Beyond English: The Impact of Prompt Translation Strategies across Languages and Tasks in Multilingual LLMs

Itai Mondshine[a], Tzuf Paz-Argaman[a], and Reut Tsarfaty[a]

[a]Bar-Ilan University, Israel,
{mondshi1, tzuf.paz-argaman, reut.tsarfaty}@biu.ac.il

Abstract

Despite advances in the multilingual capabilities of Large Language Models (LLMs) across diverse tasks, English remains the dominant language for LLM research and development. This has led to the widespread practice of *pre-translation*, i.e., translating non-English task prompts into English before inference. *Selective pre-translation*, a more surgical approach, focuses on translating specific prompt components. However, its current use is sporadic and lacks a systematic research foundation. Consequently, the optimal *selective pre-translation* strategy for various multilingual settings and tasks remains unclear. In this work, we aim to uncover the optimal setup for *selective pre-translation* by systematically assessing its use. Specifically, we view the prompt as a modular entity, composed of four functional parts: instruction, context, examples, and output, either of which could be translated or not. We evaluate pre-translation strategies across 35 languages covering both low and high-resource languages, on various tasks including Question Answering (QA), Natural Language Inference (NLI), Named Entity Recognition (NER), and Abstractive Summarization. Our experiments show the impact of factors as similarity to English, translation quality, and the size of pre-trained data, on the model performance. We suggest practical guidelines for choosing optimal strategies in various multilingual settings.[1]

1 Introduction

Large language models (LLMs) demonstrate impressive capabilities across various natural language processing tasks, including machine translation (Kocmi et al., 2023), natural language understanding (Saba, 2024) and complex reasoning tasks (Huang and Chang, 2022). These exceptional capabilities of LLMs stem, to a large extent, from the vast amounts of data they were trained on (Kaplan et al., 2020). Current LLMs are primarily trained on English data but also include data from other languages, i.e., GPT-3 was trained on 119 languages, but only 7% of the tokens are from non-English languages.[2] With over 7,000 languages spoken worldwide (Anderson, 2010), the increasing pace of globalization has amplified the need for LLMs that understand and respond in diverse languages.

One common strategy to respond to a task presented in a language different than English is *pre-translation*, which involves translating the complete prompt into English before querying the model (Ahuja et al., 2023; Shi et al., 2022), allowing to leverage the robust capabilities acquired in English across different languages. At the same time, this approach introduces complexities and risks of information loss (Nicholas and Bhatia, 2023). Also, it is unclear whether this approach is uniformly effective across languages and tasks, especially tasks requiring region-specific or culturally-apropriate knowledge.

In contrast to pre-translation, recent studies show that *direct inference*, i.e., prompting the model directly in the (non-English) *source* language spoken by the user, outperforms pre-translation for tasks like QA (Intrator et al., 2024). However, it is unclear whether this approach is optimal, considering that the model was trained on limited data in the source language. It is also unclear how much information is shared across languages during pre-training. Be that as it may, as we show in Sec. 3.2.1, direct inference results still remain suboptimal.

In view of these shortcomings, various studies propose to use *selective pre-translation*, a more nuanced method compared to the de-facto standard

[1]We launched a user-friendly HuggingFace Space for generation and use of selective pre-translation prompts https://huggingface.co/spaces/naacl-anonymous/selective_pre_translation. Appendix D provides further details and illustrations.

[2]https://github.com/openai/gpt3/blob/master/dataset_statistics

Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025), pages 81–104

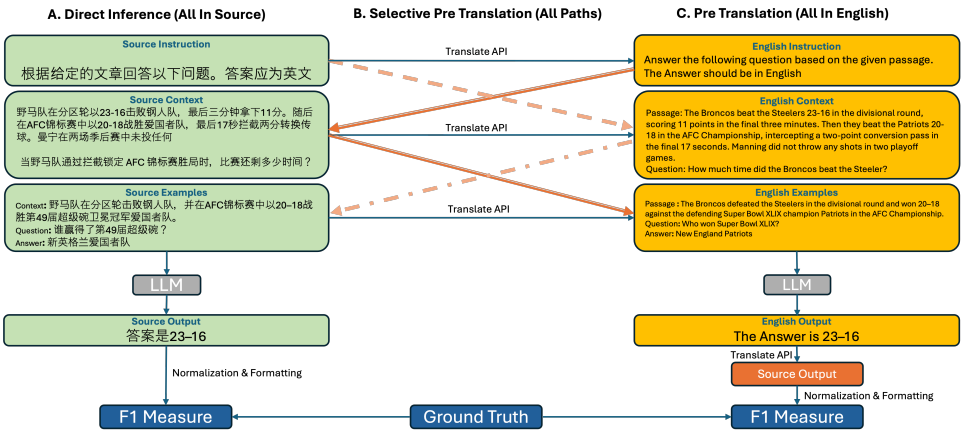


Figure 1: Prompting Strategies: *Direct Inference*, *Selective Pre-Translation*, and *Pre-Translation*

pre-translation approach, which calls for translating only specific parts of the prompt (Ahuja et al., 2023; Kim et al., 2023; Kim et al.). For example, Liu et al. (2024) show that translating only the context to English outperforms direct-inference in summarization and NLI. Ahuja et al. (2023) translated few-shot examples to English while keeping the context in the source language. Kim et al. (2023) used the selective approach when prompting different cross-lingual compositions of in-context examples. However, the selective approach lacks a systematic evaluation of more complex setups, e.g., instruction in English and output in the source language. Consequently, the efficacy of *selective pre-translation* and the optimal prompt configurations for various multilingual settings and tasks remain unclear. To fill this gap, in this paper we set out to examine the impact of selective pre-translation, a commonly used method, across diverse tasks, in order to devise effective prompting strategies for multilingual LLMs.

Concretely, we define a formal *configuration* for a prompt — consisting of four functional parts: instruction, context, examples, and output — either of which could be *selectively* pre-translated or not (see also Winata et al. (2021); Ahuja et al. (2023)). We exhaustively assess all configurations of cross-lingual prompt translation into English from different source languages. Figure 1 presents an overview of our approach, demonstrating the various *selective pre-translation* strategies compared against *pre-translation* and *direct inference* in the source language.

Through a comprehensive evaluation involving 35 languages, four tasks, six dataset collections, and three models, our results demonstrate that *selective pre-translation* consistently outperforms both *pre-translation* and *direct inference* in the source language, establishing the efficacy of *selective pre-translation* strategies (Section 3). Additionally, we analyze the considerations in determining which component to translate, and illustrate the optimal strategies across tasks and languages with varying resource levels. Moreover, we examine how factors such as language similarity to English, training size, and language script affect task performance, and show the effectiveness of selective pre-translation method in mitigating various translation issues, by choosing which prompt components to translate (Section 4).

More specifically, our findings demonstrate that in extractive tasks such as QA or NER, where the output overlaps with the provided context and no generation is needed, the model is either agnostic to the context language in the case of high-resource languages or prefers context in the source language in the case of low-resource languages. Surprisingly, we have discovered that low-resource languages yield better results even when the model's output is required in English, e.g., in NER (Section 3). Moreover, we show that translation quality significantly affects model performance and that the *selective pre-translation* approach essentially mitigates the negative effect of suboptimal translations, which are in turn specifically problematic in lower-resourced languages (Section 4).

All in all, our extensive and systematic evaluation of pre-translation strategies facilitates generalization across a broader range of languages and tasks, beyond the specific ones herein, towards more robust LLM-use in multilingual settings.

2 The Proposal: Formalizing Prompts *Selective Pre-Translation* Strategies

Current practices of *prompting* generative LLMs, such as the GPT models family (Ouyang et al., 2022) and Gemini (Team et al., 2023), uncover two remarkable capabilities in performing language processing tasks: (i) *chain of thought* (Wei et al., 2022), where LLMs solve complex tasks through a series of intermediate reasoning steps, and (ii) *in-context learning* (Brown et al., 2020), allowing the model to adapt to new tasks based on limited examples, without weights updates. These capabilities are built on top of the notion of a *prompt*, which serves as a prefix for the LLM's response. These capabilities are powered by the complex nuanced structure of nowadays prompts, consisting of four components: *instruction, context, examples*, and *output*.

Let us first define these four components, as follows. The ***Instruction*** (I) provides a natural language guidance to the model, explaining the task to be performed. The ***Context*** (X) represents the task data that the model operates on in performing the task. ***Examples*** (E) are optional illustrations of context:output pairs, that can be used for in-context learning. Overall, we define $\langle I\ X \rangle$ as a zero-shot prompt and a few-shot prompt as $\langle I\ X\ E \rangle$. The prompt is processed by a model M to yield an ***Output*** (O), where the instruction can include a request for the model to generate the output in a specific language or format.

Each component, i.e., the instruction, the context, the examples, or the output, may be pre-translated or not. We denote a pre-translation decision $l \in \{e, s\}$, where e stands for English pre-translation and s for the source language. Standardly the prompt is composed as $\langle I\ X\ E \rangle$ and is delivered to the model M, which in turn emits an output O. We define a specific *pre-translation configuration*[3] as $c = \langle \mathrm{I}^{l_i}, \mathrm{X}^{l_x}, E^{l_e}, O^{l_o} \rangle$ where the subscript $l \in \{e, s\}$ indicates the language of the component. Having defined the pre-translation configurations, we evaluate them in different settings.

3 Selective Pre-Translation Evaluation

3.1 Experimental Setup

Goal We set out to compare *selective pre-translation* to both *pre-translation* and *direct inference*, and to assess the impact of the selected configuration on task performance across languages.

Affinity	Class	*Range (% of tokens)*	*Avg. #tokens (M)*	*STD*
High Resource	A	$p \geq 0.1\%$	1,240	1,156
Medium Resource	B	$0.01\% < p < 0.1\%$	72	49
Low Resource	C	$0\% < p \leq 0.01\%$	5.07	5.41
Unrepresented	D	$p = 0\%$	0	0

Table 1: Language categorization based on the percentage (p) of tokens per language in GPT-3's training data. Avg. token count (millions), STD: standard deviation.

Prompt Configuration We assess *selective pre-translation* in both zero-shot and few-shot settings. In the zero-shot settings, with no examples, we considered 2^3 configurations. For the few-shot scenario, with four components, each is either translated to English or retains the source language, we get 2^4 configurations. All in all we experiments with 24 configurations per language and task.[4]

Prompt Creation And Output Normalization Based on the prompt configuration, we used the Google Translate API[5] to translate the components that required translation. After querying the model, we normalized and formatted its output, then translated it to match the language of the gold standard. See Appendix A.1 for further implementation.

