmark, and discuss the factors contributing to performance scores, including language family, language resourcedness, task, and comprehension versus generation directions. With ChiKhaPo, we hope to enable and encourage the massively multilingual benchmarking of LLMs.[1]

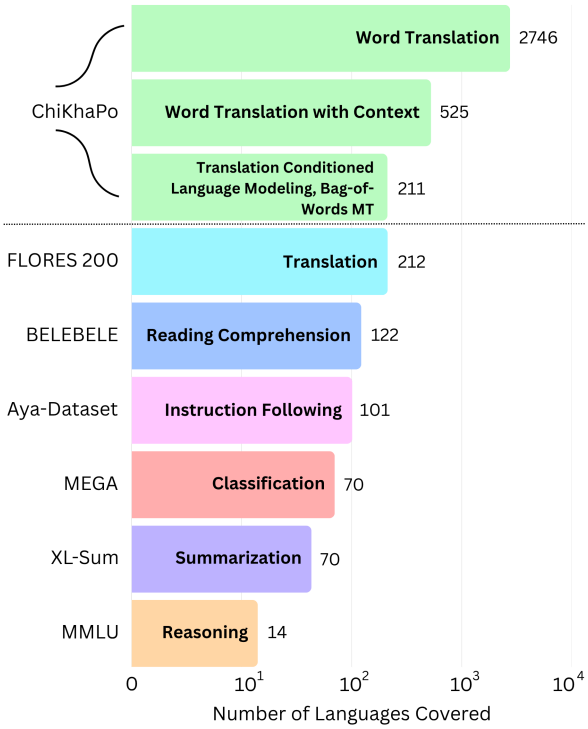

Figure 1: ChiKhaPo evaluates basic lexical competence with several tasks, covering an order of magnitude more languages than existing multilingual benchmarks.

1 Introduction

Benchmarks are crucial for not only measuring but steering progress in NLP (Ruder, 2021). While LLMs are capable of impressive feats of complex reasoning and content generation (DeepSeek-AI et al., 2025; Bercovich et al., 2025; Chen et al., 2025), these capabilities are restricted to a few dozen high-resource languages (HRLs) among 3800+ written languages and dialects in the world (Aji et al., 2022; Ebrahimi et al., 2022). The availability of evaluation benchmarks reflects this problem, with the most multilingual of these at the time of writing being FLORES+ (NLLB Team et al., 2024), which tests machine translation (MT) for 212 languages. For the rest of the world's languages, we have no way to assess even basic LLM capabilities.

We introduce ChiKhaPo, a benchmark that measures basic lexical comprehension and generation abilities in LLMs on a massively multilingual scale.[2] ChiKhaPo includes 4 tasks $\times$ 2 evaluation directions. The tasks provide various perspectives on lexical competence, and the evaluation directions measure model ability for lexical *comprehension* (`X`$\rightarrow$`model`) and *generation* (`model`$\rightarrow$`X`) per

[1]We release our dataset, code for our experiments, and package for running our benchmark.
Dataset: huggingface.co/datasets/ec5ug/chikhapo
Code: github.com/ec5ug/chikhapo/
Python package: pypi.org/project/chikhapo/

[2]The name is inspired by the Hokkien saying that progress is made step-by-step: *chit kha-po, chit kha-in*.

task. The tasks include 1) **word translation (WT)**, involving direct prompting for word translation, 2) **word translation with context (WTWC)**, involving direct prompting for word translation with source context cues, 3) **translation-conditioned language modeling (TCLM)**, involving next word generation given source and target language context in a natural machine translation setting, and 4) **bag-of-words machine translation (BOW MT)**, involving word generation as part of a sentence-level translation task. Each task and direction is evaluated at the word level for a target language.[3]

ChiKhaPo's subtasks make use of existing lexicons, monolingual data, and bitext. In particular, WT relies solely on lexicons, and WTWC additionally requires monolingual data. Both resources are widely available for many languages (Kamholz et al., 2014; ImaniGooghari et al., 2023); thus, ChiKhaPo covers 2700+ and 500+ languages for these tasks respectively, which surpasses the coverage of any existing benchmark (see Figure 1). We also show that performance on WT is correlated with sentence-level MT performance, providing a simple proxy in the absence of bitext.

We evaluate 6 state-of-the-art multilingual LLMs on our benchmark. We provide an analysis of the factors affecting their performance, such as subtask, language resourcedness, and language family, and thus highlight several avenues of focus for improving the broad multilingual competence of LLMs.

ChiKhaPo aims to fill two important gaps in current benchmarks. First, it evaluates *core lexical abilities* in LLMs and allows us to track the "atomic" word-level competence of an LLM in a given language. Second, it does so on a *massively multilingual scale*. With this work, we hope to draw attention to the pressing issue of language inequity in NLP (Joshi et al., 2020), and promote the massively multilingual evaluation of LLMs.

2 Related Work

LLM evaluation benchmarks Most existing benchmarks that LLMs are evaluated on focus on English and other high-resource languages (Grattafiori et al., 2024; Aryabumi et al., 2024; Qwen et al., 2025). Popular benchmark suites include BIG-Bench (Srivastava et al., 2023)—a collection of 200 tasks testing various kinds of comprehension and generation—and HELM (Liang et al., 2023), a framework that standardizes LLM reasoning and generation and provides metrics beyond accuracy (e.g. calibration). Datasets such as XNLI (Conneau et al., 2018) and XCOPA (Ponti et al., 2020) measure reasoning skills with classification-style tasks, whereas natural language generation is evaluated with datasets for summarization, machine translation, and instruction following, such as XL-SUM (Hasan et al., 2021), FLORES+ (NLLB Team et al., 2024), BOUQuET (Andrews et al., 2025; Team et al., 2026), and the Aya Evaluation Suite (Singh et al., 2024).

In Appendix Table 4, we list 20+ commonly used datasets in LLM multilingual benchmarking. These datasets test a collection of relatively complex tasks and cover a limited number of languages.

Lexical evaluation McCarthy (2002) first introduced *lexical substitution*, the task of choosing an appropriate substitute for a word given a context to test word sense disambiguation systems. Prior lexical substitution benchmarks are overwhelmingly English (McCarthy and Navigli, 2007; Kremer et al., 2014; Lee et al., 2021) These benchmarks are small and manually designed.

In implementing ChiKhaPo, we adopted the approach of Mihalcea et al. (2010) who coined the term *cross-lingual lexical substitution*, and evaluated lexical understanding using translations rather than paraphrases. Martínez et al. (2024) uses expert-designed vocabulary tests to perform a fine-grained evaluation of LLMs; however, the benchmark is limited to English and Spanish.

As far as we know, our work is the first to design a lexical competence benchmark with a massively multilingual scope using existing resources.

3 Tasks

ChiKhaPo's suite of tasks centers on lexical semantics, the branch of semantics concerned with word meaning. A word has two meanings: grammatical and lexical. While grammatical meaning refers to the word's function in a language (e.g. plurality, tense), we focus on the word's *lexical meaning*, or the denotative meaning of the base word (Pustejovsky, 2016).

Given the English-centricness of LLMs (Wendler et al., 2024), we treat the model's ability to translate a word *into English* as a proxy for its comprehension of the word (`X→model`), and its

[3] In this paper, the term "target language" refers to the language being evaluated, which may not be the language being generated. We use the terms "source-side" and "target-side" instead to refer to the input and output languages of the model.

Task	Comprehension: `X` → `model`	Generation: `model` → `X`
Word Translation	**Input**: Translate the following text from Malay to English: `ujan`. **Correct Output**: `rain` **Model Output**: `rain` **Score**: `scores["ujan"] += 1`	**Input**: Translate the following text from English to Afrikaans: `attacked`. **Correct Output**: `aangeval` **Model Output**: `aangeval` **Score**: `scores["aangeval"] += 1`
Word Translation with Context	**Input**: In 'Minonke konam phoro isi sonturi aghaipo aro anang pen Jisu yok honsi kido, aro alok hel, labadi chiklik hel aro ajat jat kachiplang theksi, anali chiphere detno, aro pulo, "Khanangsi labang arlengpo Arnam Aso kido . " ', the word '`kido`' means ____ in English. **Correct Output**: `letter` **Model Output**: `child` **Score**: `scores["kido"] += 0`	**Input**: In 'After the match, King of Clay said, "I am just excited about being back in the final rounds of the most important events. I am here to try to win this." ', the word 'win' means ____ in Basque. **Correct Output**: `aurrea hartu` **Model Output**: `ganar` **Score**: `scores["aurrea hartu"] += 0`
Translation-Conditioned Language Modeling	**Input**: Translate the following text into English: Dyula: `Aka dugutaga se'n fei, Iwasaki ye kassara chaman le sôrô.` English: `During his trip, Iwasaki` **Reference Translation**: During his trip, Iwasaki `ran` into trouble on many occasions. **Model Output**: $\mathrm{P}[\texttt{ran} \mid \texttt{input}] = 0.567$ **Score**: `scores["kassara"] += 0.567`	**Input**: Translate the following text into Iloko. English: `"We now have 4-month-old mice that are non-diabetic that used to be diabetic," he added.` Iloko: `"Addaan kami ti 4-a-bulan a` **Reference Translation**: "Addaan kami ti 4-a-bulan a **babbao** a dati ket diabetic ngem saan itan," nainayonna. **Model Output**: $\mathrm{P}[\texttt{babbao} \mid \texttt{input}] = 0.351$ **Score**: `scores["babbao"] += 0.351`

Bag-of-Words Machine Translation	**Input**: Translate into English: `Los trabalhadors devon sovent obtenir l'aprobacion de sos superiors`.
	Reference Translation: `Workers must often get their superiors' approval`
	Model Output: `Workers often need to obtain their superiors' approval`
	Score:
	`scores["los"] += 1`
	`scores["trabalhadors"] += 1`
	`scores["devon"] += 0`
	`scores["sovent"] += 1`
	`scores["obtenir"] += 1`
	`scores["l'aprobacion"] += 1`
	`scores["de"] += 0`
	`scores["sos"] += 1`
	`scores["superiors"] += 1`
		`Workplace harmoney is crucial`
		Reference Translation: Ukusebenza ngokubambisana endaweni yokusebenzela kubalulekile
		Model Output: `Ukuzwana endaweni yokusebenza kubalulekile`
		`scores["ukusebenza"] += 0`
		`scores["ngokubambisana"] += 0`
		`scores["endaweni"] += 1`
		`scores["yokusebenzela"] += 1`
		`scores["kubalulekile"] += 1`

Table 1: Example task prompts, model outputs, and vocabulary-based scores. These scores are aggregated over target language words as per § 3. For the `X` → `model` direction, English output words are aligned to input language words for scoring; alignment is shown via coloring.

	Vocabulary Size		Total Word Count		Number of Languages	
Task	X→model	model→X	X→model	model→X	X→model	model→X
WT	4.8K ± 39K	4.8K ± 39K	4.8K ± 39K	9.4K ± 74K	2746	2746
WTWC	2.4K ± 4.8K	8.2K ± 18K	410K ± 630K	9700K ± 19000K	525	525
TCLM	7.4K ± 11K	6.8K ± 1.6K	90K ± 140K	21K ± 5.4K	211	211
BOW MT	7.4K ± 11K	6.8K ± 1.6K	90K ± 140K	21K ± 5.4K	211	211

Table 2: Vocabulary size: Mean and standard deviation of the number of unique words over languages per subtask. Total word count: Mean and standard deviation of total word count per language, relevant for tasks where a single word can be tested in multiple contexts. Vocabulary size and total word count are expressed in the thousands (K). Large standard deviations are caused by HRL outliers.

ability to generate the word when translating *from English* as a proxy for its generation capability for that word (model→X).

We design 8 subtasks: 4 tasks in two directions each (X→model, model→X), aimed at examining various facets of lexical capabilities in LLMs, and described in detail below. For all subtasks, we calculate our metrics over target language words (i.e. not English words $w^E_{(i)}$). More specifically, we assign the i^{th} word $w^X_{(i)}$ in the target language X a score $s(w^X_{(i)}) \in [0, 1]$. We calculate aggregate scores for language $L_{(\lambda)}$ over its vocabulary and for a model $M_{(\kappa)}$ over languages:

$$s(L_{(\lambda)}) = \left(\frac{1}{|V|} \sum_{i=1}^{|V|} s(w^X_{(i)}) \right) \times 100\%$$

$$s(M_{(\kappa)}) = \frac{1}{|L|} \sum_{\lambda=1}^{|L|} s(L_{(\lambda)})$$

We describe below how word scores $s(w^X_{(i)})$ are calculated in each of our 8 subtasks. Table 1 displays example inputs, outputs, and associated scores for each subtask. We also provide more examples per task in Appendix E. We list dataset sizes and number of supported languages for each subtask in Table 2.

3.1 Word Translation

In this task, we directly prompt a model to translate an input word either into or out of English for every term within a bilingual lexicon.

3.1.1 Scoring

For a given model output, we check for equivalence against all translation equivalents of the source word from our lexicon Ξ, using $\Xi(w_{(i)})$ to refer to the set of equivalents of $w_{(i)}$. Note that requiring answers to be an exact match to lexicon translations is unfairly strict, as the model may output a different morphological form of the correct equivalent or extraneous text around the correct answer. Given these considerations, we use additional string-matching heuristics, such as `inflection` and `substring`, among others, to determine if the model output is equivalent to the reference. We also check for `synonymy` using the English WordNet (Miller, 1994) in the X→model direction. See Appendix E for more examples and an analysis of the false positive and negative rates of these heuristics across tasks.

X→model Given a word to translate $w^X_{(i)}$ and the model prediction $\hat{w}^E_{(i)}$, we compute the binary correctness variable $\alpha^{\text{WT}}_{\text{X→model}}(w^X_{(i)})$,

$$\begin{aligned}\alpha^{\text{WT}}_{\text{X→model}}(w^X_{(i)}) = \texttt{exact\_match}(\hat{w}^E_{(i)}, \Xi(w^X_{(i)})) \\ \vee \texttt{inflection}(\hat{w}^E_{(i)}, \Xi(w^X_{(i)})) \\ \vee \texttt{substring}(\hat{w}^E_{(i)}, \Xi(w^X_{(i)})) \\ \vee \texttt{inflection\_in\_substring}(\hat{w}^E_{(i)}, \Xi(w^X_{(i)})) \\ \vee \texttt{synonym}(\hat{w}^E_{(i)}, \Xi(w^X_{(i)}))\end{aligned}$$

$$s^{\text{WT}}_{\text{X→model}}(w^X_{(i)}) = \alpha^{\text{WT}}_{\text{X→model}}(w^X_{(i)}) \in \{0, 1\}$$

where $\texttt{exact\_match}(\hat{w}^E_{(i)}, \Xi(w^X_{(i)})) = 1$ if $\hat{w}^E_{(i)}$ matches with *any* of the references in $\Xi(w^X_{(i)})$ (analogously for other heuristics).

model→X Given an English word $w^E_{(i)}$ and model prediction $\hat{w}^X_{(i)}$, we calculate binary accuracy for $w^E_{(i)}$ analogously to above, without considering synonymy as we lack WordNets in our target LRLs. Recall that model scores are computed in terms of target language vocabulary, not English words.