Models We conducted experiments on several LLMs: (1) Standard generative models—GPT-3.5-turbo (Ouyang et al., 2022), Mistral-8x7B (Jiang et al., 2023), and Gemini-1.0-pro (Team et al., 2023) — with context sizes of 16k, 32k, and 8k, respectively; (2) multilingual - bloomz-7b1-mt (Muennighoff et al., 2022), with a 2k context.[6]

Language Selection and Categorization We selected ~11 languages per task, ensuring a balanced representation across resource levels (high, medium, low). Due to the lack of precise pre-training distribution for the LLMs we use, we employed the GPT-3 distribution as a proxy, as it is the only distribution publicly shared, to our knowledge.[7] The GPT-3's multilingual coverage enables us to categorize languages into classes based on their data ratios. Following Lai et al. (2023), we categorized the tested languages into four classes based on data ratio: High-Resource (A), Medium-

[3] See Appendix A.2 for specific configuration examples.

[4] NLI has 12 configurations, with output always in the instruction language, due to its particular, fixed, output format.

[5] pypi.org/project/easygoogletranslate/

[6] See Appendix A.1 for details on the models we used.

[7] https://github.com/openai/gpt-3/blob/master/dataset_statistics

Task	*Dataset*	*Languages*
NLI	XNLI	***High***: Spanish, German, Chinese ***Medium***: Greek, Turkuish, Arabic ***Low***: Bulgarian, Hindi, Thai, Swahili, Urdu
QA	XQuAD	***High***: German, Russian, Romanian ***Medium***: Arabic, Greek, Vietnamese
	IndicQA	***Low***: Hindi, Malayalam, Bengali, Telugu ***Unrepresented***: Assamese
NER	MasakhaNER	***Unrepresented***: Bambara, Ese, Hausa, Yoruba
	WikiANN	***High***: French, Chinese, Italian, Portuguese, Swedish ***Medium***: Serbian, Slovak
Summarization	XL-Sum	***High***: French, Japanese, Spanish, Portuguese ***Medium***: Korean, Turkish, ***Low***: Azerbaijani, Nepali, Persian, Uzbek

Table 2: Experiment Setup: Tasks, datasets, languages. Languages are separated by their resource-type affinity.

Resource (B), Low-Resource (C), and Unrepresented (D).[8] Class D, which includes languages unseen during training. Table 1 summarizes this classification with basic properties.[9]

Tasks and Datasets We assess model performance on 4 tasks, NLI, QA, NER, and Summarization, which we detail in turn. (i) ***Natural Language Inference (NLI)*** determines if a hypothesis entails, contradicts, or is neutral to a premise. We use the XNLI dataset (Conneau et al., 2018), with sentence pairs in 11 languages, and measured prediction accuracy. (ii) ***Question Answering (QA):*** We focus on extractive QA, where the model identifies the answer span in a given context. We evaluated performance on XQuAD (Artetxe et al., 2019) and IndicQA (Doddapaneni et al., 2022) for Indic languages, using the F1 score to assess performance. (iii) ***Named Entity Recognition (NER):*** We sampled languages from two datasets: WikiANN (Pan et al., 2017), which includes Wikipedia sentences annotated with Location, Person, and Organization tags in 176 languages; and MasakhaNER (Adelani et al., 2021), for African languages. While both datasets use the BIOSE scheme to delineate entity boundaries, we recast the task as generative, prompting the model to generate the list labeled named entities for a given input context. Model performance has been evaluated using F1 scores. (iv) ***Abstractive Text Summarization*** involves generating short summaries of long contexts, rather than extracting existing sentences. We used the XL-Sum dataset (Hasan et al., 2021), which offers news article summaries in diverse languages. We sampled 10 languages from the dataset and evaluated with ROUGE. We conducted experiments on a sample of 250 examples[10] from the test sets for each language.[11] In total, the datasets we use encompass 35 languages across 4 tasks. Table 2 lists the datasets used, covering ~11 languages per task, ensuring a balanced representation of {Low,Mid,High} resource categories for each task.

Analysis Methods To analyze the empirical results and detect the most influential components, we use three methods: (i) *Correlation analysis* – Assessing the relationship between the model's prediction scores and the language selection per component. (ii) *Association Rule Learning (ARL)* and *the Apriori algorithm* – While correlation analysis provides a preliminary understanding of the relationship between individual components and model performance, it does not capture non-linear relationships, i.e., the combined effect of multiple translation decisions on performance. To address this limitation, we utilize ARL with the Apriori algorithm (Piatetsky-Shapiro, 1991).[12] (iii) *Performance Gap* – We computed the average difference of k configuration pairs c_i and c_j such that they differ only in the language of one component, e.g., $\langle I^{\mathbf{e}}, X^e, E^e, O^e\rangle$ and $\langle I^{\mathbf{s}}, X^e, E^e, O^e\rangle$. We then calculated this average to determine the performance gap for each specific task: $\frac{1}{k}\sum_{\langle i,j\rangle=1}^{k}(\text{Eval}(c_i) - \text{Eval}(c_j))$, where Eval() denotes the task evaluation score, and k is the number of distinct pairs.

3.2 Results

In Section 3.2.1, we present the results of *selective pre-translation* demonstrating their advantage over both *direct inference* (source language only) and *pre-translation* (English only). Subsequently, in Section 3.2.2, we identify the optimal configurations for each task and analyze the impact of each component on the overall performance, emphasizing key considerations for effective prompting. We start off with GPT-3.5-Turbo and proceed to verify that our results generalize to other models.

3.2.1 Selective Pre-Translation Advantage

Table 3 shows each language's highest-performing configuration score among all 24 distinct configurations.[13] Additionally, we display the improvement

[8] See Table 12 for list of languages, codes and data ratios.

[9] Alternative criteria such as speakers ratio, as proposed by Joshi et al. (2020), do not reflect language diversity in LLMs, which is affected by availability of data rather than speakers.

[10] We selected 250 that followed the can fit into the context of the model, i.e., $< 16K$.

[11] For tasks without public test sets (XQUAD, IndicQA), we used the validation data.

[12] Appendix A.3.1 details the algorithm and implementation.

[13] Appendix E displays the full-fledged table of results.

Question Answering (QA, F1)					Summarization (ROUGE)					Named Entity Recognition (NER, F1)					Natural Language Inference (NLI, Acc.)				
Lng	*Cls.*	↑*Top*	*Src.* (%)	*Eng.* (%)	*Lng*	*Cls.*	↑*Top*	*Src.* (%)	*Eng.* (%)	*Lng*	*Cls.*	↑*Top*	*Src.* (%)	*Eng.* (%)	*Lng*	*Cls.*	↑*Top*	*Src.* (%)	*Eng.* (%)
en	A	0.77	N.A	N.A	en	A	30.23	N.A	N.A	en	A	0.65	N.A	N.A	en	A	0.69	N.A	N.A
de	A	0.85	18%	9%	fr	A	35.12	16%	10%	sr	B	0.77	52%	265%	sw	C	0.73	58%	28%
hi	C	0.82	32%	182%	ja	A	32.47	17%	14%	it	A	0.75	9%	41%	bg	C	0.72	57%	8%
ar	B	0.74	84%	138%	fa	C	29.34	21%	0%	sk	B	0.72	15%	36%	el	B	0.71	24%	30%
vi	B	0.73	0%	58%	es	A	28.28	10%	3%	po	A	0.72	18%	20%	es	A	0.69	20%	18%
ro	A	0.69	0%	9%	po	A	27.40	8%	0%	fr	A	0.72	23%	24%	ar	B	0.67	28%	23%
ru	A	0.69	6%	305%	tr	B	20.87	18%	0%	hau	C	0.70	62%	51%	hi	C	0.64	59%	8%
el	B	0.69	0%	2%	ne	C	19.58	31%	28%	ee	D	0.68	46%	81%	de	A	0.64	19%	8%
bn	D	0.68	44%	423%	as	D	15.79	17%	7%	sv	A	0.68	12%	9%	zh	B	0.63	16%	4%
as	D	0.56	138%	450%	uz	C	15.72	58%	24%	zh	B	0.63	90%	121%	th	B	0.57	49%	10%
te	C	0.53	210.10%	253.30%	ko	B	11.84	36.99%	11.78%	bam	D	0.33	33.25%	80.24%	ur	C	0.57	29.96%	9.08%
ml	C	0.49	104.30%	600.00%						yor	D	0.32	66.02%	49.19%	tr	B	0.57	0.00%	8.14%

Table 3: For each Language, we present the top-performing selective configuration score over all other configurations (*Top*) along with its relative improvement (%) over *direct inference* (*Src*) and *pre-translation* (Eng).

QA						Summarization						NER						NLI				
lng	***cls.***	***instruction***	***context***	***examples***	***output***	***lng***	***cls.***	***instruction***	***context***	***examples***	***output***	***lng***	***cls.***	***instruction***	***context***	***examples***	***output***	***lng***	***cls.***	***instruction***	***context***	***examples***
ru	A	-0.08**	**0.35****	0.12**	0.09**	ja	A	**-0.33****	-0.08	-0.02*	0.00	fr	A	-0.11*	0.10*	-0.01	0.01	de	A	-0.03	-0.02	-0.01
de	A	-0.03**	**0.30****	0.08	0.03*	fr	A	0.01	0.020	-0.04	0.06	it	A	0.02	0.04	-0.04	0.01	es	A	-0.03	0.02	-0.03*
ro	A	-0.03	0.12**	0.04	0.02	po	A	-0.08*	0.05*	-0.03*	0.10*	po	A	-0.15	0.09*	0.1	0.01	el	B	-0.04	0.01	0.07
vi	B	0.04	**0.40****	0.10**	0.10	es	A	-0.09*	0.03*	-0.03	0.05	sv	A	-0.11*	0.06*	-0.03**	0.01	zh	B	0.01	-0.06	-0.06
ar	B	-0.07**	**0.20****	0.13**	0.04*	tr	B	-0.14**	0.10	-0.1	-0.03*	zh	B	-0.26**	**0.44****	0.00	0.07	ar	B	0.00	-0.02	-0.04
el	B	-0.06	**0.48****	0.03	0.07*	ko	B	-0.10**	0.13	0.01	0.05	sr	B	-0.26**	**0.44****	0.09**	0.05	th	B	-0.03	0.03	-0.14*
bn	C	-0.10**	**0.38****	0.03	0.03	uz	C	**-0.42****	0.14	0.03	-0.12*	sk	B	-0.11**	**0.30****	-0.1*	0.01	tr	B	-0.02	0.00	0.02
ma	C	-0.14**	**0.30****	0.01	0.03	fa	C	**-0.37****	0.05	-0.07**	-0.04	bam	D	0.03	**0.44****	-0.11	0.02	ur	C	0.01	0.01	-0.08*
te	C	-0.10**	**0.38****	0.03	0.03	ne	C	**-0.35****	-0.09	0.07**	-0.14	ewe	D	-0.01	**0.38****	-0.12**	0.01	bg	C	0.01	0.05	-0.13*
hi	C	-0.07**	**0.30****	0.05	0.01	az	C	**-0.30****	0.04	-0.00	-0.05	yo	D	-0.01	**0.36****	0.01*	0.03	sw	C	0.12	-0.06	-0.09
as	D	-0.04**	**0.30****	0.06	0.06							hau	D	-0.04	**0.30****	0.08*	0.02	hi	C	-0.03	-0.09	-0.09**

Table 4: For each language, we present the Point-biserial correlation (τ) between the individual component's language selection (English/Source), and the model performance score across all the configurations samples that use it. Positive $|\tau|$ values correlate with the source language, and negative $|\tau|$ values correlate with English. Significant correlations are indicated by *p < 0.05 and **p < 0.01. Bold values denote correlations ($|\tau| > 0.3, p < 0.01$).