Suppose the word $w^X_{(m)}$ has $K = |\Xi(w^X_{(m)})|$ English translations. We define $s^{\text{WT}}_{\texttt{model}\to\texttt{X}}(w^X_{(m)}) \in [0,1]$ as

$$s^{\text{WT}}_{\texttt{model}\to\texttt{X}}(w^X_{(m)}) = \frac{1}{K} \sum_{w^E_{(i)} \in \Xi(w^X_{(m)})} \alpha^{\text{WT}}_{\texttt{model}\to\texttt{X}}(w^E_{(i)})$$

3.2 Word Translation with Context

Although a model may not understand or produce a word in isolation, it may do so given its natural context. In this task, we provide additional context for the source-side language word in the form of a sentence containing it, and then prompt the LLM to perform word translation.

This task requires monolingual data in the target language and in English for the X→model and model→X directions respectively. We evaluate on all words in the available monolingual data that also have an entry in our bilingual lexicon. Note that the number of evaluated words may therefore differ by direction.

3.2.1 Scoring

The word to be translated $w_{(i)}$ may appear in several sentences. We define $C(w_{(i)})$ to be the number of times a word appears and $w_{(i,r)}$ to be the rth occurrence of word $w_{(i)}$.

X→model We compute $\alpha^{\text{WTWC}}_{\texttt{X}\to\texttt{model}}(w^X_{(i,r)})$ for a single occurrence similarly to $\alpha^{\text{WT}}_{\texttt{X}\to\texttt{model}}(w^X_{(i)})$. We then average over occurrences to compute:

$$s^{\text{WTWC}}_{\texttt{X}\to\texttt{model}}(w^X_{(i)}) = \frac{1}{C(w^X_{(i)})} \sum_{r=1}^{C(w^X_{(i)})} \alpha^{\text{WTWC}}_{\texttt{X}\to\texttt{model}}(w^X_{(i,r)})$$

model→X We evaluate WTWC model→X similarly to WT model→X with

$$\alpha^{\text{WTWC}}_{\texttt{model}\to\texttt{X}}(w^E_{(i,r)}) = \alpha^{\text{WT}}_{\texttt{model}\to\texttt{X}}(w^E_{(i)}) \in \{0,1\}$$

To account for $C(w^E_{(i)})$ occurrences of $w^E_{(i)}$, we compute:

$$\beta^{\text{WTWC}}_{\texttt{model}\to\texttt{X}}(w^E_{(i)}) = \frac{1}{C(w^E_{(i)})} \sum_{r=1}^{C(w^E_{(i)})} \alpha^{\text{WTWC}}_{\texttt{model}\to\texttt{X}}(w^E_{(i,r)})$$

$$s^{\text{WTWC}}_{\texttt{model}\to\texttt{X}}(w^X_{(m)}) = \frac{1}{K} \sum_{w^E_{(i)} \in \Xi(w^X_{(m)})} \beta^{\text{WTWC}}_{\texttt{model}\to\texttt{X}}(w^E_{(i)})$$

3.3 Translation-Conditioned Language Modeling

WT and WTWC prompt the model directly to comprehend or generate a word and utilize a binary accuracy metric for a given output. In TCLM, we design a soft measure of the model's capability to do so given a sentence-level translation task. We utilize parallel sentence pairs t^X-t^E in target language X and English, respectively. Given the entire source sentence and a partial translation up to the word of interest, we observe the generation probability of the correct word.

Because this task deals with generation probabilities rather than observed outputs, we caution that the scores reported in each evaluation direction may not directly correspond to observed behavior. It may also not be comparable across models, as different models may have different generation distribution shapes. Similar to perplexity, this metric may be more useful in comparing various checkpoints of a single model.

3.3.1 Scoring

model→X We define the word of interest $w^X_{(i,r)}$ that appears at index n in sentence t^X. We provide the model with the complete sentence t^E as well as the left context of $w^X_{(i,r)}$, denoted as $t^X_{<n}$. In $\alpha^{\text{TCLM}}_{\texttt{model}\to\texttt{X}}(w^X_{(i,r)}) \in [0,1]$, we observe the generation probability of $w^X_{(i,r)}$:

$$\alpha^{\text{TCLM}}_{\texttt{model}\to\texttt{X}}(w^X_{(i,r)}) = P(w^X_{(i,r)} | t^E, t^X_{<n})$$

$$s^{\text{TCLM}}_{\texttt{model}\to\texttt{X}}(w^X_{(i)}) = \frac{1}{C(w^X_{(i)})} \sum_{r=1}^{C(w^X_{(i)})} \alpha^{\text{TCLM}}_{\texttt{model}\to\texttt{X}}(w^X_{(i,r)})$$

Intuitively, this is a language-modeling-like task; however, pure language modeling without the source-side English sentence to guide the model has a higher entropy at every word, since the model may choose to continue with different concepts (not necessarily $w^X_{(i)}$) as reasonable continuations. We use the sentence translation task to constrain the semantic scope of what the model might generate, thereby measuring the model's ability to generate a word broadly conditioned on its underlying concept.[4] Note that this evaluation does not require bilingual lexicons.

[4]We note that observing generation probabilities in this way is not a perfect measure of this ability. While the model knows the sentence-level semantics of the target language text as well as the left context up to the word of interest, it may still choose a different continuing formulation of the target-side sentence, leading to an unfairly low score.

X→model We now have t^E on the output side, and are interested in evaluating the comprehension of various words $w^X_{(m)}$ in the source-side t^X sentence. For every $w^E_{(i,r)}$ occurring at index n of sentence t^E, we calculate

$$\alpha^{\text{TCLM}}_{\texttt{X}\to\texttt{model}}(w^E_{(i,r)}) = P(w^E_{(i,r)}|t^X, t^E_{<n}) \in [0,1]$$

The intuition is similar to the model→X case: we are interested in evaluating the model's ability to comprehend a word in a natural setting, and use the generation probability of its English equivalent given a restricted semantic scope. However, recall once again that ChiKhaPo scores are computed in terms of the vocabulary of the *target language* X, not English. We therefore have the additional problem of finding the language X word in t^X that maps to or "produced" $w^E_{(i,r)}$. We use our existing lexicons in conjunction with statistical alignments with `FastAlign` (Dyer et al., 2013) to identify this mapping. We define an alignment as $\mathcal{A}(w^X_{(m)}) = \{w^E_{(i,r)}\}$ where $\mathcal{A}$ denotes alignments for sentence t^X-t^E. We define $\mathcal{F}$ as a union of $\Xi(w^X_{(m)})$ and $\mathcal{A}(w^X_{(m)})$, prioritizing the former. For every $w^X_{(m)} \in t^X$, we calculate:

$$\beta^{\text{TCLM}}_{\texttt{X}\to\texttt{model}}(w^E_{(i)}) = \frac{1}{C(w^E_{(i)})} \sum_{r=1}^{C(w^X_{(i)})} \alpha^{\text{TCLM}}_{\texttt{X}\to\texttt{model}}(w^E_{(i,r)})$$

$$s^{\text{TCLM}}_{\texttt{X}\to\texttt{model}}(w^X_{(m)}) = \frac{1}{|\mathcal{F}|} \sum_{w^E_{(i)}\in\mathcal{F}} \beta^{\text{TCLM}}_{\texttt{X}\to\texttt{model}}(w^E_{(i)})$$

3.4 Bag-of-Words Machine Translation

Given a sequence-level machine translation task, metrics such as BLEU (Papineni et al., 2002) and CHRF (Popović, 2015) measure translation quality by assessing the exact match n-gram or character-gram overlap between model outputs and reference translations. Given our lexical focus, we instead formulate a coarser evaluation metric. Given a sentence-level MT task, we are interested in evaluating whether the target language words were correctly produced (model→X) or translated correctly to English equivalents (X→model), regardless of the syntax of the output or the appropriateness of the morphological form of the word.

3.4.1 Scoring

model→X Given a parallel sentence pair t^X-t^E, we prompt $M_{(\kappa)}$ to translate t^E to target language X. For every $w^X_{(i)} \in t^X$, we check whether the predicted sentence $\hat{t}^X$ contains $w^X_{(i)}$. We calculate:

$$\alpha^{\text{BOW MT}}_{\texttt{model}\to\texttt{X}}(w^X_{(i,r)}) = \texttt{exact\_match}(\hat{t}^X, w^X_{(i)}) \vee \texttt{inflection}(\hat{t}^X, w^X_{(i)})$$

$$s^{\text{BOW MT}}_{\texttt{model}\to\texttt{X}}(w^X_{(i)}) = \frac{1}{C(w^X_{(i)})} \sum_{r=1}^{C(w^X_{(i)})} \alpha^{\text{BOW MT}}_{\texttt{model}\to\texttt{X}}(w^X_{(i,r)})$$

X→model Given t^X-t^E, we prompt $M_{(\kappa)}$ to translate t^X into English. We check whether the predicted sentence $\hat{t}^E$ contains $w^E_{(i)} \in t^E$.

$$\alpha^{\text{BOW MT}}_{\texttt{X}\to\texttt{model}}(w^E_{(i,r)}) = \texttt{exact\_match}(\hat{t}^E, w^E_{(i)}) \vee \texttt{inflection}(\hat{t}^E, w^E_{(i)}) \vee \texttt{synonym}(\hat{t}^E, w^E_{(i)})$$

Similarly as in TCLM, we generate the English alignments $\mathcal{F}$ for $w^X_{(m)}$ and compute its score:

$$\beta^{\text{BOW MT}}_{\texttt{X}\to\texttt{model}}(w^E_{(i)}) = \frac{1}{C(w^E_{(i)})} \sum_{r=1}^{C(w^E_{(i)})} \alpha^{\text{BOW MT}}_{\texttt{X}\to\texttt{model}}(w^E_{(i,r)})$$

$$s^{\text{BOW MT}}_{\texttt{X}\to\texttt{model}}(w^X_{(m)}) = \frac{1}{|\mathcal{F}|} \sum_{w^E_{(i)}\in\mathcal{F}} \beta^{\text{BOW MT}}_{\texttt{X}\to\texttt{model}}(w^E_{(i)})$$

4 Data Sources and Languages

Task	X→model	model→X
WT	lexicons	lexicons
WTWC	lexicons, monolingual datasets	lexicons, monolingual datasets
TCLM	lexicons, bitext	bitext
BOW MT	lexicons, bitext	bitext

Table 3: Data type required in each task

Data sources Table 3 lists the types of data required for each task, as per the task description above. We use **lexicons** created by amalgamating GATITOS (Jones et al., 2023), Intercontinental Dictionary Series (Bibiko, 2023), and PanLex (Kamholz et al., 2014) data. For a given word, we used translations from the first two if available, and fallback to PanLex. See Appendix D for language coverage of lexicons. We use **monolingual data** from GLOTLID (Kargaran et al., 2023) which covers 1665 languages, and **parallel data** from FLORES+ (NLLB Team et al., 2024), which

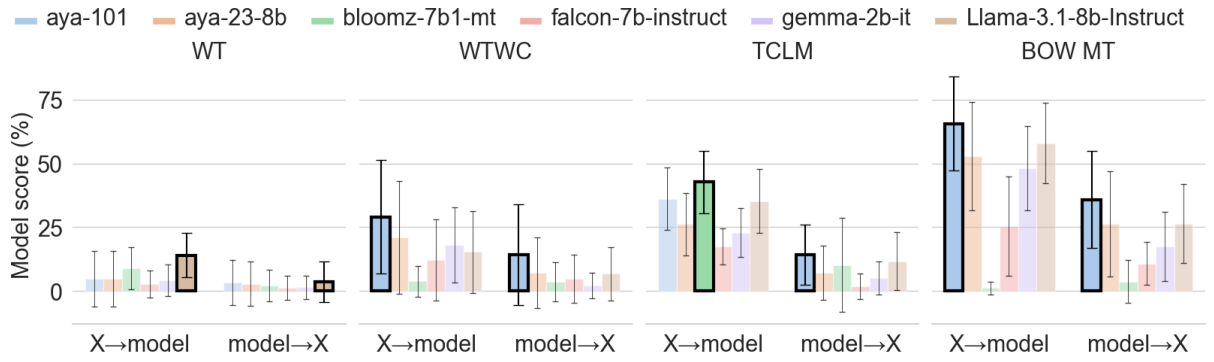

Figure 2: Model scores across subtasks, with std. deviation over languages. Best performing model is highlighted.

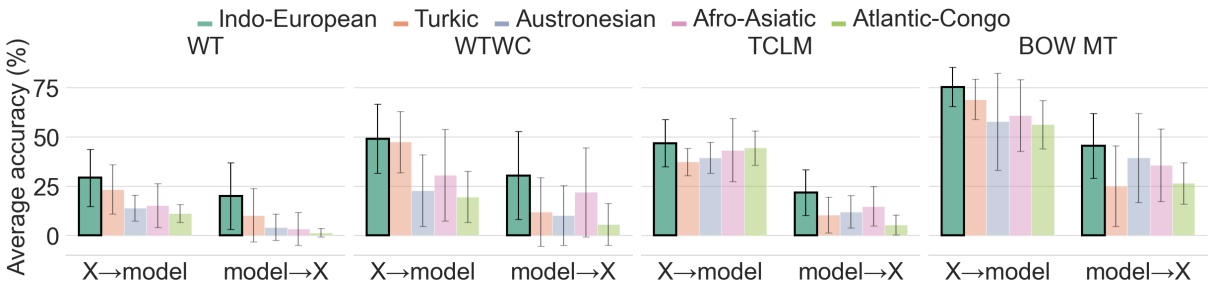

Figure 3: We compute the score of a language family as the average of its constituent languages, with the best-performing language family highlighted. Error bars represent the standard deviation within the language family. The Indo-European family has consistently higher scores than other families.

covers 212 languages. We discard languages with fewer than 100 entries in the target language lexicon. In WTWC, TCLM, and BOW MT we discard languages where our lexicons cover less than 100 unique words from monolingual or parallel data.

Languages See Appendix B for details concerning the distribution of languages over language families as covered by each task, geographic spread, and code conventions used.

5 Experimental Setup

We evaluated six multilingual open-source models: `aya-101` (Üstün et al., 2024), `aya-23-8b` (Aryabumi et al., 2024), `bloomz-7b1-mt` (Muennighoff et al., 2023), `falcon-7b-instruct` (Almazrouei et al., 2023), `gemma-2b-it` (Team et al., 2024), and `Llama-3.1-8B-Instruct` (Grattafiori et al., 2024). We list key characteristics of these models in Table 24. See Appendix F for prompts used per subtask and Appendix I for details on GPU hours required to run each subtask.

Given the number of languages and size of dataset, we present a `lite` version of the benchmark, which uses a subset of the data as follows. We cap the number of vocabulary entries per language for WT and WTWC `X→model` at 300. We use 30% of the available data for TCLM and BOW MT. All reported language scores are computed over a minimum of 100 words per language.

6 Results and Discussion

See the performance of tested models on all 8 subtasks in Figure 2. See detailed results in Appendix G, including the language score distribution per subtask and model as well as sampled language scores. Broadly, we observe that models have significant room for improvement; i.e. **our benchmark is a challenging measure of multilingual performance**.

We train a decision tree to predict language scores per task based on a series of features, including model, language resourcedness, script, language family, and others. We find the top features that determine task performance for a given language are evaluation direction, whether the language is supported by the model, and resource level of the language (see § G.1 for decision trees and ranked feature importances). We discuss these features in more detail below.

Evaluation direction Models evaluated in the X→model direction exhibit higher scores than in the model→X direction, i.e. even if a model can comprehend a word in an LRL, it might not be capable of generating it. This finding is consistent with previous literature that finds a considerable gap between NLU and NLG, or the out-of-X direction and the into-X direction in MT (Belinkov et al., 2017; Kandimalla et al., 2022).

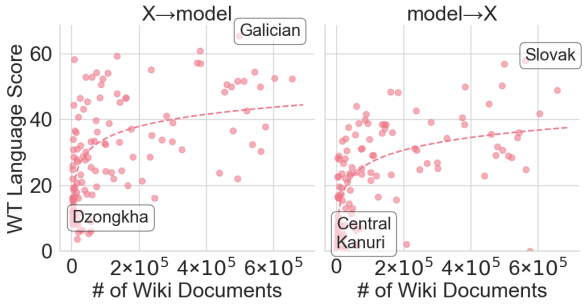

Figure 4: Comparison of the number of Wikipedia documents—a proxy for resource level—and language performance for the task WT. See § G.4 for other tasks.

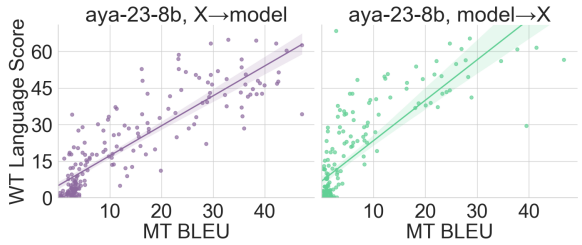

Figure 5: WT scores are strongly correlated with sentence-level MT BLEU scores.

Language family and resource In Figure 3, we draw attention to the performance gap between Indo-European languages and underrepresented Austronesian and Atlantic-Congo languages.

Naturally, there is a lot of variation between model performance on languages within a single family, depending on other potential factors such as the resourcedness of the language and whether it is supported by the model. In Figure 4, we show the relationship between resource level and WT performance. This is roughly logarithmic, with the bulk of LRLs performing significantly worse than HRLs, large improvements for mid-resource languages, and gains saturating for HRLs. In sum, **we highlight the scope of improvement for SOTA models on underrepresented language families and low-resource languages**.

Model aya-101 achieves the highest average score on five of the eight subtasks. Compared to other models, aya-101 is unique in that it employs an encoder-decoder architecture, is larger (13B parameters), and instruction-tuned on 101 languages. (See Table 24). These qualities may contribute to its performance.

Task Note that while WT, WTWC, and BOW MT all report accuracy metrics over a vocabulary set, model scores are not directly comparable across tasks as they are computed over different vocabularies, as per resource requirements for each task. That being said, we observe generally higher scores for WTWC than WT in Figure 2. This indicates that models are able to utilize and benefit from the additional context provided in the former.

We also see that models generally show higher scores for BOW MT than for WT and WTWC. BOW MT uses a sentence-level machine translation setup, which instruction-tuned models may be more familiar with as opposed to direct prompts concerning word meaning as used in WT and WTWC. BOW MT also allows the model to generate the previous context of the word of interest in the output translation, potentially priming the model better in terms of semantic context as well as language of generation.

As discussed in § 3, TCLM is less directly interpretable and comparable across models than the other tasks and is better employed during model development. **By including subtasks of different difficulties and settings, our benchmark allows for various perspectives and a nuanced understanding of lexical competence.**

Correlation with MT While machine translation is a good measure of natural language understanding (Iyer et al., 2023), sentence-level translation datasets are expensive to create and curate. In Figure 5, we demonstrate that there is a strong linear correlation between BLEU scores on machine translation performance with FLORES+ and scores from WT, for available languages in FLORES+ (0.873 and 0.769 in the X→model and model→X evaluation direction respectively). Given that WT covers 2700+ languages as opposed to the 212 covered by FLORES+, **our benchmark can provide a cheap proxy in the absence of machine translation data**.

7 Conclusion

We introduce ChiKhaPo, a massively multilingual benchmark testing lexical competence, that draws on existing available resources such as lexicons, monolingual data, and bitext. ChiKhaPo consists of 8 subtasks that provide various perspectives on lexical comprehension and generation skills. We evaluate SOTA models on our benchmark and find that these have a long way to go for low-resource languages. With this work, we hope to promote the massively multilingual evaluation of LLMs as one step towards addressing language inequity in NLP.

8 Limitations

Benchmark coverage and quality While ChiKhaPo covers 2700+ languages, this is restricted to the relatively simple word translation task; other tasks have more typical coverage compared to existing benchmarks. Further, ChiKhaPo is heavily reliant on public lexicons, specifically PanLex, which is known to be noisy for many languages. We note that while ChiKhaPo currently relies heavily on PanLex, GLOTLID, and FLORES+, it can integrate other lexicons and monolingual and parallel datasets in the future. Notably, the BOUQuET dataset (Andrews et al., 2025; Team et al., 2026), covering 275 languages, was released after the completion of this work, and constitutes an additional high-quality source of parallel data.

Lexical coverage, sense disambiguation, and synonymy Our benchmark is limited by the available annotations in the lexicons we work with. This results in a number of shortcomings and avenues for future improvement. Several languages may only have a few hundred entries in available lexicons. Further, models may output valid variants or synonyms that are not documented in our lexicons, potentially resulting in false negatives in WT. Our lexicons also do not annotate word sense. This limitation may become problematic, e.g. in WTWC where only a particular word sense should be marked correct given a sentence.

Morphological, syntactic, and complex semantic skills are out of scope. Our benchmark focuses on evaluating lexical understanding in models. However, basic skills in a language also include understanding and producing appropriate morphological forms and appropriate word orders for utterances. Although these are important dimensions of the evaluation, we currently lack resources in the target languages to evaluate these skills in our benchmark.

We hope that our experiments and benchmark motivate the further collection and refinement of lexicons, as well as other such resources in low-resource languages. In doing so, ChiKhaPo can enable richer evaluations of the basic linguistic skills of LLMs on a massively multilingual scale.

Ethics Statement

We do not expect any negative ethical consequences of this work, which presents a benchmark for the multilingual evaluation of large language models. We use publicly available datasets to design our benchmark, and provide results on open-source models. Our benchmark and Python package are in accordance with the licenses of each constituent dataset (see Appendix C) and include functionalities for downloading the data as well as running evaluations. This work used LLMs for coding assistance only.

9 Acknowledgments

We would like to thank Drs. David Yarowsky and Karen Livescu for helpful discussions and feedback on this paper. We also thank the anonymous reviewers for their feedback.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, and 1 others. 2023. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models. *Preprint*, arXiv:2311.16867.

Sotiris Anagnostidis and Jannis Bulian. 2024. How susceptible are llms to influence in prompts? *Preprint*, arXiv:2408.11865.

Pierre Andrews, Mikel Artetxe, Mariano Coria Meglioli, Marta R. Costa-jussà, Joe Chuang, David Dale, Mark Duppenthaler, Nathanial Paul Ekberg, Cynthia Gao, Daniel Edward Licht, Jean Maillard, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Eduardo Sánchez, Ioannis Tsiamas, Arina Turkatenko, Albert Ventayol-Boada, and Shireen Yates. 2025. BOUQuET : dataset, benchmark and open initiative for universal quality evaluation in translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27515–27535, Suzhou, China. Association for Computational Linguistics.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, and 2 others. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

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

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? *arXiv preprint arXiv:1704.03471*.

Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, and 1 others. 2025. Llama-nemotron: Efficient reasoning models. *arXiv preprint arXiv:2505.00949*.

Hans-Jörg Bibiko. 2023. Cldf dataset derived from key and comrie's "intercontinental dictionary series" from 2023.

Mingyang Chen, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z. Pan, Wen Zhang, Huajun Chen, Fan Yang, Zenan Zhou, and Weipeng Chen. 2025. Research: Learning to reason with search for llms via reinforcement learning. *Preprint*, arXiv:2503.19470.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *Preprint*, arXiv:1809.05053.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm. Accessed: 2023-06-30.

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, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 644–648.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten,

Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2025. Glottolog 5.2. Available online at http://glottolog.org, Accessed on 2025-09-25.

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Vivek Iyer, Pinzhen Chen, and Alexandra Birch. 2023. Towards effective disambiguation for machine translation with large language models. *arXiv preprint arXiv:2309.11668*.

Alexander Jones, Isaac Caswell, Orhan Firat, and Ishank Saxena. 2023. GATITOS: Using a new multilingual lexicon for low-resource machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 371–405, Singapore. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

David Kamholz, Jonathan Pool, and Susan Colowick. 2014. PanLex: Building a resource for panlingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3145–3150, Reykjavik, Iceland. European Language Resources Association (ELRA).

Akshara Kandimalla, Pintu Lohar, Souvik Kumar Maji, and Andy Way. 2022. Improving english-to-indian language neural machine translation systems. *Information*, 13(5):245.

Amir Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. Glotlid: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, page 6155–6218. Association for Computational Linguistics.

Gerhard Kremer, Katrin Erk, Sebastian Padó, and Stefan Thater. 2014. What substitutes tell us - analysis of an "all-words" lexical substitution corpus. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 540–549, Gothenburg, Sweden. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Mina Lee, Chris Donahue, Robin Jia, Alexander Iyabor, and Percy Liang. 2021. Swords: A benchmark for lexical substitution with improved data coverage and quality. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4362–4379, Online. Association for Computational Linguistics.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, and 31 others. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, and 2 others. 2021. Few-shot learning with multilingual language models. *CoRR*, abs/2112.10668.

Yile Liu, Ziwei Ma, Xiu Jiang, Jinglu Hu, Jing Chang, and Liang Li. 2025. Maxife: Multilingual and cross-lingual instruction following evaluation. *arXiv preprint arXiv:2506.01776*.

Gonzalo Martínez, Javier Conde, Elena Merino-Gómez, Beatriz Bermúdez-Margaretto, José Alberto Hernández, Pedro Reviriego, and Marc Brysbaert. 2024. Establishing vocabulary tests as a benchmark for evaluating large language models. *PLOS ONE*, 19(12):1–17.

Diana McCarthy. 2002. Lexical substitution as a task for WSD evaluation. In *Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 089–115. Association for Computational Linguistics.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 task 10: English lexical substitution task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic. Association for Computational Linguistics.

Rada Mihalcea, Ravi Sinha, and Diana McCarthy. 2010. SemEval-2010 task 2: Cross-lingual lexical substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 9–14, Uppsala, Sweden. Association for Computational Linguistics.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

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. 2023. Crosslingual generalization through multitask finetuning. *Preprint*, arXiv:2211.01786.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Bolette Pedersen, Nathalie Sørensen, Sussi Olsen, Sanni Nimb, and Simon Gray. 2024. Towards a Danish semantic reasoning benchmark - compiled from lexical-semantic resources for assessing selected language understanding capabilities of large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16353–16363, Torino, Italia. ELRA and ICCL.

Edoardo M. Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. *arXiv preprint*.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

James Pustejovsky. 2016. *Lexical semanics*, page 33–64. Cambridge Handbooks in Language and Linguistics. Cambridge University Press.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2019. The MuCoW test suite at WMT 2019: Automatically harvested multilingual contrastive word sense disambiguation test sets for machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 470–480, Florence, Italy. Association for Computational Linguistics.

Annette Rios, Mathias Müller, and Rico Sennrich. 2018. The word sense disambiguation test suite at WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 588–596, Belgium, Brussels. Association for Computational Linguistics.

Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. 2017. Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19, Copenhagen, Denmark. Association for Computational Linguistics.

Sebastian Ruder. 2021. Challenges and opportunities in nlp benchmarking.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, and 14 others. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *Preprint*, arXiv:2402.06619.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adri Garriga-Alonso, and 1 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

The Omnilingual MT Team, Belen Alastruey, Niyati Bafna, Andrea Caciolai, Kevin Heffernan, Artyom Kozhevnikov, Christophe Ropers, Eduardo Sánchez, Charles-Eric Saint-James, Ioannis Tsiamas, Chierh Cheng, Joe Chuang, Paul-Ambroise Duquenne, Mark Duppenthaler, Nate Ekberg, Cynthia Gao, Pere Lluís Huguet Cabot, João Maria Janeiro, Jean Maillard, and 12 others. 2026. Omnilingual MT: Machine translation for 1,600 languages.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. *Preprint*, arXiv:2402.10588.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. Paws-x: A cross-lingual adversarial dataset for paraphrase identification. *arXiv preprint arXiv:1908.11828*.

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Preprint*, arXiv:2306.05179.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. 2024. ∞bench: Extending long context evaluation beyond 100k tokens. *Preprint*, arXiv:2402.13718.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin 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. 2024. Aya model: An instruction finetuned open-access multilingual language model. *Preprint*, arXiv:2402.07827.