Resource	Model	QA				NER				Summarization				NLI		
Component		I.	X.	E.	O.	I.	X.	E.	O.	I.	X.	E.	O.	I.	X.	E.
High	GPT	N	S	S	S	N	S	S	S	S	S	N	N	N	S	E
	Gemini	S	S	S	S	N	S	S	S	E	E	Z	N	N	N	E
	Mixtral	N	S	S	S	N	S	S	S	S	S	Z	S	N	S	S
Low	GPT	N	S	S	S	N	S	S	E	E	E	S	E	N	E	S
	Gemini	S	S	S	S	E	S	S	E	S	S	Z	N	N	N	E
	Mixtral	N	S	S	S	E	S	S	E	E	E	E	E	N	S	E
High	Bloomz	S	S	S	S	S	S	S	S	E	E	E	E	E	E	E
Low	Bloomz	S	S	S	S	S	S	S	S	E	E	E	E	E	E	E

Table 5: The Top-performing configurations based on Apriori-based Association Rules for High/Low resource level. Comparing Standard LLMs (GPT/Gemini/Mixtral) and Multilingual LLM. $confidence > 0.8, support > 0.15$. S / E - source/English language, Z - zero-shot (no examples), N - neutral (same performance for English/Source).

(%) over *direct inference* and *pre-translation* for each language. The results indicate that 92% of the tested languages show an improvement over the basic *pre-translation* configuration. Particularly for low-resource languages like Malayalam and Telugu, the gains with *selective pre-translation* are substantial, exceeding 200% in relative improvement. Overall, when comparing selective pre-translation to *complete pre-translation*, the average improvement in low-resource languages is 65% greater than the average improvement in high-resource languages.

The results further reveal that 90% of the languages show improvement over basic *direct inference*. Similar to the pre-translation approach, low-resource languages like Telugu and Assamese demonstrate a relative improvement of over 100%. High-resource languages also show impressive improvement, albeit smaller, e.g., French and Portuguese show an improvement of over 20% in NER.

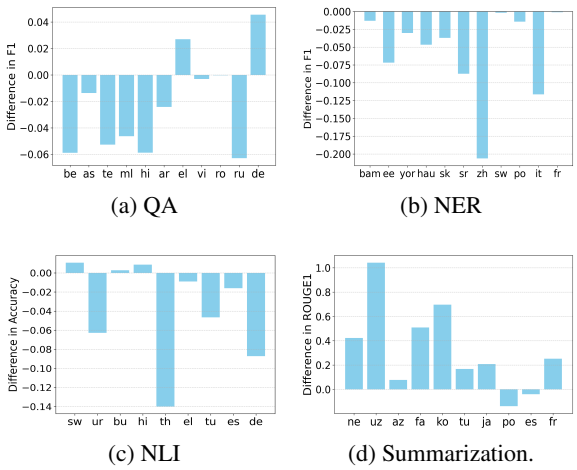
Overall, the table shows that *selective pre-translation* can outperform both *pre-translation* and *direct inference*, particularly for languages considered low-resource during pre-training.

3.2.2 The Holy Grail of Optimal Configuration

Having established the advantage of *selective pre-translation* in general, we now study the effects of component language selection on model performance and provide general guidelines for multilingual scenarios.[14] Table 4 shows the Point-biserial correlation between individual component selection and model performance for all the 24 configurations per language/task.[15] Table 5 presents the top-performing configurations, based on the highest-scoring apriori rules for multiple component selections, henceforth *optimal configurations*.

[14]For the instruction component language, except for a slight preference for English as demonstrated in Table 4, we did not observe a strong affinity for any language selection.

[15]We calculate the correlation between performance and a binary vector indicating whether the component is in English.



(a) QA (b) NER

(c) NLI (d) Summarization.

Figure 2: Performance Gap Analysis for the Examples language (English minus Source). Left to right X-axes order indicates Low to High Resource Level. Y-axes indicate language preference: positive values for the source language, and negative for the English language.

Context Language Impact Table 4 shows that source language selection correlates most with model performance score in extractive tasks, such as QA (average of 0.33) and NER (average of 0.32), particularly for low-resource languages (Class C/D), which demonstrate a 70% higher correlation coefficient compared to high-resource languages. In contrast, in tasks like abstractive summarization and NLI, we found no correlation (average of 0.05) to context language selection. Our rule-association analysis in Table 5 further underscores the importance of source-language context in extractive tasks, especially with low-resource languages.

Examples Impact In general, the top-performing configurations of the GPT model in Table 5 show that the optimal configurations are those that include examples, i.e., a few-shot rather than zero-shot setup (Appendix A.4.1 further supports incorporating examples in prompts, especially for high-resource languages). Concretely concerning the language selected for the examples, the optimal configurations in Table 5 show that extractive tasks as NER perform better with source-language examples, possibly due to NER's dependence on region-specific or culturally-relevant knowledge. Also, the performance gap analysis in Figure 2, shows that, for extractive tasks, prompts with examples in the source language perform better than those with English examples, especially for low-resource languages (See 2(a)/2(b)).

Output Language Impact Unlike context and examples, the output depends on the model generations's grammaticality and fluency. The best-performing configurations for the GPT model in Table 5 indicate that for extractive tasks, source-language output is beneficial across all languages. Interestingly, despite context mismatches, NER in low-resource languages also benefits from English output. For generative tasks such as summarization, model output in English performs better due to the model's stronger capabilities in English, *even though* we back-translate the output to source prior to evaluation. Thus, while it is fine in such generative tasks to instruct the model to generate outputs in the source language for high-resource languages, it appears better to generate in English in the low-resource case.

3.3 Beyond Configuration: Key Factors

Having analyzed the impact of the components' language selection, we discuss key additional factors influencing the efficacy of our approach.[16]

Pre-Training Data Size Impact Table 3 presents the optimal prompt configuration scores per language and task. For QA, summarization, and NER, the general trend indicates that even for the optimal pre-translation configuration, classes A and B (High-Medium resource) achieve better results than classes C and D (Low resource), However, a few exceptions exist, i.e., in Hausa and Ewe (Class C/D) we see better results on the NER task compared to Swedish and Chinese (Class A/B). For the NLI task we found no trend where a class C/D languages outperform A/B languages. So, while pre-trained data distribution matters, selective pre-translation can help low-resource languages match the results of higher-resource ones in specific tasks.

Linguistic Similarity to English Impact We used pre-computed syntactic similarities to English from the URIEL dataset (Littell et al., 2017) and calculated the Pearson correlation for each task between the best-performing configuration scores (top score per language) and the syntactic similarity of these languages to English. A moderate correlation (0.42) for summarization shows that syntactic similarity to English positively correlates with performance. NER also shows moderate correlations, suggesting models better identify entities

[16] See also Appendix A.4.3 for script impact.

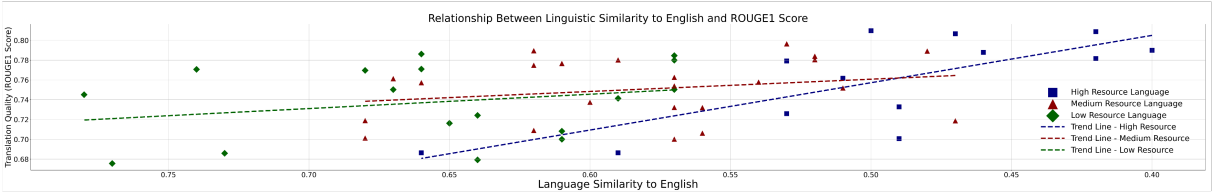


Figure 3: Scatter plot showing the relationship between syntactic similarity to English (further right is more similar) and translation quality (ROUGE) for four language resource subsets (represented as distinct four colored shapes). Each dot represents a different language. Positive linear regression shows an upward trend.

when texts share syntactic features with English. Appendix A.4.3 further details these results.

Standard Model Impact In the previous section, we assessed the selective pre-translation strategies using the GPT model. In this section we check whether these strategies generalize to other LLMs. Table 5 displays the optimal configurations per task and language for Gemini and Mixtral. We see that the *preference for source language in extractive tasks* (for context, examples, and output) holds across all three models. Additionally, outputting in English while keeping the *context in the source language for NER* in low-resource languages is consistent. In NLI, models are agnostic to instruction language. However, surprisingly, in abstractive summarization, we found no clear pattern.[17]

Multilingual Model Impact In addition to the standard LLMs, we evaluated BLOOMZ-7b1-mt, known for its multilingual capabilities (Muennighoff et al., 2022). Table 5 displays the optimal configurations for BLOOM across all tasks. We found no distinction between resource types for this model. As shown, the preference for source language in QA is relevant here as well. Interestingly, NER can be answered in the source language, highlighting its multilingual strength. However, for generative tasks and NLI, this multilingualism diminishes, as the model tends to favor English prompts. Overall, the top-performing configurations for BLOOMZ indicate that it performs better with single-language prompts rather than with selective pre-translation prompts.

4 Translation: Key to Pre-Translation

In the previous Section we examined various factors affecting model performance for selective pre-translation strategies. Since translation forms the foundation of selective pre-translation, this Section

[17] See Appendix E for additional results.

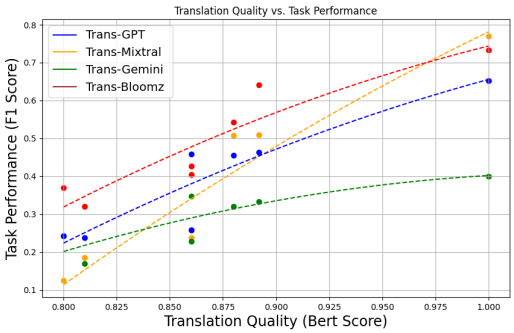


Figure 4: Correlation between translation quality (BERTScore) and accuracy (F1) for Pre-Translation-Zero-shot prompting, each dot is a different language.

focuses on a key question: Are these factors primarily due to the limitations of LLMs, or are due to the quality of the pre-translations themselves? To address this, we first isolate the impact of these factors on translation quality through a controlled experiment. Subsequently, we investigate how translation quality, independently of other factors, influences downstream tasks in our setup.

4.1 Experimental Setup

Factors Affecting Translation Quality We evaluated Google Translate performance on the FLORES-200 validation set (Guzmán et al., 2019), analyzing 91 languages across all resource levels, each with 997 sentences paired with their English translations. We compared machine translations to human-generated references using: (1) n-gram matching metrics – Meteor (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), and BLEU (Papineni et al., 2002), and (2) neural network-based evaluation metrics – BertScore (Zhang et al., 2019) and Comet (Rei et al., 2020). Additional experiments using other translation models are in Appendix C.1.

Impact on Downstream tasks We used the QA dataset XQuAD, with 300 parallel sentences in English and other languages. We compared the

zero-shot *pre-translation* using output in the source language with the *selective pre-transaltion* optimal approach using GPT-3.5-Turbo (0125).

4.2 Results

Factors Affecting Translation Quality Our analysis focuses on the factors influencing model performance discussed in Sec. 3 and we explore their impact on translation quality. We focus on two factors: ***language resource levels (high/low)*** and ***linguistic similarity to English***. First, we found that the average quality for high-resource languages was higher compared to low-resource languages (0.75 vs 0.73), although correlation between resource level and quality wasn't significant. As for *linguistic similarity to English* we used the pre-computed linguistic similarities from the URIEL dataset (Littell et al., 2017)[18] and calculated the correlation between syntactic similarity to English and the translation quality for each language. Figure 3 shows a positive correlation (coefficient = 0.33, p-value = 0.01) between syntactic similarity to English and translation quality (ROUGE-1). This correlation is particularly strong for high-resource languages (coefficient=0.73, p-value=0.004). Additional correlation results are detailed in Appendix C.2.

Impact On Downstream Tasks To isolate translation quality as the sole direct factor influencing model performance, we translated the entire prompt into English. Since the original input differed only in language, not content, any variation in the processed input can be attributed solely to the quality of the translation. This approach allows us to directly measure the correlation between translation quality and model performance across various tasks. Figure 4 shows the correlation between translation quality (BERTScore) and model performance (accuracy) for each language. Our results show that *higher translation quality goes hand in hand with improved task performance*. The overall Pearson correlation is 0.233 ($p < 0.001$). However, when assessing the same tasks with *selective pre-translation* instead of a *completely pre-translated* prompt, we found a low correlation of 0.05 ($p< 0.05$) between the translation quality and task performance, while selective pre-translation outperforms the fully translated prompt. This disparity shows that the *selective pre-translation* method effectively neutralizes translation issues. By strategically choosing which prompt components to translate, we can make pre-translation useful for languages with lower translation quality.

In sum, our findings demonstrate that factors influencing downstream tasks, such as high resource level and similarity to English, are positively correlated with translation quality. We further show that *selective pre-translation* can mitigate the negative effects of poor translation quality. These two findings underscore the importance of investing in high-quality translation, and on the other hand, prioritizing the *selective pre-translation* approach in languages where machine translation is sub-optimal.

5 Related Work

With over 7,000 languages spoken globally (Anderson, 2010), the growing use of diverse languages have fueled the demand for multilingual LLMs. Progress in this field stems from two primary efforts: (1) developing dedicated monolingual models for low-to-medium-resource languages (Seker et al., 2022; Cui et al., 2023; Andersland, 2024), and (2) creating multilingual LLMs with pre-trained data encompassing multiple languages (Qin et al., 2024; Jiang et al., 2024).

The ability of Multilingual LLMs to operate in different languages (Raffel et al., 2020; Conneau et al., 2019; Chowdhery et al., 2023) comes from two sources: (1) training or fine-tuning on multilingual data in order to achieve multilingual proficiency (Xue et al., 2020; Chen et al., 2021; Le Scao et al., 2023; Shaham et al., 2024; Muennighoff et al., 2022), and (2) utilizing prompting techniques to harness the model's inherent multilingual capabilities without modifying parameters during inference (Brown et al., 2020). This latter approach has gained popularity due to its efficiency and applicability to a wider range of use cases.

For the latter, to improve the multilingual capabilities of LLMs researchers developed various prompting methods. Huang et al. (2023) introduced XLT, a cross-lingual prompt that directs LLMs to function as experts in a specific language through a process involving problem-solving and cross-lingual thinking. Zhao and Schütze (2021) employed discrete and soft prompting techniques and showed that few-shot non-English prompts outperform finetuning in cross-lingual transfer. Shi et al. (2022) found that chain-of-thought (CoT) prompts lead to multilingual reasoning abilities in LLMs, even in under-represented languages. An-

[18] https://github.com/antonisa/lang2vec

other strategy is *pre-transaltion* which translates the entire prompt to English (Chowdhery et al., 2023; Qin et al., 2023; Ahuja et al., 2023). A more nuanced approach, *selective pre-translation*, translates part of the prompt into English, for instance, Liu et al. (2024) translated only the instruction, and Ahuja et al. (2023) translated the few shot examples. While these use cases lack a systematic research foundation, in this study, we systematically study pre-translation configurations to provide evidence-based recommendations for optimal use.

6 Conclusion

In this work we formalize and comprehensively assess selective pre-translation prompting strategies for LLMs in multilingual settings. With four tasks, six dataset collections, three models, and 35 languages, we deliver the first systematic evaluation, to our knowledge, of all existing prompt configurations of pre-translation. We demonstrate that *selective pre-translation* consistently outperforms both *pre-translation* of the entire prompt and *diret-inference* in the source language, establishing the efficacy of *selective pre-translation* in both the high- and low-resource cases. Additionally, we show that translation quality significantly affects performance and that selective pre-translation can mitigate the negative effects of suboptimal translations.

Limitations

Subset of LLMs This study aims to systematically assess the effectiveness of various prompting strategies across different tasks and LLMs. Due to resource limitations, it was not possible to evaluate more advanced models such as GPT-4 or GPT-4o. However, we endeavored to cover several LLMs representing different architectures. Additionally, the choice of Bloom as our multilingual model is based on previous works (Bawden and Yvon, 2023; Nezhad and Agrawal, 2024). We make our evaluation framework, code, configurations, and execution pipeline, for open public use, allowing to extend the investigation to more and newer models.

LLM Adherence and Impact on the Output In our evaluation, we attempted to influence the output by instructing the model to generate a response in a specific language. However, the model occasionally did not follow these instructions, producing output in a different language, which could impact the results. Appendix B provides error analysis of the various issues we encountered.

Evaluation Metrics based on n-gram matching, such as ROUGE (Lin, 2004), are commonly used for evaluating summarization quality in English. However, these metrics can be problematic when applied to morphologically rich languages (MRL) such as Persian, which have more flexible word order compared to English. Additionally, their morphological richness means that the same concept can be expressed in multiple ways due to variations in prefixes, suffixes, and root conjugations.

Translation Quality's Impact on Downstream tasks Our analysis of the impact of translation quality impact on downstream tasks in section 4.2 was constrained by the scarcity of datasets with parallel splits for English and other languages, limiting our evaluation to the QA task. Future research should incorporate a wider array of datasets and tasks to validate and expand upon our findings.

Pretrained Data Distribution Details In our experiments, we evaluated four models and grouped the languages based on GPT-3's pre-training data distribution information. Ideally, we would split the languages according to each model's data distribution. However, to our knowledge, only GPT-3's pre-training data distribution is publicly shared. Explicitely testing different language distributions is desired but resource intensive, and is left for future research.

Acknowledgements

This research has been funded by a grant from the Israel Science Foundation (ISF) grant number 670/23 as well as a grant from the Israeli Ministry of Science and Technology (MOST), and a KAMIN grant from the Israeli Innovation Authority, for which we are grateful. We are further grateful for a generous VATAT grant to the BIU NLP team which contributed resources for computation, annotation and human evaluation in this project.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. 2021. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. 1993. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216.

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

Michael Andersland. 2024. Amharic llama and llava: Multimodal llms for low resource languages. *arXiv preprint arXiv:2403.06354*.

Stephen R Anderson. 2010. How many languages are there in the world. *Linguistic Society of America*, pages 1–12.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of bloom. *arXiv preprint arXiv:2303.01911*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Guanhua Chen, Shuming Ma, Yun Chen, Li Dong, Dongdong Zhang, Jia Pan, Wenping Wang, and Furu Wei. 2021. Zero-shot cross-lingual transfer of neural machine translation with multilingual pretrained encoders. *arXiv preprint arXiv:2104.08757*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2022. Indicxtreme: A multi-task benchmark for evaluating indic languages. *arXiv preprint arXiv:2212.05409*.

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

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. *arXiv preprint arXiv:2106.13822*.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. *arXiv preprint arXiv:2305.07004*.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Yotam Intrator, Matan Halfon, Roman Goldenberg, Reut Tsarfaty, Matan Eyal, Ehud Rivlin, Yossi Matias, and Natalia Aizenberg. 2024. Breaking the language barrier: Can direct inference outperform pre-translation in multilingual llm applications? *arXiv preprint arXiv:2403.04792*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

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*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Sunkyoung Kim, Dayeon Ki, Yireun Kim, and Jinsik Lee. Translating qa is enough: A key to unlocking in-context cross-lingual performance. In *ICML 2024 Workshop on In-Context Learning*.

Sunkyoung Kim, Dayeon Ki, Yireun Kim, and Jinsik Lee. 2023. Boosting cross-lingual transferability in multilingual models via in-context learning. *arXiv preprint arXiv:2305.15233*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, et al. 2023. Findings of the 2023 conference on machine translation (wmt23): Llms are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2024. Is translation all you need? a study on solving multilingual tasks with large language models. *arXiv preprint arXiv:2403.10258*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Sina Bagheri Nezhad and Ameeta Agrawal. 2024. What drives performance in multilingual language models? *arXiv preprint arXiv:2404.19159*.

Gabriel Nicholas and Aliya Bhatia. 2023. Lost in translation: Large language models in non-english content analysis. *arXiv preprint arXiv:2306.07377*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1946–1958.

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.

Gregory Piatetsky-Shapiro. 1991. Discovery, analysis, and presentation of strong rules. *Knowledge Discovery in Data-bases*, pages 229–248.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers. *arXiv preprint arXiv:2404.04925*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Walid S Saba. 2024. Llms' understanding of natural language revealed. *arXiv preprint arXiv:2407.19630*.

Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Greenfeld, and Reut Tsarfaty. 2022. Alephbert: Language model pre-training and evaluation from sub-word to sentence level. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46–56.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. *arXiv preprint arXiv:2401.01854*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. *arXiv preprint arXiv:2109.07684*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Mengjie Zhao and Hinrich Schütze. 2021. Discrete and soft prompting for multilingual models. *arXiv preprint arXiv:2109.03630*.