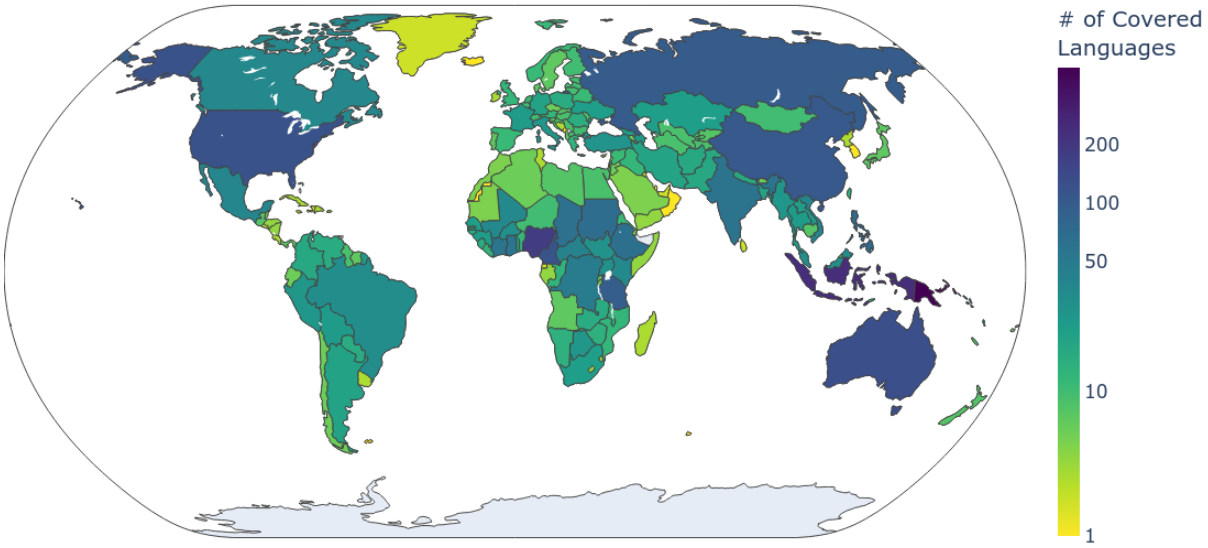

Figure 6: Drawing on Glottolog data (Hammarström et al., 2025), the choropleth map above illustrates the geographic distribution of the languages covered in at least one task in ChiKhaPo. Specifically, we note the country of origin. Countries with the highest number of languages include: Papua New Guinea where 498 languages originate, Indonesia 240, and Nigeria 182.

A Prior Work

Table 4 lists SOTA multilingual benchmarks as well as past work performed in lexical substitution, a task that is closest with our current work. The listed benchmarks exhibit limited language coverage.

B Languages

B.1 Geographic Spread

Figure 6 notes the country of origin of languages represented in at least one task. In observing the geographic spread of this task, we see that we attain coverage in all world countries.

B.2 Language Families

In Table 5, Table 6, and Table 7, we report the number of languages in Glottolog families in each of the four tasks.

For all tasks, Indo-European languages are well-represented. Nuclear Trans New Guinea languages are well-represented in WT, while Atlantic-Congo languages are well-represented in WT and WTWC.

B.3 Language code conventions

In WT, we represent each translation by its ISO code, regardless of the translation's script or geographic origin. For example, Achinese may be written in Arabic or Latin script. However, this distinction in script is not made in WT as PanLex —a major data source—classifies word translations only by their ISO code. Consequently, translations to and from the language Achinese falls under the ISO code ace.

The data sources of WTWC, TCLM, and BOW MT differentiate languages by script. For example, Achinese in Arabic script is evaluated separately from Achinese in Latin script. We adopt this distinction by script for these three tasks.

Benchmark	Task	No. of Langages
SOTA Benchmarks		
FLORES-200 (NLLB Team et al., 2024)	Translation	212
BELEBELE (Bandarkar et al., 2024)	Reading Comprehension	122
Aya Evaluation Suite (Singh et al., 2024)	Instruction Following	101
MEGA (Ahuja et al., 2023)	Generation, Classification	70
XL-Sum (Hasan et al., 2021)	Summarization	43
MaXIFE (Liu et al., 2025)	Instruction Following	23
Aya Expanse, m-Arena Hard (Dang et al., 2024)	Instruction Following	23
WikiLingua (Ladhak et al., 2020)	Summarization	18
MMMLU (Hendrycks et al., 2020)	Reasoning	14
XNLI (Conneau et al., 2018)	Inference	14
XCOPA (Ponti et al., 2020)	Classification	11
XStoryCloze (Lin et al., 2021)	Reasoning	11
TyDiQA (Clark et al., 2020)	Question Answering	11
GSM8K (Cobbe et al., 2021)	Mathematical Reasoning	10
M3Exam (Zhang et al., 2023)	Question Answering	9
PAWS-X (Yang et al., 2019)	Paraphrase Identification	6
MLQA (Lewis et al., 2020)	Question Answering	7
XWinograd (Muennighoff et al., 2023)	Coreference Resolution	6
Dolly (Conover et al., 2023)	Instruction Following	3
∞Bench (Zhang et al., 2024)	Long Context Reasoning	2
Lexical Understanding		
MuCoW (Raganato et al., 2019)	Lexical Substitution	12
ContraWSD (Rios Gonzales et al., 2017)	Lexical Substitution	3
Cross-lingual Lexical Substitution Task (Mihalcea et al., 2010)	Lexical Substitution	2
TOEFL, StuVoc, LexTale (Martínez et al., 2024)	Lexical Substitution	2
Word Sense Disambiguation Test Suite (Rios et al., 2018)	Lexical Substitution	2
Danish Semantic Reasoning Benchmark (Pedersen et al., 2024)	Lexical Substitution	1
ChiKhaPo	**Lexical Comprehension and Generation**	**2746**

Table 4: Language coverage across text benchmarks that evaluate multilingual NLU and NLG capabilities.

Language Family	WT	WTWC	TCLM	BOW MT
Atlantic-Congo	483	85	33	33
Austronesian	483	103	21	21
Nuclear Trans New Guinea	225	15	0	0
Indo-European	184	123	73	73
Afro-Asiatic	134	21	19	19
Pama-Nyungan	71	1	0	0
Tai-Kadai	38	2	3	3
Sino-Tibetan	38	11	9	9
Mande	32	3	2	2
Nakh-Daghestanian	29	10	0	0
Uralic	28	18	4	4
Nuclear Torricelli	27	0	0	0
Sepik	27	2	0	0
Austroasiatic	26	5	3	3
Athabaskan-Eyak-Tlingit	26	4	0	0
Turkic	25	27	14	14
Artificial Language	18	10	2	2
Central Sudanic	16	0	0	0
Quechuan	16	11	1	1
Uto-Aztecan	16	2	0	0
Dogon	16	0	0	0
Timor-Alor-Pantar	16	1	0	0
Nilotic	15	6	3	3
Algic	15	1	0	0
Ta-Ne-Omotic	14	0	0	0
Hmong-Mien	14	1	0	0
Otomanguean	13	1	0	0
Kru	13	0	0	0
Angan	12	0	0	0
Arawakan	11	1	0	0
Khoe-Kwadi	10	1	0	0
Dravidian	10	1	4	4
Pano-Tacanan	10	1	0	0
Surmic	10	0	0	0
Heibanic	10	0	0	0
Nyulnyulan	9	0	0	0
Anim	9	0	0	0
Mayan	9	5	0	0
Gunwinyguan	8	0	0	0
Tupian	8	3	1	1
Yam	8	0	0	0
Dagan	8	1	0	0
Cariban	8	3	0	0
Ramu	8	1	0	0
South Bird's Head	7	0	0	0
Nubian	7	0	0	0
Bosavi	7	0	0	0
Pomoan	7	0	0	0
Kadugli-Krongo	6	0	0	0
Mailuan	6	0	0	0
Ndu	6	0	0	0
Saharan	6	0	2	2
Siouan	6	0	0	0
Left May	6	0	0	0
Koiarian	6	2	0	0
Japonic	6	1	1	1
Kiwaian	6	0	0	0
Tungusic	6	0	0	0
Lower Sepik	5	0	0	0
Eleman	5	1	0	0
Cochimi-Yuman	5	0	0	0
Narrow Talodi	5	0	0	0
South Bougainville	5	0	0	0
Yeniseian	5	0	0	0

Table 5: Distribution of languages in Glottolog language families across all tasks.

Language Family	WT	WTWC	TCLM	BOW MT
Muskogean	5	1	0	0
Miwok-Costanoan	5	0	0	0
Eskimo-Aleut	5	2	0	0
East Strickland	5	0	0	0
Salishan	5	0	0	0
Yareban	5	1	0	0
Mataguayan	5	1	0	0
Suki-Gogodala	4	0	0	0
Lengua-Mascoy	4	0	0	0
Eastern Trans-Fly	4	0	0	0
Kartvelian	4	2	1	1
Abkhaz-Adyge	4	3	0	0
Koman	4	0	0	0
Ijoid	4	0	0	0
Mangarrayi-Maran	4	0	0	0
Eastern Jebel	4	0	0	0
Songhay	4	2	0	0
Maban	4	0	0	0
Tuu	4	0	0	0
Iroquoian	4	1	0	0
Dajuic	4	0	0	0
Guaicuruan	4	0	0	0
Chumashan	4	0	0	0
Mirndi	4	0	0	0
North Bougainville	4	0	0	0
Tangkic	3	0	0	0
South Omotic	3	0	0	0
Kuliak	3	0	0	0
Kwalean	3	0	0	0
Kxa	3	0	0	0
Kamula-Elevala	3	0	0	0
Kolopom	3	0	0	0
Chibchan	3	1	0	0
Iwaidjan Proper	3	0	0	0
Bookkeeping	3	0	0	0
Mongolic-Khitan	3	3	1	1
West Bomberai	3	0	0	0
Chocoan	3	0	0	0
Jarrakan	3	0	0	0
Maningrida	3	1	0	0
Nuclear-Macro-Je	3	1	0	0
Dizoid	3	0	0	0
Tucanoan	3	0	0	0
Walioic	3	0	0	0
Tamaic	3	0	0	0
Konda-Yahadian	2	0	0	0
Rashad	2	0	0	0
Keram	2	0	0	0
Haida	2	0	0	0
Mixe-Zoque	2	0	0	0
Yanomamic	2	0	0	0
Bogia	2	0	0	0
Caddoan	2	0	0	0
Kunimaipan	2	0	0	0
Pahoturi	2	0	0	0
Baibai-Fas	2	0	0	0
Kayagaric	2	0	0	0
Sign Language	2	0	0	0
Katla-Tima	2	0	0	0
Yangmanic	2	0	0	0
Kresh-Aja	2	0	0	0
Piawi	2	0	0	0
Kwomtari-Nai	2	0	0	0
Arafundi	2	0	0	0

Table 6: Distribution of languages in Glottolog language families across all tasks.

Language Family	WT	WTWC	TCLM	BOW MT
Somahai	2	0	0	0
Bunaban	2	0	0	0
Kaure-Kosare	2	0	0	0
Bayono-Awbono	2	0	0	0
Giimbiyu	2	0	0	0
Bulaka River	2	0	0	0
Teberan	2	1	0	0
Mombum-Koneraw	2	0	0	0
Worrorran	2	0	0	0
Manubaran	2	0	0	0
Chonan	2	0	0	0
Barbacoan	2	0	0	0
Amto-Musan	2	0	0	0
Turama-Kikori	2	0	0	0
Maiduan	1	0	0	0
Chicham	1	0	0	0
Koreanic	1	1	1	1
Lakes Plain	1	0	0	0
Gumuz	1	0	0	0
Aymaran	1	1	1	1
Temeinic	1	0	0	0
Chukotko-Kamchatkan	1	0	0	0
Kawesqar	1	0	0	0
Huitotoan	1	0	0	0
Misumalpan	1	0	0	0
Kiowa-Tanoan	1	0	0	0
Wakashan	1	0	0	0
Arawan	1	1	0	0
Garrwan	1	0	0	0
Tarascan	1	1	0	0
Chinookan	1	0	0	0
Saliban	1	0	0	0
East Kutubu	1	0	0	0
Totonacan	1	0	0	0
Sahaptian	1	0	0	0
Zamucoan	1	0	0	0
Tsimshian	1	0	0	0
Ainu	1	0	0	0
Tequistlatecan	1	0	0	0
Great Andamanese	1	0	0	0
Peba-Yagua	1	0	0	0
Naduhup	1	0	0	0
Pidgin	1	0	0	0
Baining	1	0	0	0
Blue Nile Mao	1	0	0	0
Furan	1	0	0	0
Nyimang	1	0	0	0
Marrku-Wurrugu	1	0	0	0
Uru-Chipaya	1	0	0	0
Huavean	1	0	0	0
Mairasic	1	0	0	0
Araucanian	1	1	0	0
Shastan	1	0	0	0
North Halmahera	1	0	0	0
Tor-Orya	1	0	0	0
Chapacuran	1	0	0	0
Yuat	1	0	0	0
Taulil-Butam	1	0	0	0
Jicaquean	1	0	0	0

Table 7: Distribution of languages in Glottolog language families across all tasks.

C Licensing

The datasets used for this study are all publicly available. **FLORES+** is released under the Creative Commons Attribution-ShareAlike 4.0 International Public License. **GLOTLID** is released under Apache 2.0. **GATITOS** is released under Creative Commons Attribution 4.0. **IDS** is released under the Creative Commons license. While **PanLex** is licensed under Creative Commons CC0 1.0 Universal, PanLex draws upon numerous sources, each of which has its own copyright status. Under Creative Commons CC0 1.0 Universal, the use of PanLex for research purposes is permitted.

D Lexicons

In Table 8, we detail the number of languages covered in our three lexicon sources. PanLex covers the most.

Lexicon Source	# of Language Pairs
GATITOS	177
IDS	240
PanLex	5731

Table 8: Breakdown of languages across lexicon sources

E Classification Heuristics

E.1 Examples across tasks

See Table 9, Table 10, Table 11, and Table 12 for examples of responses that were classified as correct and incorrect. We further define the implementation of each classification heuristic.

Prompt	Model Response	Ground Truth	Classification
Correct			
Translate the following word from Papiamento to English. Respond with a single word. Word:conoci. Translation:	know	know	exact match
Translate the following text from English to Spanish: Egyptian.	Egipto.	egipcio, de egipto	inflection
Translate the following text from English to Spanish: Charm.	El encanto.	encanto	substring
Translate the following word from French to English. Respond with a single word. Word:un des quadruplés. Translation:	one of the quadruplets	quad, quadruplet	inflection in substring
Translate the following word from Dutch to English. Respond with a single word. Word:smerig. Translation:	disgusting	dirty, soiled, squalid	synonym
Incorrect			
Translate the following text from English to Yagaria: hill.	hill	ae, moa, aemo	echo
Translate the following word from English to Gane. Respond with a single word. Word: road Translation:	path	lolan	outputted in source language
Translate the following text from Djamindjung to English: mayili.	mayil. mayil. mayil. mayil. mayil.	poison	gibberish

Table 9: Examples of WT prompts and responses.