A Selective Pre-Translation Evaluation

A.1 Experimental Setup

Models To query GPT-3.5-turbo (0125), we used the Azure platform via the API[19]. For Mixtral-8x7B-287 Instruct-v0.1, we utilized the API platform provided by Together.ai[20]. For Gemini-1.0-pro, we accessed the API through Google AI Studio[21]. Lastly, for bloomz-7b1-mt, we used deployed the model on Hugging Face[22]

[19] https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models

[20] https://www.together.ai/

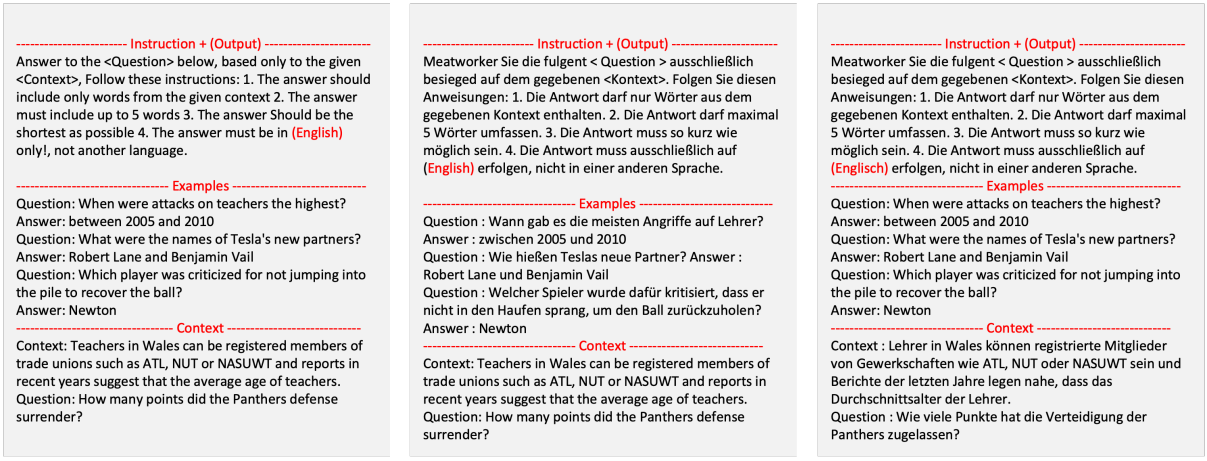
[21] https://aistudio.google.com

[22] https://huggingface.co/bigscience/bloomz-7b1-mt

All Configurations

Instruction	*Context*	*Examples*	*Output*
Source	Source	-	English
English	Source	-	English
Source	Source	Source	English
English	English	English	Source
Source	English	Source	English
English	Source	Source	Source
English	Source	-	Source
Source	Source	Source	Source
English	Source	English	Source
Source	Source	-	Source
English	English	Source	Source
English	Source	English	English
English	English	-	English
Source	English	-	English
Source	English	-	Source
Source	Source	English	English
English	English	-	Source
English	Source	Source	English
Source	Source	English	Source
Source	English	Source	Source
English	English	English	English
Source	English	English	English
English	English	Source	English
Source	English	English	Source

Table 6: All Valid Configurations (24)



(a) <English,English,English,English> (b) <German,English,German,English> (c) <German,German,English,English>

Figure 5: Examples of 3 configurations of German. Each configuration is in the following format <*Instruction,Context,Examples,Output*>

.

Prompt Creation For constructing the prompts we used the LangChain library[23] which enables us to build and validate prompts dynamically for both zero-shot and few-shot templates. For creating the instructions, we initially used ChatGPT to generate them and then fine-tuned them based on quality analysis from our experiments.

Normalization And Formatting Before evaluation, we normalized the model's output, with each task following a unique normalization process. For the QA task, for example, we converted the text to lowercase and removed punctuation, articles, and extra whitespace. In the Summarization task, we removed prefixes like 'The Summary:'. For the NER task, we converted the model's output into a list of tuples, each in the format (Tag, Entity). After normalization, additional formatting was applied if necessary. For instance, in the NER task, we transformed the normalized output into a list in the BIOSE format, identifying the entities in the original sentence and converting each entity prediction to its correct format based on its position (e.g., B-ORG for the first entity tagged as 'ORG').

A.2 Configuration Format

We define a specific *selective pre-translation configuration* as $C_i = \langle \mathrm{I}^l, \mathrm{X}^l, E_n^l, O^l \rangle$, $n \geq 0$, $l \in \{e, s\}$. Each configuration contains 4 components: instruction, context, examples, and output. Figure 5 displays examples for 3 configurations in the German language. See Table 6 for a list of all the configurations.

Python Libraries In Use For evaluation of the different models, we used the most common ROUGE package for non-English papers[24]. for loading and processing the data, we used NumPy[25] For help with writing the code, we used assistance from ChatGPT.

A.3 Analysis Methods

A.3.1 Rule Association And Apriori Algorithm

Association rule mining, one of the most important and well-researched techniques of data mining, was first introduced by Agrawal et al. (1993). It aims to extract interesting correlations, frequent patterns, associations, or casual structures among sets of items in the transaction databases or other data repositories.

Apriori algorithm The Apriori algorithm is a popular approach for mining association rules. It works by identifying frequent itemsets, which are groups of items that appear together in a dataset with a frequency above a specified threshold. The algorithm then generates association rules from these frequent itemsets, highlighting the likelihood of one item being present given the presence of another item. Apriori uses a bottom-up approach,

[23] https://pypi.org/project/langchain/

[24] https://github.com/csebuetnlp/xl-sum/tree/master/multilingual_rouge_scoring

[25] https://pypi.org/project/numpy/

gradually building larger itemsets from smaller ones while pruning those that do not meet the minimum support threshold.

In our analysis, we reported the following measures: (i) **Support**: $s(X) = \frac{\sigma(X)}{N}$, where $\sigma(X)$ is the number of transactions in which X appears and N is the total number of transactions.

(ii) **Confidence**: $c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$, measures the probability of occurrence of itemset Y with itemset X.

Implementation Details To implement the Rule Association algorithm, we created a DataFrame for each task's results using pandas DataFrames[26]. Each DataFrame contains the results for all the configurations for every language. Subsequently, we binned each score column into three bins - high, medium, and low, based on the 30th and 60th percentiles. Later, we merged all the data frames based on the configuration name. Then we used the apriori algorithm from the efficient-apriori[27] library, which produces two outputs - itemsets and rules. Later, we filtered weak rules (support > 0.05 & confidence > 0.75).

A.4 Prompting

Question Answering Answer the following <Question> based only on the given <Context>. Follow these instructions:

- Include only words from the given context in your answer.
- Keep the answer as short as possible.
- Provide the answer in *expected output language*.

Named Entity Recognition You are an NLP assistant whose purpose is to perform Named Entity Recognition (NER). You need to assign each entity a tag from the following:

1. PER means a person.
2. ORG means an organization.
3. LOC means a location entity.

The output should be a list of tuples in the format:

$$[(\text{Tag}, \text{Entity}), (\text{Tag}, \text{Entity})]$$

for each entity in the sentence. The entities should be in the *expected output language*.

[26] https://pypi.org/project/pandas/

[27] https://pypi.org/project/efficient-apriori/

Summarization Write a summary of the given <Text> The output should be in *expected output language*. The output must be up to 2 sentences maximum.

Natural Language Inference You are an NLP assistant whose purpose is to solve Natural Language Inference (NLI) problems. NLI is the task of determining the inference relation between two texts: entailment, contradiction, or neutral. Your answer should be one word from the following: entailment, contradiction, or neutral.

A.4.1 The Holy Grail of Optimal Configuration

Few-Shot Examples Impact Figure 6 demonstrates that for all tasks, using a few-shot setting over a zero-shot setting yields better results. Interestingly, For all tasks, except for NLI, high-resource languages achieved better improvement when considering a few-shot setting over low-resource languages.

Output Selection Effects Figure 7 demonstrates that while in extractive QA the output should be in the source language, and in the summarization task, the output should be in English; in NER, the output is ambiguous.

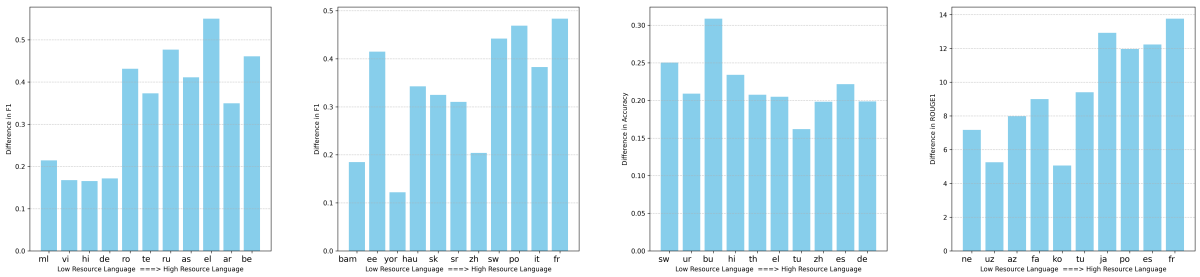
A.4.2 Factors Explaining Performance

A.4.3 Script Impact

Figure 8 presents the performance improvement achieved by the highest-performing prompt configuration among all configurations compared to the pre-translation prompt, for each language. Notably, the language family (as categorized by scripts) reveals a relatively even distribution of performance gains within the same language family. For example, languages using the Cyrillic script show greater improvement than those using the Latin script. Interestingly, languages in the same script family sometimes show varying results; for example, Spanish and Ewe belong have Latin script, but Ewe shows greater improvement over Spanish.

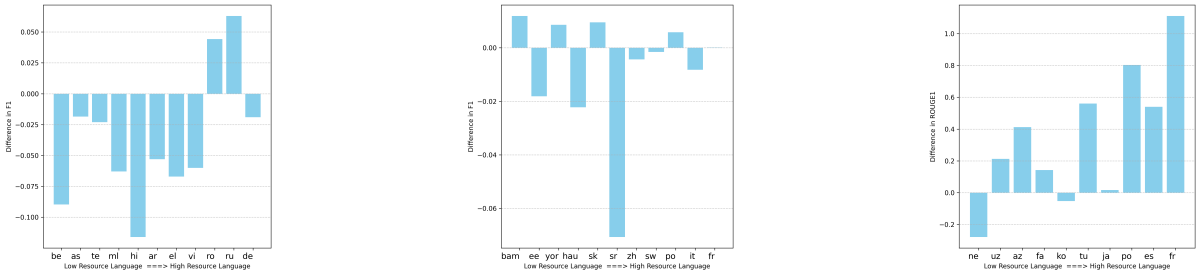
Linguistic Similarity To English We used the lang2vec[28] library to obtain syntactic similarity scores for each language. The Pearson correlation was calculated based on two vectors: one representing language similarities (ranging from 0 to 1) and the other representing model performance scores for each language across tasks. This correlation

[28] https://github.com/antonisa/lang2vec



(a) QA (b) NER (c) NLI (d) Summarization

Figure 6: Few-Shot and Zero-Shot Performance Gap (Few-Shot - Zero-Shot) for each task/language.



(a) QA (b) NER (c) Summarization

Figure 7: Output Performance Gap (English - Source) for each task/language

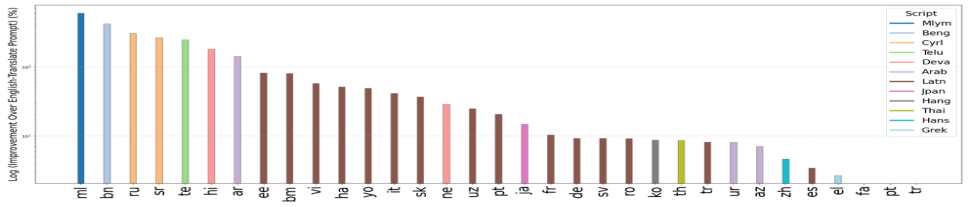


Figure 8: Percentage improvement over *pre-translation* approach, when using the highest configuration for each task For GPT-3.5-Turbo. The bars are color-coded based on the language family script.

Task	*Model Output*	*Expected Output*	*Explnation*	*Phenoenen*
NER	['LOC: 新北市', 'LOC: 平溪']	[(LOC, '新北市',), (LOC, '平溪',)]	List of strings, instead of list of tuples.	**Format Inconsistency**
	[PER: Hiei]\n- [PER: Hinata]\n	['PER', 'Hiei'), ('PER', 'Hinata')]	New line between each entity.	**Format Inconsistency**
	Ner Tags: ['PER: LL Cool J']	[(PER: LL Cool J)]	Redundant words in the beginning.	**Extraneous information**
	[] (No entities found in the sentence)	[]	Redundant words in the end.	**Extraneous information**
	Since the last sentence is in English, I will provide the NER tags in English as well	[(PER: Д а в и дIV Г р у з и с к и)]	Refusing to output in the desired language.	**Unwarranted Refusal**
QA	[The united states]	The united states	List of string instead of a string.	**Format Inconsistency**
	The question cannot be answered as the answer is not provided in the given context	[Luke Kuechly]	Insufficent information	**Unwarranted Refusal**
NLI	The second statement neutral because it does not provide any information that contradicts	neutral	Unnecessary justification for the choice.	**Extraneous information**
	vinculación	entailment	Spanish word for entailment instead of English.	**Wrong Language**
Summarization	Resumo: O ministro de Emergências da Rússia, Sergei Shoigu ...	O ministro de Emergências da Rússia, Sergei Shoigu ...	Redundant words ('Resumo' - Summary in Portuguese) in the beginning.	**Extraneous information**

Table 7: Error analysis of unexpected model outputs and observed in various tasks/languages.

Model	QA	NER	Summarization
GPT	56	60	96
Mixtral	60	61	78
Gemini	63	61	96

Table 8: Percentage of success of expected output languages for each model/task

Model	QA	NER	Summarization	NLI
GPT	0.14*	0.28**	0.42**	0.01
Mixtral	0.13*	0.2**	0.31**	0.08
Gemini	0.1*	0.19**	0.25**	0.01

Table 9: Pearson correlation between linguistic syntactic similarity to English and task performance for GPT, Mixtral, and Gemini. * $p < 0.05$, ** $p < 0.01$

was calculated at the instance level. Table 9 shows a positive correlation between model performance and syntactic similarity to English, especially for the summarization task, indicating that syntactic similarity to English significantly improves performance in this task. Additionally, NER also exhibits positive correlations, suggesting that models can better identify and classify entities in languages that share syntactic features with English.

B Error Analysis

B.1 Format Issues

Automatic evaluation requires consistent output formatting, especially in tasks like Named Entity Recognition (NER), which must adhere to a predefined format rather than free text. A common practice involves prompting the model to generate results in a specific format, such as a list of tuples representing entities and their types (e.g., $(Loc, NewYorkCity)$. However, achieving perfect consistency can be challenging. Models may not always adhere to the requested format, leading to difficulties in evaluation.

Qualitative Analysis We analyzed unexpected model outputs in various tasks and languages. For each task, we noted common phenomena observed and the expected model output. The results in Table 7 reveal that for the NER task, due to its rigid format, the model exhibited many error types. The models showed phenomena such as format inconsistency and extraneous introduction, which require a more generative normalization method to handle. An interesting phenomenon that made our modular selective pre-translating approach difficult to implement is unwarranted refusal, where the model refuses to output in the required language.

B.2 Incorrect Output Language

Table 8 summarizes the percentage of accurately outputted language for all tasks (except NLI, due to its index-based format) across all models. The results reveal that in extractive tasks such as extractive QA and NER, where the output overlaps with the context, the model struggles the most to output in the desired language. However, in abstractive summarization, a generation task, the model had better success.

C Translation: Key to Pre-Translation

C.1 Machine Translation Engines Comparison

To evaluate machine translation tools, we compared Google Translate API and Bing Translator. We excluded multilingual LLMs from consideration, as *zhu2023multilingual* found that these models still lag behind commercial systems like Google Translate, especially for low-resource languages. As shown in Table 11 and Figure 12, Google Translate outperformed Bing Translator across all metrics, demonstrating superior performance. Notably, Welsh and Maltese, both low-resource languages, achieved the highest scores. .

C.2 Linguistic Similarity To English

The results in Table 10 demonstrate the correlation between the syntactic similarity to English of the language and the ROUGE translation score of the language. The results show that the most significant

Group	Pearson Correlation	P-value
High Resource	0.73	0.05
Medium Resource	0.48	0.07
Low Resource	0.06	0.78
Extremly Low Resource	-0.34	0.30

Table 10: Correlation between syntactic similarity to English and the ROUGE score (by language subset).

correlation was observed in languages belonging to the high-resource category, and this correlation decreases as the class of the language becomes low-resource.

.

D Selective Pre-Translation Prompt Generator

We have launched a space on Hugging Face. The space makes it easy for the community to receive recommended configurations based on the type of task and language. In Figure 9 , we can see an overview of the application and an example of a recommended configuration. Figures 10 and 11 provide examples of generating prompts for zero-shot and few-shot settings.

E Detailed Results

The results across all tasks, languages and models are included in our benchmarking exercise are provided in Table 15 (for XQuAD), 13 (for indciQA), 16 (for WikiANN), 17 (for MasakhNER), 18 (for XL-Sum), 14 (for XNLI). The result of the correlation for Gemini are included in Table 19, for Mixtral in Table 20, and for bloomz in Table 21

	Google Translate API			Bing Translator		
Language	*ROUGE*	*Meteor*	*BLEU*	*ROUGE*	*Meteor*	*BLEU*
Welsh	0.86	0.86	0.63	0.85	0.85	0.61
Maltese	0.84	0.83	0.59	0.83	0.82	0.56
Danish	0.81	0.79	0.51	0.81	0.79	0.49
Swedish	0.81	0.80	0.51	0.80	0.79	0.51
Portuguese	0.81	0.79	0.52	0.80	0.78	0.50
Catalan	0.80	0.79	0.49	0.78	0.77	0.45
Spanish	0.79	0.64	0.30	0.69	0.64	0.30
Serbian	0.79	0.77	0.48	0.03	0.07	0.01
Bulgarian	0.79	0.77	0.45	0.75	0.72	0.37
French	0.79	0.77	0.48	0.78	0.76	0.48
Nepali (macrolanguage)	0.79	0.78	0.46	0.74	0.72	0.38
Macedonian	0.78	0.77	0.46	0.72	0.70	0.35
Swahili (macrolanguage)	0.78	0.79	0.51	0.74	0.74	0.43
Hebrew	0.78	0.77	0.47	0.76	0.75	0.44
German	0.78	0.76	0.46	0.79	0.76	0.46
Indonesian	0.78	0.77	0.46	0.78	0.77	0.44
Romanian	0.78	0.76	0.45	0.77	0.75	0.43
Panjabi	0.78	0.77	0.46	0.74	0.72	0.4
Bosnian	0.78	0.76	0.45	0.75	0.73	0.39
Hindi	0.78	0.76	0.45	0.76	0.74	0.41
Turkish	0.77	0.75	0.43	0.76	0.73	0.41
Armenian	0.77	0.75	0.43	0.68	0.65	0.28
Irish	0.77	0.76	0.47	0.76	0.74	0.42
Gujarati	0.77	0.76	0.44	0.73	0.69	0.35
Telugu	0.77	0.76	0.44	0.73	0.71	0.38
Slovak	0.76	0.74	0.42	0.75	0.72	0.40
Italian	0.76	0.68	0.34	0.72	0.68	0.34
Galician	0.76	0.74	0.43	0.74	0.71	0.39
Estonian	0.76	0.74	0.41	0.74	0.71	0.37
Czech	0.76	0.74	0.42	0.76	0.73	0.4
Marathi	0.76	0.74	0.41	0.72	0.69	0.35
Uzbek	0.75	0.74	0.39	0.67	0.63	0.28
Urdu	0.75	0.72	0.39	0.71	0.68	0.33
Ukrainian	0.75	0.73	0.41	0.74	0.72	0.4
Malayalam	0.75	0.73	0.4	0.71	0.68	0.34
Sinhala	0.75	0.73	0.39	0.69	0.66	0.32
Bengali	0.74	0.73	0.39	0.74	0.70	0.36
Croatian	0.74	0.72	0.39	0.73	0.70	0.36
Lao	0.74	0.73	0.39	0.69	0.66	0.30
Haitian	0.74	0.73	0.41	0.67	0.65	0.30
Hungarian	0.74	0.72	0.38	0.74	0.71	0.37
Kazakh	0.73	0.72	0.38	0.67	0.62	0.27
Russian	0.73	0.70	0.38	0.72	0.69	0.36
Vietnamese	0.73	0.72	0.39	0.72	0.71	0.36
Slovenian	0.73	0.71	0.38	0.69	0.67	0.32
Zulu	0.73	0.74	0.43	0.65	0.65	0.32
Tamil	0.73	0.71	0.37	0.70	0.68	0.33
Finnish	0.73	0.70	0.36	0.72	0.68	0.33
Kannada	0.72	0.71	0.37	0.71	0.68	0.33
Lithuanian	0.72	0.70	0.36	0.68	0.63	0.28
Icelandic	0.72	0.70	0.37	0.72	0.7	0.36
Southern Sotho	0.72	0.71	0.40	0.62	0.6	0.27
Korean	0.71	0.68	0.32	0.69	0.66	0.31
Basque	0.71	0.68	0.34	0.67	0.63	0.27
Thai	0.71	0.66	0.3	0.69	0.65	0.28
Chinese	0.70	0.67	0.31	0.68	0.64	0.29
Georgian	0.70	0.66	0.31	0.64	0.58	0.21
Xhosa	0.70	0.70	0.38	0.63	0.63	0.28
Dutch	0.70	0.66	0.31	0.72	0.67	0.34
Japanese	0.69	0.66	0.30	0.69	0.65	0.29
Polish	0.69	0.65	0.30	0.68	0.64	0.29
Burmese	0.69	0.65	0.30	0.63	0.58	0.22
Khmer	0.68	0.65	0.30	0.64	0.59	0.23
Kinyarwanda	0.68	0.67	0.34	0.61	0.6	0.23
Samoan	0.67	0.65	0.33	0.62	0.59	0.26
Somali	0.66	0.66	0.32	0.59	0.57	0.22
Faroese	0.62	0.60	0.28	0.65	0.62	0.28
Lingala	0.60	0.59	0.24	0.60	0.58	0.23
Azerbaijani	0.31	0.29	0.05	0.11	0.10	0.00
Fijian	0.16	0.16	0.02	0.51	0.47	0.12

Table 11: Comparsion between Google Translate API and Bing Translator.