Prompt	Model Response	Ground Truth	Classification
Correct			
In “Zot dir li ankor, “Kisannla to ete? Rann enn repons pou nou al inform bann dimoun ki finn avoy nou. Twa, ki to ena pou dir lor tomem?””, the word “ankor” means ____ in English.	again	again, still	exact match
In “A Comissão tem de dar apoio a esses países mas deve também utilizá-los como exemplo para obrigar outros países a fazerem o mesmo.”, the word “países” means ____ in English.	countries	country	inflection
In “Around 11:29, the protest moved up Whitehall, past Trafalgar Square, along the Strand, passing by Aldwych and up Kingsway towards Holborn where the Conservative Party were holding their Spring Forum in the Grand Connaught Rooms hotel.”, the word “were” means ____ in Hungarian.	a) voltak	voltak	substring
Sentence: Dit moedigt men aan om zelfs veur aafstande vaan ‘n paar honderd meter de auto te nómme (aafstande die door ‘t wegenetwerk ‘n paar kilometer kenne wère). Define “auto” in one English word:	cars, trucks, motorcycles, bicycles, scooters, mopeds, motorbikes	car	inflection in substring
“In “Laylak ez zituen arrosak espero.”, the word “zituen” means ____ in English.”	they will have	had	synonym
Incorrect			
What does “in” mean in Kambera in the sentence “She didn’t trust in the LORD.”? Meaning (one word):	in	hudalu, coda, hu dalu, nu dalu	echo
In “The protest started around 11:00 local time (UTC+1) on Whitehall opposite the police-guarded entrance to Downing Street, the Prime Minister’s official residence.”, the word “time” means ____ in Hmong Daw.	The protest started around 11:00 local time (UTC+1) on Whitehall opposite the police-guarded entrance to Downing Street, the Prime Minister’s official residence.	sij hawm	outputted source sentence
In “The couple has to fill an application form and submit it along with two photographs of their wedding ceremony and an invitation card for the same.”, the word “same” means ____ in Akawaio.	similar	mari	outputted in source language
In “Christopher Garcia, a spokesperson of the Los Angeles Police Department, said the suspected male offender is being investigated for trespassing rather than vandalism.”, the word “male” means ____ in Mountain Koiali.	gender	mo, ovaite	gibberish

Table 10: Examples of WTWC prompts and response

Prompt	Next Word	Probability
Translate the sentence into English. Standard Latvian:Mums tagad ir 4 mēnešus vecas peles, kas nav diabēta slimnieces, bet kuras agrāk bija diabēta slimnieces, viņš piebilda. English: “We now have 4-month-old	mice	2.37×10^{-4}
Translate the sentence into English. Czech:Dr. Ehud Ur, profesor medicíny na Dalhousieově univerzitě v Halifaxu v Novém Skotsku a zároveň předseda klinické a vědecké divize Kanadské diabetické asociace upozornil, že výzkum je teprve ve svých počátcích. English: Dr. Ehud Ur,	professor	0.367
Translate the following text into Sundanese. English:USA Gymnastics and the USOC have the same goal — making the sport of gymnastics, and others, as safe as possible for athletes to follow their dreams in a safe, positive and empowered environment. Sundanese: Organisasi Senam Nasional AS sareng USOC gaduh tujuan anu sami - ngusahakeun olahraga	sename	0.903

Table 11: Examples of TCLM prompts. LLMs may be surprisingly brittle to punctuation and spacing, but we find that these result in minor variations of our results.

Prompt	Model Response	Ground Truth
Translate into English: Edgar Vagwilla wa li valula ko kuoko kuenda ko njembue, osimbu okuti Kristoffer Schneider wa sukila oku pelaliwa vocipala.	Edgar Vagwilla was injured in the knee to the ankle, while Kristoffer Schneider was taken to the hospital for a knee injury.	Edgar Veguilla received arm and jaw wounds while Kristoffer Schneider was left requiring reconstructive surgery for his face.
Translate into Kikuyu: Perry stated that he would “return to Texas to assess the results of tonight’s caucus, determine whether there is a path forward for myself in this race”, but later said that he would remain in the race and compete in the January 21 South Carolina primary.	Perry alitangaza kwamba angerudi Texas “kuchunguza matokeo ya usiku huu wa kampeni, kuamua kama kuna njia ya mbele kwangu katika mbio hii”, lakini baadaye alisema kwamba angebaki katika mbio na kushiriki katika January 21 South Carolina primary.	Perry akiuga atĩ we “Nĩ egũcoka Texas kũrora maumĩrĩra ma mũcemanio wa atangoria ũtukũ ũcio, kũrora kana kwĩna gacĩra gakwa ga kũhotithia gũthiĩ na mbere gĩcindano-inĩ kĩu”, no thutha ũcio akiuga nĩ egũthiĩ na mbere na gĩcindano kĩu na nĩ egũcindana ithurano-inĩ cia kĩambĩrĩria mweri mĩrongo ĩrĩ na ũmwe South Carolina.

Table 12: Examples of BOW MT prompts and responses. Exact matches on words are colored in green, inflections in violet, and synonyms in blue.

E.1.1 What counts as "correct"

exact_match If the model prediction matched any uncased, unpunctuated ground-truth answer, the prediction was marked as an exact_match.

inflection We make use of the Python package fuzzywuzzy, a package that uses Levenshtein distance to perform fuzzy string matching. We classify a model prediction as an inflection should it achieve a fuzzywuzzy [5] similarity score of at least 75.

substring We mark a prediction as substring should any of the ground-truth answers exist as a word/phrase of the model's prediction, irrespective of punctuation or case.

inflection_within_substring We denote a model prediction as inflection_within_substring if any inflected form of the ground truth, as defined above, is contained within the model prediction, ignoring punctuation and case.

synonym We designate a model prediction as a synonym if it belongs to any WordNet synset of the ground truth answers. The usage of WordNet restricts this classification type to the X→model direction in WT, WTWC, and BOW MT.

E.1.2 Error classification

We designate the following categories of incorrect responses.

echo A prediction is an echo if it matches the word to be translated, ignoring casing and punctuation.

outputted_in_source_language If the prediction does not satisfy any of the above classification types but can be found on the source side of a translation lexicon, the prediction is marked as outputted_in_source_language.

gibberish Should the prediction fail to fall into any of these classification type, the prediction is marked as gibberish.

E.2 Manual Evaluation

Task	False Positive	False Negative
WT	2.5%	2.5%
WTWC	10.7%	1.7%
BOW MT	5.7%	1.7%

Table 13: To achieve these results, we evaluated 283 samples from WT, 121 from WTWC, and 229 from BOW MT.

To perform manual evaluation, we randomly selected a language for each model-evaluation direction pair and annotated at least 10 responses from it. Table 13 highlights low false positives and negatives are across our evaluations. This suggests that the evaluation metrics applied to our models are reliable.

[5] https://pypi.org/project/fuzzywuzzy/

F Prompt Exploration

We recognize that LLMs are sensitive to the prompts used for each task (Anagnostidis and Bulian, 2024).

We evaluated our six models on a series of "candidate" prompts: prompts that clearly delineate the word to translate as well as any additional context. We ran these small evaluations in Spanish as we assumed that if the model could not accurately perform the task in an HRL, such as Spanish, a model would be unlikely to do so in an LRL.

We list models and the candidate prompts they were matched with in the sections below. All our experiments use deterministic generation for decoding.

F.1 Word Translation

Our candidate prompts stress succinctness in the translation. We emphasized that the model translation be one word to make parsing simpler.

X→model

Prompt 1: We assigned the prompt below to `aya-23-8b`, `falcon-7b-instruct`, and `Llama-3.1-8B-Instruct` for WT in the `X→model` direction.

```
Translate the following word from {target language} to English. Respond
with a single word.
Word: {word}
Translation:
```

Prompt 2: We assigned the prompt below to `aya-101` and `bloomz-7b1-mt`.

```
Translate the following text from {target language} to English: {word}.
```

Prompt 3: We assigned the prompt below to `gemma-2b-it`.

```
Translate '{word}' from {target language} into English. Respond in one word.
```

model→X

Prompt 1: We assigned the prompt below to `aya-23-8b`, `falcon-7b-instruct`, `Llama-3.1-8B-Instruct`.

```
Translate the following word from English to {target language}. Respond
with a single word.
Word: {word}
Translation:
```

Prompt 2: We assigned the prompt below to `aya-101` and `bloomz-7b1-mt`.

```
Translate the following text from English to {target language}: {word}.
```

Prompt 3: We assigned the prompt below to `gemma-2b-it`.

```
Translate '{word}' from English to {target language}. Answer in one word:
```

F.2 Word Translation with Context

A common error we encountered involved models translating the entire sentence rather than a specific word. Consequently, our prompts emphasized translating a sole word.

X→model

Prompt 1: We assign the prompt below to `aya-101`.

```
What does '{word}' mean in English in the sentence '{sentence}'? Meaning
(one word):
```

Prompt 2: We assign the prompt below to `aya-23-8b` and `falcon-7b-instruct`.

```
In '{sentence}', the word '{word}' means ____ in English.
```

Prompt 3: We assign the prompt below to `bloomz-7b1-mt` and `Llama-3.1-8B-Instruct`.

```
Sentence: {sentence}
Define '{word}' in one English word:
```

Prompt 4: We assign the prompt to `gemma-2b-it`.

```
Sentence: {sentence}
English definition of '{word}'
```

`model`→`X`

Prompt 1: We assign the prompt below to `aya-101`.

```
What does '{word}' mean in {target language} in the sentence '{sentence}'?
Meaning (one word):
```

Prompt 2: We assign the prompt to `aya-23-8b`, `falcon-7b-instruct`, `gemma-2b-it`, `Llama-3.1-8B-Instruct`.

```
In '{sentence}', the word '{word}' means ____ in {target language}.
```

Prompt 3: We assign the prompt below to `bloomz-7b1-mt`.

```
Define '{word}' in '{sentence}' in {target language}:
```

F.3 Translation-Conditioned Language Modeling

Prompt construction depended on model architecture. Because `aya-101` uses an encoder-decoder architecture, the first n words in the target translation are fed into the decoder rather than encoded as a prompt. The remaining five models utilized decoder architecture; the target translation of the first n words was part of the prompt.

`X`→`model`

Prompt 1: We assign the prompt below to `aya-101`.

```
Translate the sentence into English:
{Target Language}:{source sentence}
English:
```

Prompt 2: We assign the prompt below to `aya-23-8b`, `bloomz-7b1-mt`, `falcon-7b-instruct`, `gemma-2b-it`, and `Llama-3.1-8B-Instruct`.

```
Translate the sentence into English.
{Target Language}:{source sentence}
English: {target translation up to index n}
```

`model`→`X`

Prompt 1: We assign the prompt below to `aya-101`.

```
Translate the following text into {target language}.
English: {source sentence}
{Target Language}:
```

Prompt 2: We assign the prompt below to `aya-23-8b`, `bloomz-7b1-mt`, `gemma-2b-it`, `falcon-7b-instruct`, and `Llama-3.1-8B-Instruct`.

```
Translate the following text into {target language}.
English:{source sentence}
Target Language: {target translation up to index n}
```

F.4 Bag-of-Words Machine Translation

When prompted to translate a sentence, model outputs often missed the objective; models provided additional context to the subject of the sentence. To avoid confusion of what was expected, we made the act of translation as explicit as possible.

X→model

Prompt 1: We assigned the prompt below to `gemma-2b-it`.

```
Sentence: {source sentence}
English translation:
```

Prompt 2: We assigned the prompt below to `Llama-3.1-8B-Instruct`.

```
What does this sentence mean in English: {source sentence}?
```

Prompt 3: We assigned the prompt below to `aya-101`, `aya-23-8b`, `bloomz-7b1-mt`, and `falcon-7b-instruct`.

```
Translate into English: {source sentence}
```

model→X

Prompt 1: We assigned the prompt below to `gemma-2b-it`.

```
Sentence: {source sentence}
{Target Language} translation:
```

Prompt 2: We assigned the prompt below to `Llama-3.1-8B-Instruct`.

```
English sentence: {source sentence}
{Target Language} translation:
```

Prompt 3: We assigned the prompt below to `aya-101`, `aya-23-8b`, `bloomz-7b1-mt`, and `falcon-7b-instruct`.

```
Translation into {target language}: {source sentence}
```

G Results in Detail

G.1 Feature Importance

We trained a decision tree regressor on several features of a language: whether the model supports a language, the language's resource level (i.e. the number of Wikipedia pages available), which model predicted the language (e.g. `bloomz-7b1-mt`, `Llama-3.1-8B-Instruct`, `falcon-7b-instruct`), which language family the language belonged to (e.g. Atlantic-Congo, Indo-European), what evaluation direction the model was assessed under, what script the language used (e.g. Latin), and the languages associated score. For task-specific decision trees, see Figure 7, Figure 8, Figure 9, and Figure 10. Table 14 averages feature importance values and enumerates them in descending order.

Feature	**Average Feature Importance**
Translation mode: `model`→`X`	0.264 ± 0.21
Supported by model	0.2425 ± 0.13
Resource level	0.19 ± 0.16
Model: `bloomz-7b1-mt`	0.148 ± 0.14
Model: `falcon-7b-instruct`	0.053 ± 0.07
Family: Indo-European	0.048 ± 0.09
Script: Latin	0.03 ± 0.03
Model: `Llama-3.1-8B-Instruct`	0.024 ± 0.03
Script: Devanagari	0.002 ± 0.0

Table 14: For each task, we trained a decision tree regressor with the language score as the label and attributes, such as the model in which the language was evaluated, the language's family, and its script, as features. Each regressor assigns importance scores for the features, ranging from 0 to 1 and reflecting their contribution to predicting the language's score. We then averaged feature importance across the four tasks and reported the features with non-zero importance scores. The overall average of each feature is depicted on the left of $\pm$ and the standard deviation on the right.

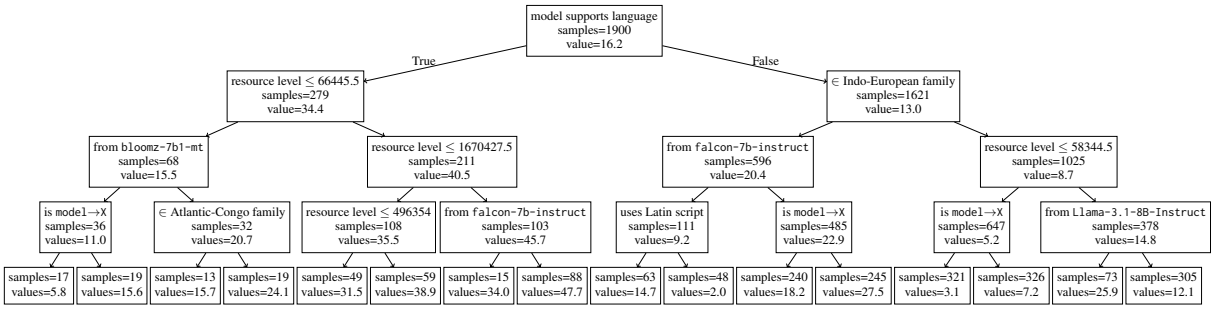

Figure 7: A decision tree trained on linguistic and task features as well as **Word Translation** language scores.

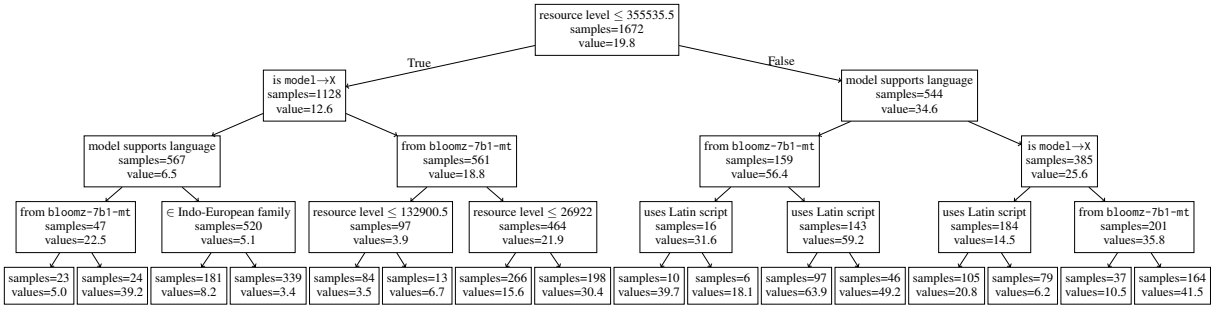

Figure 8: A decision tree trained on linguistic and task features and **Word Translation with Context** language scores.

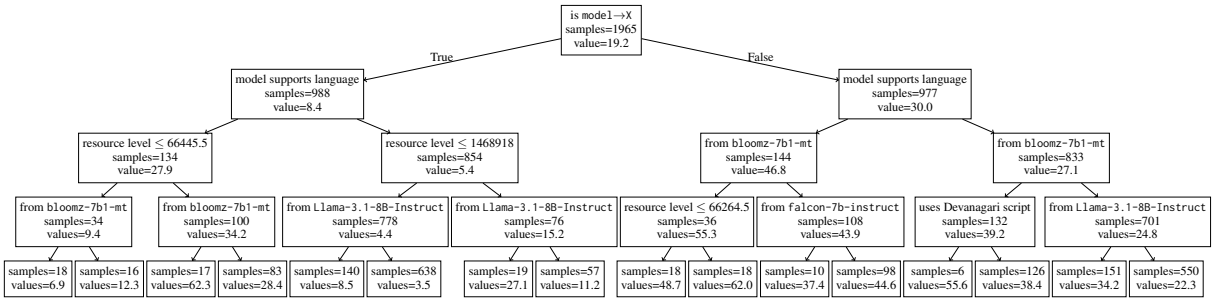

Figure 9: A decision tree trained on linguistic and task features as well as **Translation-Conditioned Language Modeling** language scores.

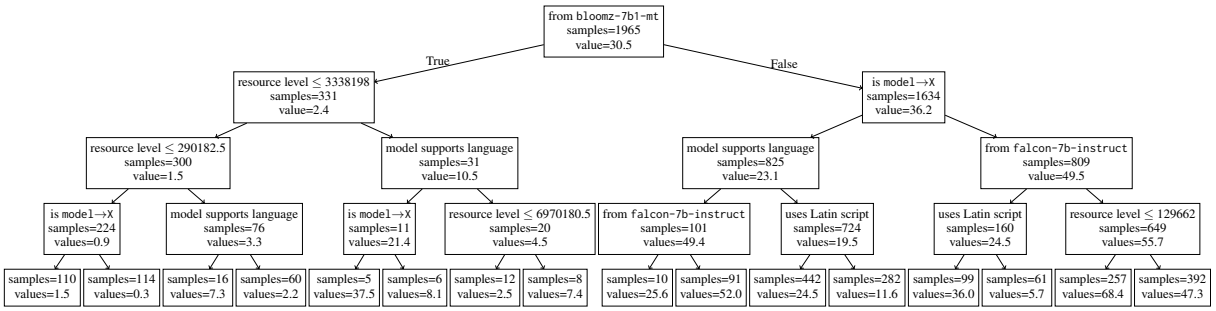

Figure 10: A decision tree trained on linguistic and task features as well as **Bag-of-Words Machine Translation** language scores.

G.2 Model Averages

Table 15 lists the model score averages across all tasks and evaluation directions.

	WT		WTWC		TCLM		BOW MT	
Model	**X→model**	**model→X**	**X→model**	**model→X**	**X→model**	**model→X**	**X→model**	**model→X**
aya-101	4.8 ± 10.9	3.4 ± 8.8	**29.0 ± 22.3**	**14.2 ± 19.8**	36.1 ± 12.2	**14.2 ± 11.7**	**65.6 ± 18.4**	**35.9 ± 19.0**
aya-23-8b	4.8 ± 10.2	2.8 ± 8.7	21.0 ± 22.2	7.2 ± 13.9	26.2 ± 12.3	7.2 ± 10.7	52.9 ± 21.2	26.4 ± 20.6
bloomz-7b1-mt	9.0 ± 8.2	2.1 ± 6.3	3.8 ± 6.0	3.6 ± 7.8	**42.7 ± 12.1**	10.2 ± 18.4	1.1 ± 2.5	3.7 ± 8.5
falcon-7b-instruct	2.8 ± 5.3	1.3 ± 4.8	12.2 ± 15.9	4.9 ± 9.4	17.6 ± 7.0	1.9 ± 5.0	25.5 ± 19.4	10.8 ± 8.5
gemma-2b-it	4.1 ± 6.2	1.5 ± 4.6	18.0 ± 14.7	2.2 ± 5.0	22.9 ± 9.5	5.0 ± 6.5	48.0 ± 16.5	17.5 ± 13.4
Llama-3.1-8B-Instruct	**14.0 ± 8.7**	**3.6 ± 7.9**	15.3 ± 15.9	6.7 ± 10.6	35.3 ± 12.5	11.7 ± 11.5	58.0 ± 15.7	26.4 ± 15.5

Table 15: Average model accuracy percentages across tasks and comprehension (X → eng) and generative settings (eng → X) using the evaluation metrics defined in Appendix E. The best-performing models for each task and direction are **bolded** and underlined.

G.3 Language Family Averages

Figure 3 shows for each task, the *best* language family average across six models. We show language family averages across all tasks and models in Table 16, Table 17, Table 18, and Table 19. While the Indo-European language family's average tends to be higher, there is more variation within the models themselves. In WT X→model, aya-101's Turkic language family average is 11.7% higher than falcon-7b-instruct's Indo-European language family average.

	Indo-European		Turkic		Austronesian		Afro-Asiatic		Atlantic-Congo	
Model	**X→model**	**model→X**	**X→model**	**model→X**	**X→model**	**model→X**	**X→model**	**model→X**	**X→model**	**model→X**
aya-101	27.4 ± 17.6	20.0 ± 16.9	23.4 ± 12.4	10.2 ± 13.5	3.6 ± 6.7	3.9 ± 7.6	5.7 ± 13.9	3.4 ± 9.8	2.0 ± 4.0	1.2 ± 3.0
aya-23-8b	25.2 ± 17.8	20.0 ± 18.0	14.5 ± 9.2	4.8 ± 8.4	4.1 ± 6.6	2.7 ± 7.1	5.7 ± 14.5	3.2 ± 10.9	1.7 ± 2.4	0.4 ± 1.0
bloomz-7b1-mt	21.9 ± 16.7	11.3 ± 13.6	9.9 ± 4.2	1.3 ± 2.4	7.7 ± 5.5	3.1 ± 5.7	9.9 ± 10.6	2.4 ± 9.2	8.7 ± 4.1	0.7 ± 1.8
falcon-7b-instruct	11.7 ± 11.7	9.3 ± 13.2	3.6 ± 2.2	1.6 ± 3.0	2.5 ± 2.8	0.8 ± 1.7	1.8 ± 1.6	0.3 ± 0.9	1.7 ± 1.9	0.3 ± 0.8
gemma-2b-it	15.6 ± 13.1	10.0 ± 12.0	8.2 ± 4.6	1.6 ± 2.7	3.1 ± 3.1	1.0 ± 2.4	3.7 ± 4.1	0.6 ± 1.5	2.4 ± 2.5	0.5 ± 1.0
Llama-3.1-8B-Instruct	29.3 ± 14.5	19.1 ± 14.8	22.1 ± 9.0	9.0 ± 9.8	13.9 ± 6.5	4.2 ± 6.7	15.2 ± 11.1	3.4 ± 8.4	11.2 ± 4.5	1.4 ± 2.1

Table 16: Averaging scores by language family, model, and evaluation direction (i.e. X→model or model→X) for the task **Word Translation**. Data is written in the format mean ± standard deviation.

	Indo-European		Turkic		Austronesian		Afro-Asiatic		Atlantic-Congo	
Model	**X→model**	**model→X**	**X→model**	**model→X**	**X→model**	**model→X**	**X→model**	**model→X**	**X→model**	**model→X**
aya-101	49.1 ± 17.4	30.4 ± 22.3	47.4 ± 15.5	12.0 ± 17.4	22.8 ± 18.2	10.1 ± 15.1	30.6 ± 23.2	21.9 ± 22.5	19.6 ± 12.9	5.7 ± 10.6
aya-23-8b	43.8 ± 19.6	17.1 ± 18.9	24.8 ± 16.9	4.5 ± 10.8	15.2 ± 16.6	4.3 ± 10.3	24.5 ± 26.7	10.8 ± 18.1	6.2 ± 4.3	1.0 ± 1.3
bloomz-7b1-mt	8.3 ± 8.6	8.8 ± 12.4	1.2 ± 0.9	0.9 ± 1.5	2.4 ± 4.4	2.0 ± 4.1	3.4 ± 3.6	4.4 ± 9.3	2.3 ± 3.0	1.2 ± 2.5
falcon-7b-instruct	27.2 ± 21.0	13.0 ± 14.6	4.9 ± 5.2	1.8 ± 2.8	9.6 ± 9.6	2.6 ± 3.9	6.6 ± 5.1	1.9 ± 3.1	4.7 ± 2.2	1.1 ± 1.0
gemma-2b-it	31.8 ± 16.3	5.9 ± 8.1	14.8 ± 8.8	0.6 ± 1.3	13.7 ± 10.8	1.0 ± 2.5	19.4 ± 15.9	0.5 ± 1.0	9.9 ± 4.2	0.4 ± 0.5
Llama-3.1-8B-Instruct	31.2 ± 16.3	14.8 ± 15.0	19.2 ± 14.1	3.2 ± 6.3	10.6 ± 11.3	5.0 ± 6.6	18.2 ± 16.0	2.7 ± 3.7	5.4 ± 4.5	2.4 ± 2.0

Table 17: Averaging scores by language family, model, and evaluation direction (i.e. X→model or model→X) for the task **Word Translation with Context**. Data is written in the format mean ± standard deviation.

	Indo-European		Turkic		Austronesian		Afro-Asiatic		Atlantic-Congo	
Language	**X→model**	**model→X**	**X→model**	**model→X**	**X→model**	**model→X**	**X→model**	**model→X**	**X→model**	**model→X**
aya-101	43.7 ± 4.6	21.7 ± 11.6	37.3 ± 6.9	10.4 ± 9.0	35.9 ± 9.8	12.0 ± 8.2	36.6 ± 12.9	14.8 ± 10.0	24.8 ± 11.3	4.7 ± 4.8
aya-23-8b	32.8 ± 11.4	12.0 ± 13.2	22.8 ± 6.7	3.0 ± 6.0	23.7 ± 8.9	5.7 ± 7.1	30.0 ± 15.9	13.1 ± 13.9	14.7 ± 1.2	1.2 ± 0.4
bloomz-7b1-mt	46.8 ± 11.9	16.2 ± 23.7	33.0 ± 1.7	0.5 ± 0.2	39.3 ± 7.8	6.4 ± 11.0	43.3 ± 16.1	9.3 ± 10.6	44.4 ± 8.7	5.4 ± 5.1
falcon-7b-instruct	20.5 ± 9.8	3.9 ± 8.0	15.3 ± 0.6	0.2 ± 0.1	17.3 ± 3.8	1.9 ± 1.6	14.4 ± 1.8	0.4 ± 0.2	14.2 ± 1.2	0.7 ± 0.2
gemma-2b-it	27.5 ± 9.9	8.1 ± 8.5	19.7 ± 3.8	1.5 ± 2.1	20.6 ± 7.2	4.8 ± 4.8	23.3 ± 9.0	4.3 ± 3.6	14.8 ± 1.5	1.5 ± 0.6
Llama-3.1-8B-Instruct	43.9 ± 6.3	19.2 ± 12.1	35.7 ± 5.9	7.8 ± 5.6	32.8 ± 11.0	8.4 ± 7.7	34.6 ± 14.0	11.3 ± 9.9	20.3 ± 5.3	2.4 ± 2.1

Table 18: Averaging scores by language family, model, and evaluation direction (i.e. X→model or model→X) for the task **Translation-Conditioned Language Modeling**. Data is written in the format mean ± standard deviation.