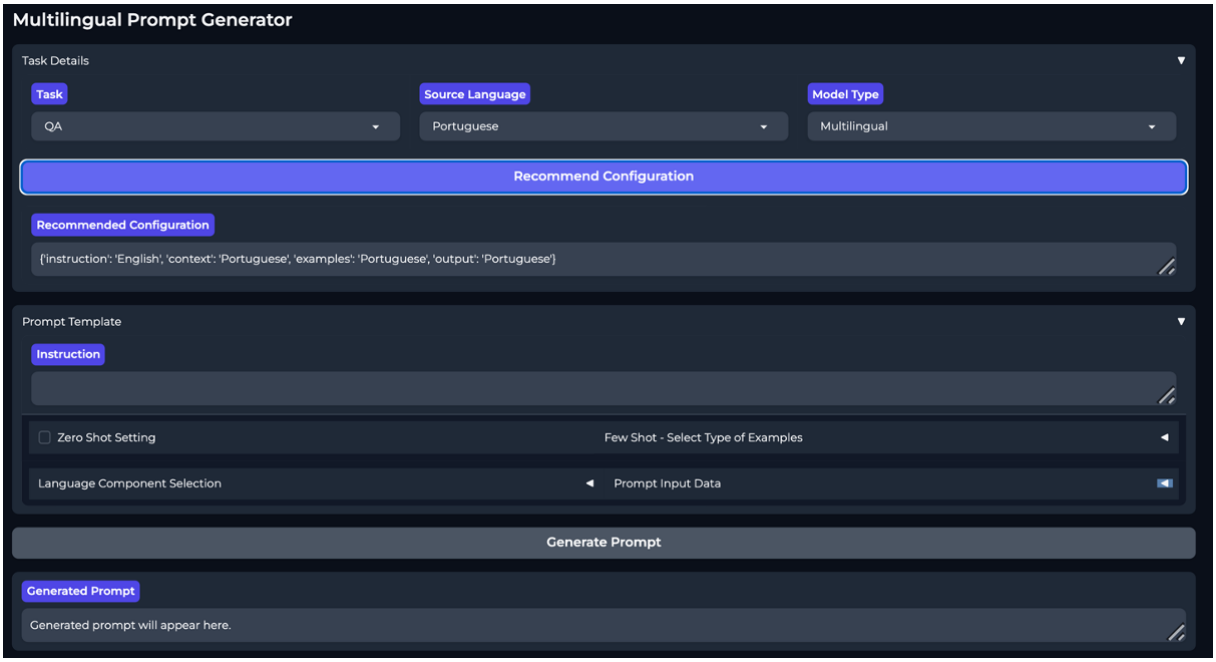


Figure 9: *Recomended Configuration:* Overview of our application. From top to bottom: Task Details—general configuration about the task (Task, Language, Model). Clicking on "Recommended Configuration" provides a suggested selective pre-translation configuration.

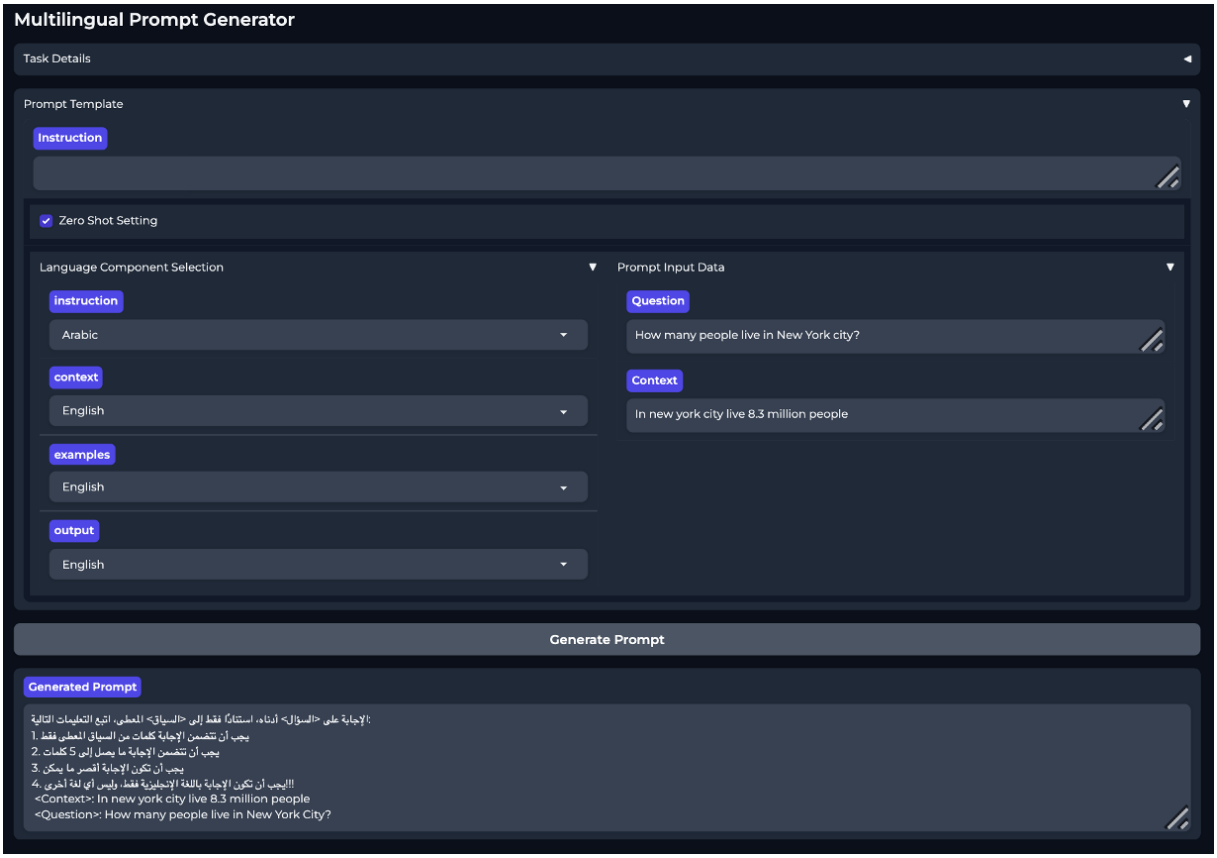


Figure 10: *Generating selective pre-translation prompt for zero-shot:* The user needs to configure the instruction (optional) and the languages for the components under "Language Component Selection": instruction, context, examples, and output. Additionally, under "Prompt Input Data," the user must configure the relevant input data or task, such as the question and context for QA in this example. Clicking on "Generate Prompt" provides a zero-shot pre-translation prompt

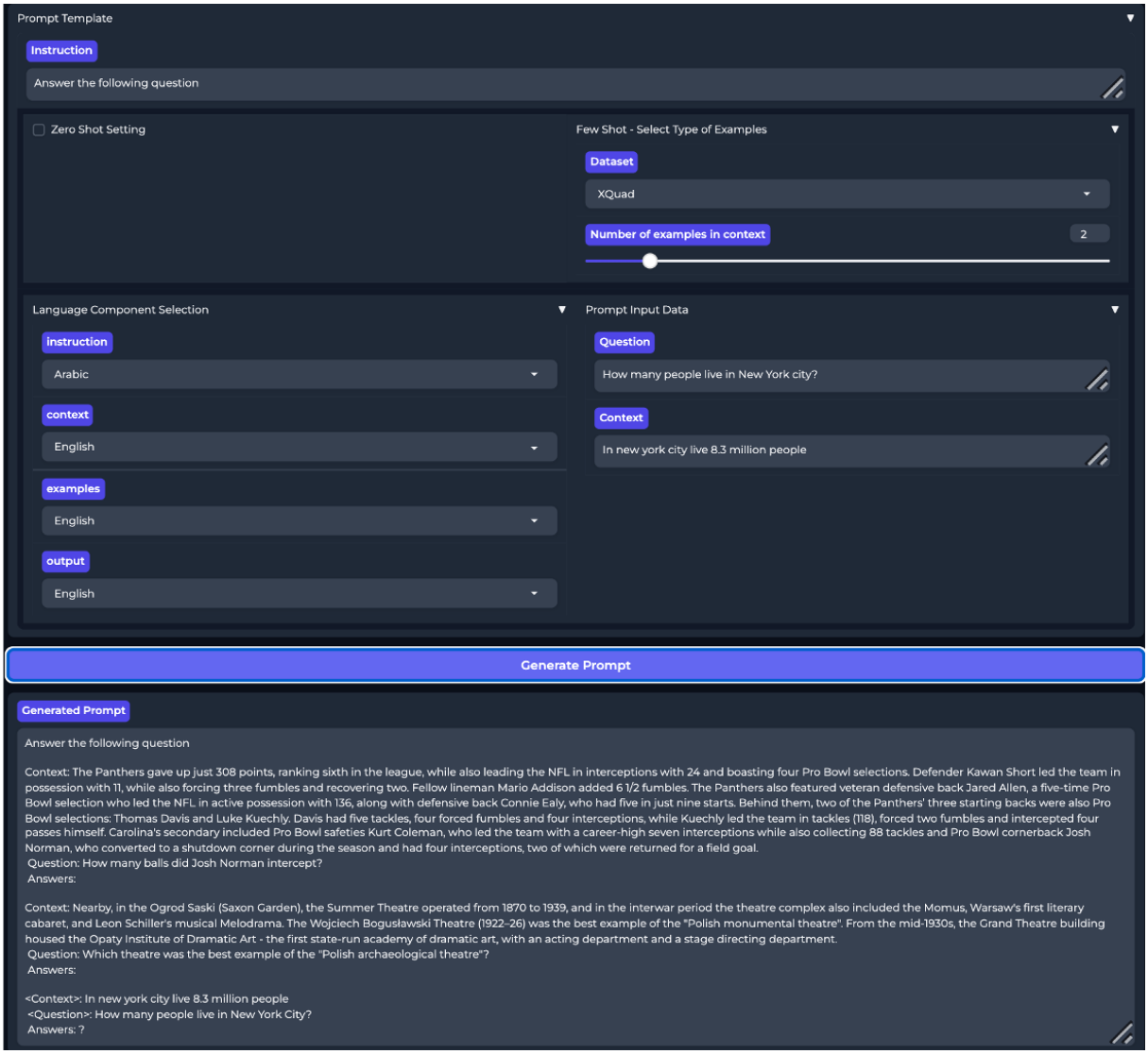
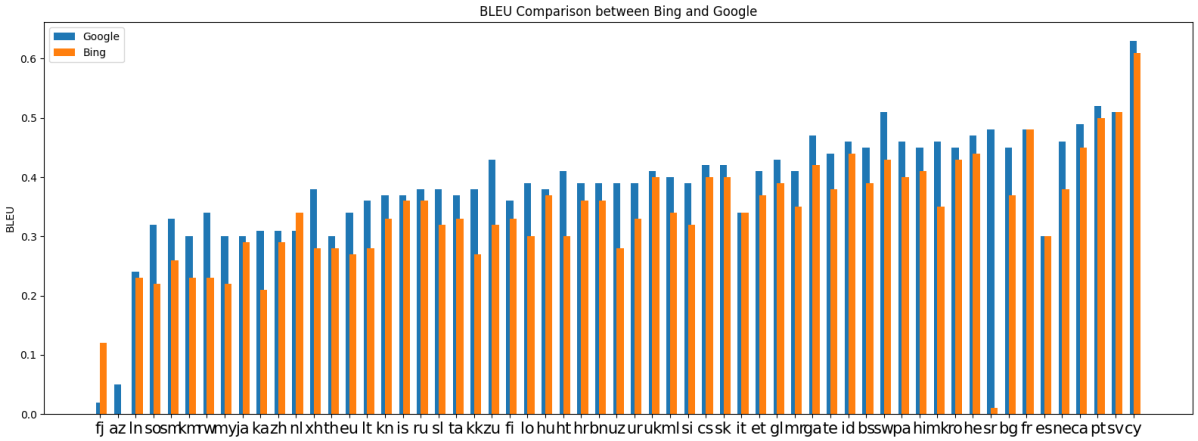
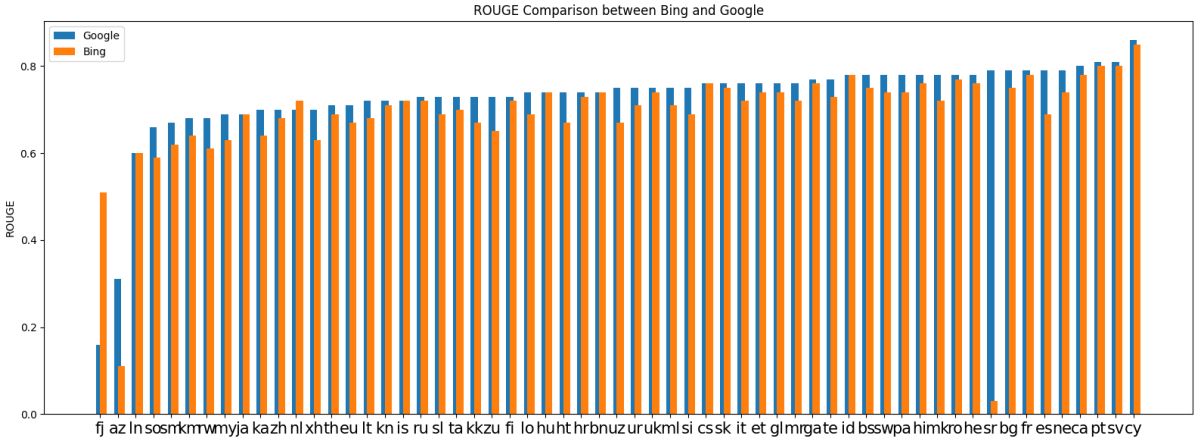