	Indo-European		Turkic		Austronesian		Afro-Asiatic		Atlantic-Congo	
Language	**X→model**	**model→X**	**X→model**	**model→X**	**X→model**	**model→X**	**X→model**	**model→X**	**X→model**	**model→X**
aya-101	75.3 ± 10.0	45.4 ± 16.4	69.0 ± 10.1	25.0 ± 20.4	57.7 ± 24.5	39.3 ± 22.6	60.8 ± 18.2	35.7 ± 18.4	56.2 ± 12.2	26.5 ± 10.5
aya-23-8b	65.1 ± 17.9	36.2 ± 21.9	48.8 ± 11.6	16.0 ± 14.8	49.9 ± 19.1	31.5 ± 21.7	53.5 ± 26.1	33.3 ± 25.7	33.1 ± 4.6	15.9 ± 5.4
bloomz-7b1-mt	2.1 ± 3.6	5.2 ± 9.3	0.0 ± 0.0	0.5 ± 0.6	0.6 ± 1.9	5.5 ± 12.8	1.1 ± 1.4	0.5 ± 0.5	0.3 ± 0.4	1.6 ± 1.4
falcon-7b-instruct	30.2 ± 25.2	12.8 ± 11.2	14.8 ± 10.6	4.3 ± 3.4	31.8 ± 16.0	14.6 ± 6.7	12.5 ± 10.1	4.9 ± 4.7	26.2 ± 4.3	11.6 ± 3.5
gemma-2b-it	56.9 ± 16.5	22.8 ± 16.1	42.4 ± 7.8	8.4 ± 7.7	45.1 ± 14.8	23.7 ± 14.3	44.0 ± 16.8	11.3 ± 5.8	34.8 ± 4.2	14.4 ± 4.1
Llama-3.1-8B-Instruct	67.8 ± 9.8	34.2 ± 15.1	62.2 ± 6.5	17.2 ± 10.1	57.3 ± 15.1	33.0 ± 16.8	53.5 ± 20.1	20.4 ± 10.9	45.1 ± 5.9	20.1 ± 5.7

Table 19: Averaging scores by language family, model, and evaluation (i.e. `X→model` or `model→X`) for the task **Bag-of-Words Machine Translation**. Data is written in the format `mean ± standard deviation`.

G.4 Resourceness

Figure 11 compares resource level against language scores across all tasks and evaluation directions.

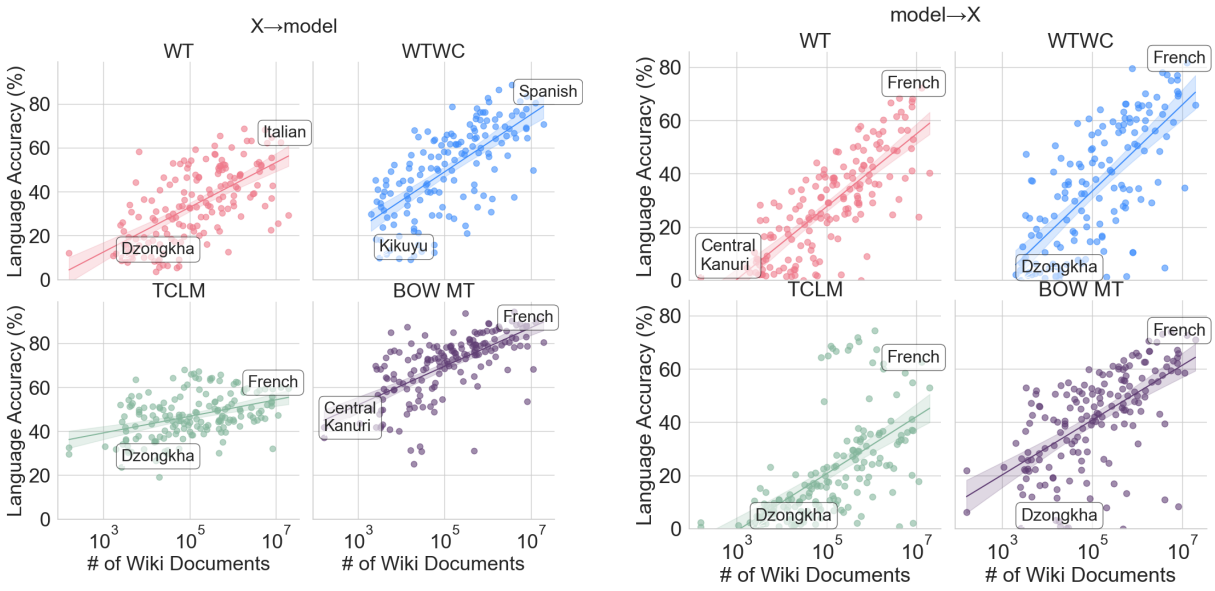

Figure 11: Comparison of the number of Wikipedia documents—a proxy for resource level—and language performance for each task. For each language, the highest score among the six evaluated models was used. Resource levels are shown on a logarithmic scale to account for their wide range. Scatterplot labels indicate the lowest-performing low-resource language and the highest-performing high-resource language. The fitted lines in each plot depict the overall trend between resource level and performance. The shaded regions represent 95% confidence band, which are consistently narrow and indicate the high precision of the fitted lines.

G.5 Sampled Languages

We sample 22 languages in our four tasks and display their scores in Table 20, Table 21, Table 22, and Table 23.

	aya-101		aya-23-8b		bloomz-7b1-mt		falcon-7b-instruct		gemma-2b-it		LLaMA-3.1-8b Instruct	
Language	X→model	model→X	X→model	model→X	X→model	model→X	X→model	model→X	X→model	model→X	X→model	model→X
Romanian	55.7	66.6	63.3	60.9	35.8	34.4	32.0	39.8	38.5	40.5	49.6	38.6
Bulgarian	52.3	48.9	41.3	42.7	24.9	5.4	6.9	8.6	25.0	10.3	43.0	4.6
Slovak	50.0	27.2	42.2	50.7	23.4	15.7	17.1	11.5	18.4	18.5	45.2	58.0
Haitian	46.6	39.8	30.2	27.3	25.2	14.4	15.4	12.2	22.8	5.4	36.0	25.6
Spanish	44.7	55.8	50.4	61.1	51.2	68.3	25.4	62.2	34.6	47.1	50.0	62.1
Awadhi	41.6	0.0	44.3	17.1	43.3	10.5	0.0	0.3	10.5	0.6	41.9	13.9
Czech	37.8	28.1	48.0	49.5	16.0	10.2	14.2	11.9	14.2	23.6	43.8	41.9
Friulian	36.7	27.8	30.8	27.3	31.8	20.3	20.5	27.9	25.5	14.3	31.2	15.6
Korean	34.1	18.4	39.3	48.9	9.3	0.1	2.4	0.0	15.2	12.8	29.6	22.5
Sundanese	32.7	34.0	16.4	23.3	21.1	20.8	6.5	2.3	6.5	7.6	20.7	16.9
Telugu	30.8	34.4	7.0	4.7	28.9	30.5	0.0	0.0	3.0	0.0	11.4	26.4
Hungarian	28.7	49.4	17.9	22.4	8.3	2.3	3.5	18.5	6.6	5.3	28.4	51.8
Balinese	25.9	22.6	27.7	23.5	25.6	23.3	8.2	1.1	7.7	5.5	25.7	15.0
Sindhi	23.7	23.8	20.7	2.4	11.7	0.0	2.0	0.0	7.3	0.6	33.3	18.8
Turkmen	17.1	2.5	17.9	4.8	11.4	6.4	6.5	2.2	9.4	3.2	20.9	15.5
Pedi	12.7	1.0	3.8	1.8	16.7	3.5	2.9	2.2	4.6	4.7	5.0	2.4
Sanskrit	11.0	23.4	13.3	23.3	11.3	7.1	1.0	0.0	4.0	0.0	19.7	19.7
Somali	8.2	22.2	5.4	12.0	11.7	2.2	2.2	0.8	6.4	1.4	11.6	6.5
Sango	5.1	0.9	7.8	0.3	6.9	2.9	2.3	0.0	6.6	0.3	11.8	0.7
Nyanja	3.4	9.2	4.0	0.1	8.1	5.4	1.0	0.4	1.7	0.2	9.7	4.0
Kabyle	0.7	0.0	0.4	0.0	3.6	0.2	0.4	1.4	2.4	0.0	3.0	0.4
Mossi	0.0	0.5	3.3	0.0	5.4	1.3	1.6	0.0	4.4	0.0	5.4	0.3

Table 20: Performance on task **Word Translation** across 22 sampled languages.

	aya-101		aya-23-8b		bloomz-7b1-mt		falcon-7b-instruct		gemma-2b-it		LLaMA-3.1-8b Instruct	
Language	X→model	model→X	X→model	model→X	X→model	model→X	X→model	model→X	X→model	model→X	X→model	model→X
Spanish	78.8	69.3	81.1	64.0	30.3	65.0	83.4	65.0	77.0	34.9	51.7	59.7
Bulgarian	76.1	60.8	70.5	33.1	7.8	3.2	29.4	9.8	46.5	1.6	41.3	5.8
Slovak	70.0	66.2	78.0	42.5	6.5	16.9	38.5	17.3	63.0	15.3	55.2	21.0
Korean	68.6	31.2	63.4	48.5	22.0	1.3	11.8	0.1	48.4	3.9	54.7	0.6
Czech	68.2	70.1	69.8	62.6	5.3	7.2	35.9	27.9	48.2	16.8	60.1	31.1
Hungarian	64.1	60.1	59.4	24.1	1.3	11.1	8.7	6.7	36.6	5.1	46.5	23.9
Romanian	63.6	68.9	65.4	59.3	16.9	8.1	58.8	29.8	54.0	13.1	32.3	33.1
Haitian	61.2	60.6	48.4	2.6	7.8	7.9	27.9	10.0	27.1	2.0	24.7	14.3
Sundanese	58.2	54.8	32.1	12.6	3.4	4.3	13.3	2.6	16.8	0.8	19.3	5.6
Turkmen	45.2	13.6	33.8	4.7	1.5	1.1	6.0	4.8	16.3	0.7	27.8	8.3
Pedi	44.9	0.6	6.5	0.3	9.0	3.7	6.5	2.8	11.7	0.5	5.9	4.5
Balinese	43.9	14.9	25.3	8.6	1.2	2.0	12.8	2.4	16.7	1.3	21.9	4.5
Friulian	42.6	22.4	43.8	12.4	3.2	7.9	28.6	18.7	21.9	6.5	22.7	16.8
Awadhi	39.2	0.0	39.1	6.2	2.4	0.0	2.6	0.0	27.8	0.0	32.2	0.0
Sindhi	32.2	32.5	30.8	1.4	1.1	0.5	3.8	0.5	16.6	0.0	40.4	0.6
Sango	29.8	1.6	6.6	1.0	1.9	0.7	6.7	0.9	11.9	0.0	3.1	2.1
Sanskrit	29.3	1.6	27.2	23.4	0.8	0.0	6.1	5.9	22.2	0.2	23.2	1.9
Nyanja	28.3	37.3	7.4	2.4	3.0	18.8	4.2	3.2	5.3	0.9	10.0	6.3
Telugu	21.0	22.4	10.9	2.1	2.4	10.3	1.7	0.1	14.1	0.0	6.7	6.9
Somali	9.7	33.1	15.9	3.1	2.2	0.7	3.9	1.6	12.8	0.1	6.9	2.9
Mossi	3.7	0.4	3.4	0.5	1.0	0.1	3.1	0.3	8.0	0.1	0.8	4.5
Kabyle	2.0	3.4	4.1	0.6	1.6	0.4	1.6	3.6	9.2	0.1	2.6	5.2

Table 21: Performance on task **Word Translation-in-Context** across 22 sampled languages.

	aya-101		aya-23-8b		bloomz-7b1-mt		falcon-7b-instruct		gemma-2b-it		LLaMA-3.1-8b Instruct	
Language	X→model	model→X	X→model	model→X	X→model	model→X	X→model	model→X	X→model	model→X	X→model	model→X
Romanian	47.3	38.4	49.1	40.6	48.5	4.3	32.8	7.5	39.7	18.9	50.7	38.5
Telugu	46.7	24.3	17.9	1.8	64.3	68.5	14.3	0.3	18.1	2.2	45.4	19.7
Bulgarian	46.6	36.8	41.1	8.2	44.7	2.5	15.1	0.8	39.7	12.8	49.9	29.1
Sindhi	46.5	21.0	18.4	1.5	37.5	1.1	12.5	0.2	16.0	1.2	40.3	10.1
Czech	46.3	33.2	47.6	35.9	38.9	2.4	23.6	2.6	39.9	16.1	49.2	33.5
Awadhi	45.8	16.6	39.7	15.8	63.5	23.0	13.8	0.2	27.0	6.1	44.5	19.2
Sundanese	45.7	15.6	27.6	5.3	46.5	5.0	18.6	1.1	21.9	3.8	41.2	8.6
Slovak	45.4	33.7	44.1	14.4	37.6	2.1	19.6	1.6	35.1	10.9	47.1	26.0
Haitian	44.6	22.4	21.9	3.2	39.9	2.3	16.8	1.2	18.6	2.4	37.3	8.1
Spanish	43.4	36.4	44.4	37.3	55.7	54.8	40.7	29.5	41.5	31.8	46.0	38.4
Hungarian	43.3	26.5	31.1	4.2	31.7	0.9	14.2	0.5	28.0	5.5	46.8	26.3
Korean	41.3	20.1	42.4	22.7	40.7	1.2	14.8	0.3	35.7	11.2	44.6	18.5
Friulian	40.9	4.4	33.6	4.8	46.2	3.0	22.1	1.5	25.5	3.1	41.4	11.0
Balinese	39.8	12.3	28.4	8.2	44.9	8.2	19.0	1.8	24.4	6.4	38.0	8.9
Pedi	38.3	6.8	14.3	1.7	57.9	8.4	13.9	1.1	14.4	2.5	20.3	3.3
Somali	36.7	10.1	15.8	2.0	30.6	1.0	14.5	0.7	14.5	1.4	22.1	2.3
Nyanja	35.6	8.7	15.0	1.6	52.8	7.0	14.6	1.1	14.9	1.7	22.1	2.9
Turkmen	35.4	2.4	22.2	1.5	31.1	0.5	15.4	0.2	17.1	0.9	31.2	3.5
Sanskrit	30.6	2.5	19.9	1.7	41.2	2.4	13.5	0.1	18.6	1.0	32.0	3.9
Mossi	12.5	1.1	14.2	0.9	34.1	1.2	14.2	0.7	14.4	1.1	16.9	0.9
Sango	11.9	2.1	11.9	1.5	31.9	2.4	11.5	0.9	11.0	1.8	13.5	1.8
Kabyle	10.6	1.5	11.4	0.7	26.6	0.8	11.2	0.4	11.5	1.0	14.6	1.4

Table 22: Performance on task **Translation-Conditioned Language Modeling** across 22 sampled languages.

	aya-101		aya-23-8b		bloomz-7b1-mt		falcon-7b-instruct		gemma-2b-it		LLaMA-3.1-8b Instruct	
Language	X→model	model→X	X→model	model→X	X→model	model→X	X→model	model→X	X→model	model→X	X→model	model→X
Spanish	87.6	57.0	88.3	58.9	9.4	32.2	79.1	30.9	85.5	51.6	79.7	51.9
Slovak	84.5	57.5	82.8	45.2	0.8	5.5	43.1	13.4	70.9	30.9	77.6	44.0
Romanian	82.8	63.5	85.6	70.2	4.8	5.4	57.0	20.6	74.4	45.1	78.7	56.2
Czech	81.6	57.6	85.9	62.6	0.7	4.7	47.9	13.5	71.8	37.1	76.0	46.7
Bulgarian	78.9	61.4	79.0	27.7	1.1	0.7	9.1	2.1	72.7	16.9	73.4	37.5
Korean	76.7	35.1	80.5	43.1	5.9	1.0	21.4	2.9	62.9	18.6	71.4	27.1
Friulian	76.0	31.7	67.3	40.9	0.9	2.5	44.6	23.0	56.9	25.9	67.5	41.9
Haitian	75.7	59.7	49.9	23.0	0.2	1.6	34.0	22.4	44.7	25.6	65.5	35.7
Telugu	74.6	42.8	45.3	3.8	3.5	9.1	2.7	1.7	44.4	3.7	23.0	21.2
Hungarian	73.9	51.9	61.3	22.3	0.3	2.9	29.2	11.0	57.1	23.7	64.7	47.8
Sundanese	73.7	52.8	57.6	50.8	0.0	5.5	34.9	16.9	49.0	24.4	64.5	35.4
Balinese	71.2	45.0	56.1	52.5	0.4	3.6	35.8	18.0	50.0	35.9	62.2	43.5
Sanskrit	66.2	15.6	51.3	17.2	0.1	0.4	3.3	1.4	47.6	7.5	55.9	11.4
Turkmen	65.9	11.1	44.0	12.9	0.0	0.5	23.8	6.9	36.0	8.5	56.2	14.7
Awadhi	65.9	3.6	76.2	50.6	0.5	0.6	4.0	1.9	56.8	19.3	71.2	32.9
Nyanja	64.0	42.7	32.2	14.9	0.0	0.6	26.1	14.3	34.2	14.6	48.3	19.0
Pedi	61.8	21.3	38.9	18.7	0.8	0.6	32.7	16.0	37.7	19.1	45.9	23.1
Sindhi	58.3	41.2	33.7	3.7	0.0	0.4	2.3	2.3	31.9	3.9	60.3	19.0
Somali	53.7	39.5	34.1	17.5	0.1	0.8	21.3	10.3	28.7	11.0	46.3	17.8
Sango	50.1	32.2	34.5	22.4	0.1	6.7	29.2	20.0	39.4	20.7	43.4	29.0
Mossi	34.2	14.4	28.5	17.0	0.0	0.6	25.2	10.0	30.9	14.1	34.2	16.0
Kabyle	19.2	9.9	19.5	8.8	0.0	0.5	16.0	7.7	20.8	9.0	31.3	11.5

Table 23: Performance on task **Bag-of-Words Machine Translation** across 22 sampled languages.

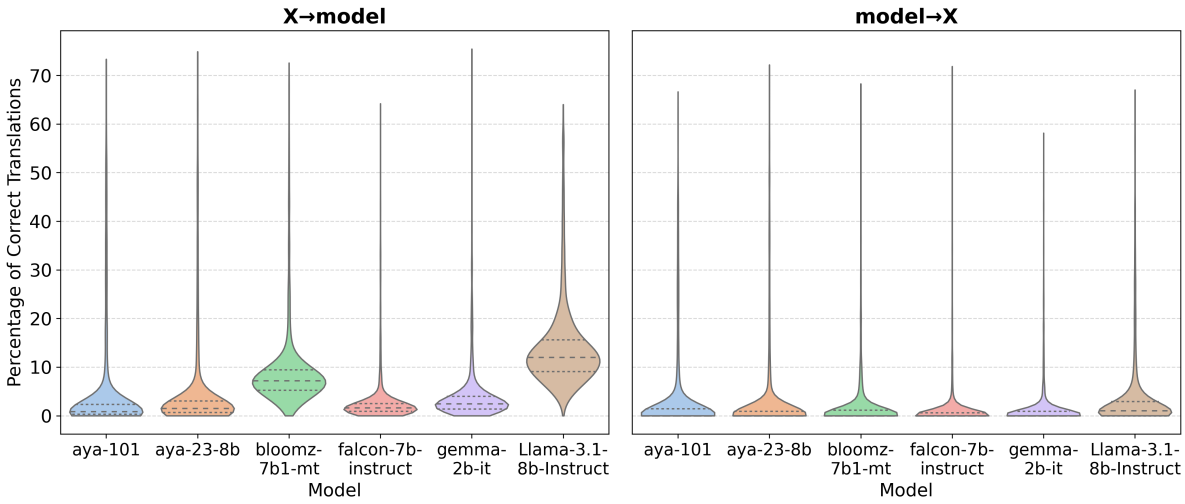

Figure 12: Model-wise performance distribution for the task **Word Translation**. Each violin depicts the distribution of scores across evaluated languages. Dotted lines indicate the first, second, and third quartiles of this distribution.

G.6 Language Score Distribution

Figure 12, Figure 13, Figure 14, and Figure 15 outline the distribution of language scores for each task.

You may notice that `bloomz-7b1-mt` performs especially badly in Figure 15. Interestingly, the model average is higher in `model`→`X` than in `X`→`model`. The model performs poorly even with HRLs, receiving a score of 9.4% for the Spanish→English translations (see Table 23). The model `bloomz-7b1-mt` had difficulty following instructions, often echoing the prompt. For example, `bloomz-7b1-mt` echoes the source sentence when tasked with translating a Swedish sentence:

Prompt
Translate into English: “Vi har nu 4 månader gamla möss som har blivit kvitt sin diabetes”, tillade han.

Model Response:
Translate into English: “Vi har nu 4 månader gamla möss som har blivit kvitt sin diabetes”, tillade han.

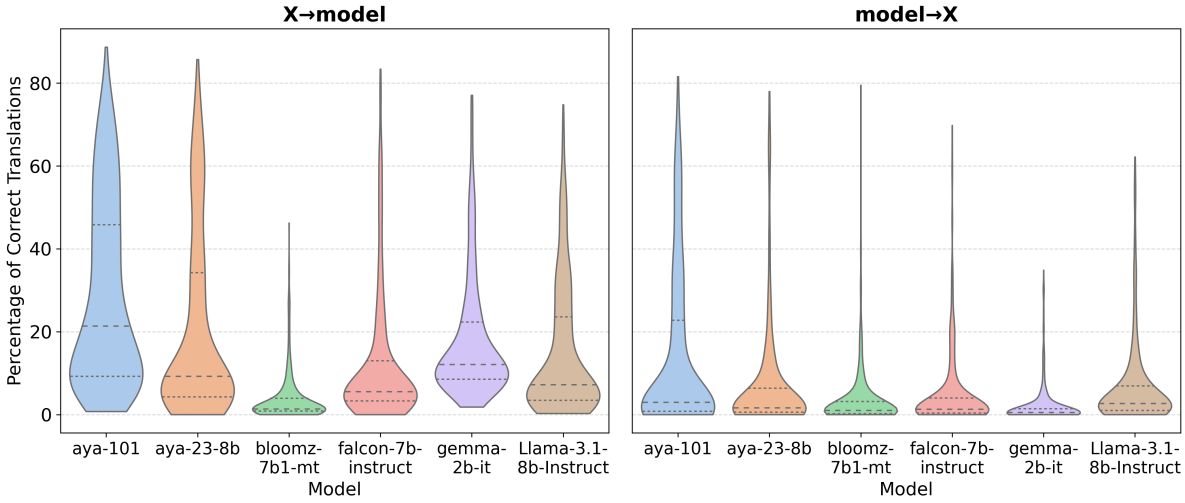

Figure 13: Model-wise performance distribution for the task **Word Translation with Context**. Each violin depicts the distribution of scores across evaluated languages. Dotted lines indicate the first, second, and third quartile of this distribution.

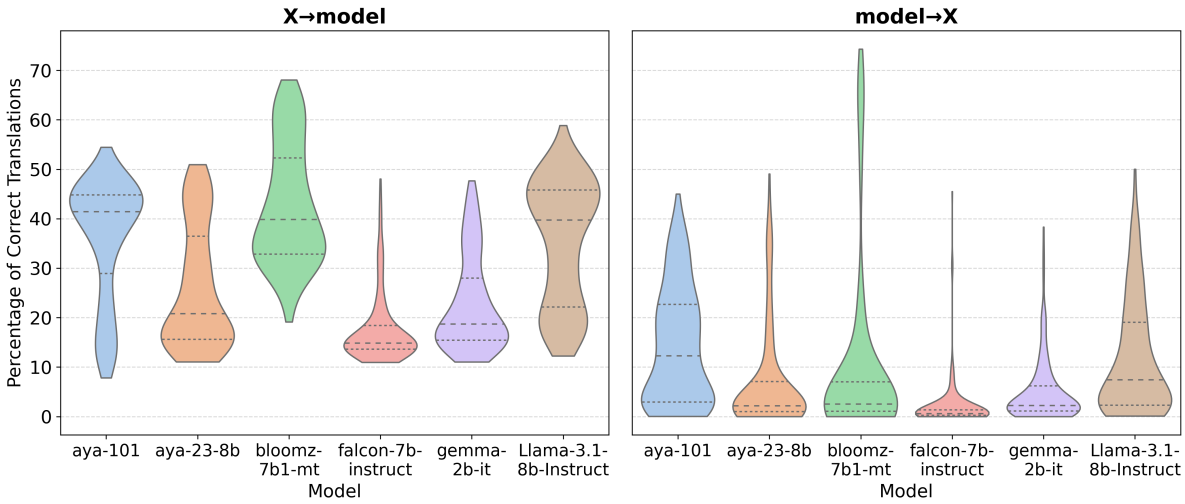

Figure 14: Model-wise performance distribution for the task **Translation-Conditioned Language Modeling**. Each violin depicts the distribution of scores across evaluated languages. Dotted lines indicate the first, second, and third quartile of this distribution.

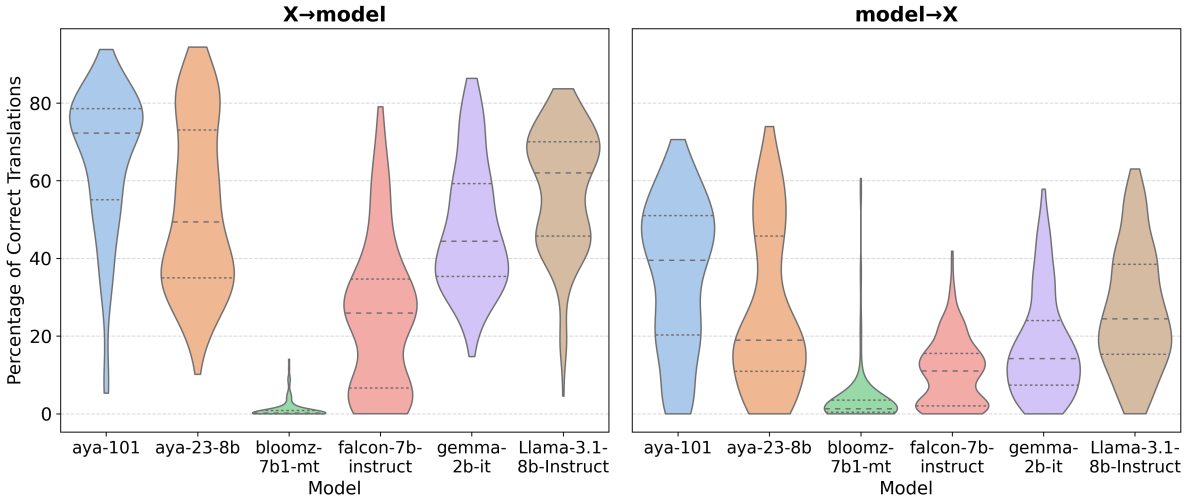

Figure 15: Model-wise performance distribution for the task **Bag-of-Words Machine Translation**. Each violin depicts the distribution of scores across evaluated languages. Dotted lines indicate the first, second, and third quartile of this distribution.

H Sampling

Due to the large size of our dataset and limited compute, we evaluated only a sample of existing data. We explain what this means in each task.

- Word Translation: We randomly sample 300 entries from the translation lexicon should more than 300 entries exist.
- Word Translation With Context: We prompt a model until we have evaluated 300 unique words.
- Translation-Conditioned Language Modeling: We prompt the model on words from the first 300 sentences in FLORES+.
- Bag-of-Words Machine Translation: Similarly to TCLM, we prompt the model on words from the first 300 sentences in FLORES+.

I Evaluation Details

We used A40s A100s, and A6000s to run evaluation on WT, WTWC, TCLM, and BOW MT. We discuss the GPU compute hours in more detail.

I.1 Compute

Word Translation We conducted evaluation for 2,746 languages $\times$ 2 evaluation directions $\times$ 6 model $=$ 32,952 evaluations. Each run takes approximately 6 minutes, resulting in 3,295.2 GPU hours.

Word Translation with Context We conducted evaluation for 525 languages $\times$ 2 evaluation directions $\times$ 6 models $=$ 6,300 evaluations. Each run takes approximately 40 minutes, resulting in 4,200 GPU hours.

Translation-Conditioned Language Modeling We conducted evaluation for 211 languages $\times$ 2 evaluation directions $\times$ 6 models $=$ 2,532 evaluations. Each run takes approximately 6 minutes, resulting in 253.2 GPU hours.

Bag-of-Words Machine Translation We conducted evaluation for 211 languages $\times$ 2 evaluation directions $\times$ 6 models $=$ 2,532 evaluations. Each run takes approximately 3 minutes, resulting in 126.6 GPU hours.

I.2 Evaluation

We tested our models on `devtest` splits of the FLORES+ dataset and version `v3.1` from GLOTLID. We also used BLEU (HuggingFace `evaluate` wrapper), and WORDNET from `nltk.corpus`.

Model	Languages Supported	Release Year	Architecture	Training Data Mixture	Rationale
`aya-101` (13B)	101	2024	encoder-decoder	multilingual templates, human annotations, synthetic data, machine translation	trained on many low-resource languages
`aya-23-8b`	23	2024	decoder	multilingual templates, human annotations, synthetic data, machine translation	outperforms `aya-101` across 23 covered languages
`bloomz-7b1-mt`	45	2023	decoder	machine translation, simplification, program synthesis, code datasets	reputed for strong cross-lingual generalization
`falcon-7b-instruct`	11	2023	decoder	instruct and chat datasets	multilingual in high-resource languages (`fra`)
`gemma-2b-it`	1	2024	decoder	web documents, code, math	English-trained model (study control)
`Llama-3.1-8B-Instruct`	8	2024	decoder	public online data	multilingual in mid- and high-resource languages

Table 24: Overview of the language models evaluated in this study, the number of languages each model supports, model release year, basic model architecture, datasets used to train and finetune the model, as well as the rationale for why the model was selected.