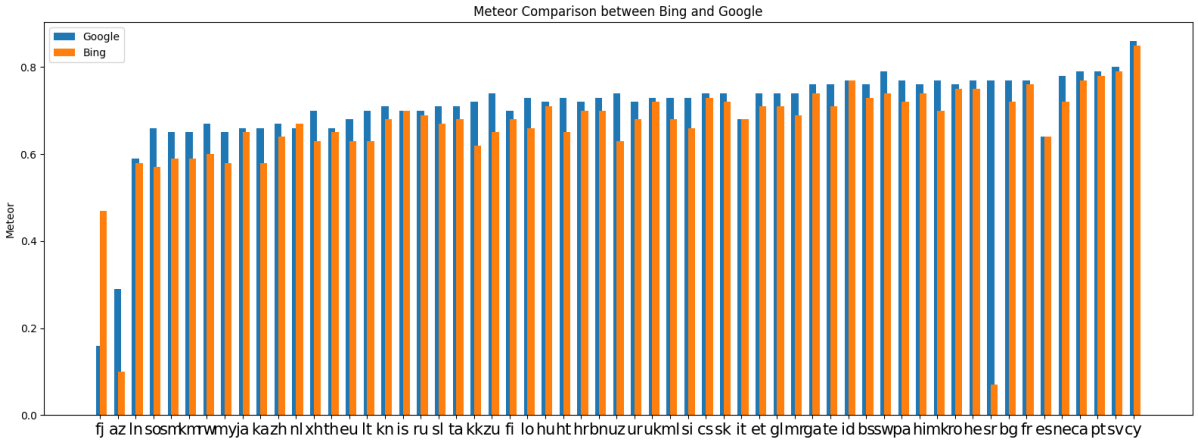
Figure 11: *Generating selective pre-translation prompt for few-shot:* Here, the user must also configure the few-shot settings: the dataset to use (from which the few-shot examples are taken) and the number of examples to use (default = 1).



(a) BLEU



(b) ROUGE



(c) Meteor

Figure 12: Google Translate API vs Bing Translator Comparsion

Language	Lang Code	Number of Tokens (M)	Percentage of Tokens	Class
English	en	181,015	92.64%	A
French	fr	3,553	1.81853%	A
German	de	2,871	1.46937%	A
Spanish	es	1,510	0.77289%	A
Italian	it	1,188	0.60793%	A
Portuguese	po	1,025	0.52483%	A
Russian	ru	368	0.18843%	A
Romanian	ro	308	0.15773%	A
Swedish	sv	221	0.11307%	A
Japanese	ja	217	0.11109%	A
Chinese	zh	194	0.09905%	B
Indonesian	id	117	0.05985%	B
Turkish	tr	116	0.05944%	B
Vietnamese	vi	83	0.04252%	B
Greek	el	62	0.03153%	B
Arabic	ar	61	0.03114%	B
Serbian	sr	53	0.02706%	B
Korean	ko	33	0.01697%	B
Slovak	sk	28	0.01431%	B
Thai	th	27	0.01372%	B
Slovenian	sl	26	0.01333%	B
Persian	fa	17	0.00856%	C
Hebrew	he	15	0.00769%	C
Hindi	hi	9	0.00483%	C
Bulgarian	bg	6	0.00303%	C
Bengali	bn	3	0.00154%	C
Malayalam	ml	3	0.00165%	C
Azerbaijani	az	2	0.00128%	C
Telugu	te	2	0.00084%	C
Uzbek	uz	1.5	0.00075%	C
Nepali	ne	1.1	0.00057%	C
Urdu	ur	0.7	0.00035%	C
Swahili	sw	0.6	0.00030%	C
Assamese	as	0	0.00000%	D
Bambara	bam	0	0.00000%	D
Ewe	ee	0	0.00000%	D
Hausa	hau	0	0.00000%	D
Yoruba	yor	0	0.00000%	D

Table 12: List of languages, language codes, number of tokens in pre-trained GPT-3 data, data ratios. The languages are grouped into four classes based on their data ratios in the GPT-3 pre-trained data: High Resource (H > 0.1%), Medium Resource (M > 0.01%), and Low Resource (L < 0.01%), and extremely low resource for unrepresented languages.

Configuration				Assamese			Bengali			Hindi			Malayalam			Telugu		
P	I	C	O	*gemini*	*gpt*	*mixtral*	*gemini*	*gpt*	*mixtral*	*gemini*	*gpt*	*mixtral*	*gemini*	*gpt*	*mixtral*	*gemini*	*gpt*	*mixtral*
S	S	Z	E	0.2	0.43	0.00	0.16	0.48	0.02	0.26	0.22	0.19	0.19	0.25	0.03	0.29	0.15	0.13
E	S	Z	E	0.15	0.03	0.00	0.11	0.03	0.02	0.21	0.30	0.12	0.13	0.13	0.14	0.12	0.10	0.10
S	S	S	E	0.59	0.35	0.10	0.65	0.52	0.28	0.74	0.68	0.37	0.46	0.36	0.04	0.53	0.23	0.19
E	E	E	S	0.25	0.00	0.01	0.25	0.10	0.02	0.53	0.26	0.19	0.02	0.06	0.07	0.27	0.14	0.09
S	E	S	E	0.69	0.51	0.70	0.71	0.67	0.44	0.80	0.79	0.62	0.47	0.41	0.32	0.46	0.53	0.39
E	S	S	S	0.2	0.00	0.01	0.08	0.02	0.18	0.50	0.5	0.34	0.12	0.06	0.11	0.16	0.12	0.14
E	S	Z	S	0.39	0.10	0.01	0.35	0.03	0.08	0.55	0.23	0.17	0.12	0.10	0.03	0.19	0.1	0.08
S	S	S	S	0.69	0.32	0.10	0.65	0.57	0.51	0.74	0.74	0.45	0.52	0.33	0.06	0.60	0.34	0.55
E	S	E	S	0.32	0.00	0.00	0.27	0.03	0.02	0.55	0.30	0.22	0.02	0.17	0.08	0.15	0.04	0.11
S	S	Z	S	0.69	0.23	0.07	0.64	0.47	0.39	0.72	0.62	0.42	0.60	0.24	0.09	0.53	0.16	0.15
E	E	S	S	0.28	0.09	0.12	0.24	0.18	0.17	0.48	0.35	0.29	0.04	0.15	0.08	0.18	0.14	0.13
E	S	E	E	0.11	0.13	0.00	0.08	0.03	0.02	0.29	0.32	0.18	0.05	0.08	0.04	0.09	0.09	0.07
E	E	Z	E	0.17	0.01	0.00	0.14	0.03	0.02	0.23	0.30	0.24	0.15	0.10	0.11	0.15	0.17	0.12
S	E	Z	E	0.54	0.37	0.07	0.64	0.43	0.15	0.44	0.64	0.68	0.48	0.2	0.42	0.59	0.4	0.16
S	E	Z	S	0.71	0.46	0.29	0.60	0.68	0.40	0.60	0.71	0.72	0.13	0.48	0.39	0.60	0.45	0.27
S	S	E	E	0.14	0.33	0.01	0.22	0.51	0.05	0.45	0.71	0.15	0.09	0.32	0.01	0.24	0.29	0.22
E	E	Z	S	0.24	0.05	0.01	0.33	0.17	0.07	0.40	0.31	0.19	0.02	0.07	0.06	0.26	0.16	0.1
E	S	S	E	0.1	0.00	0.00	0.10	0.01	0.02	0.33	0.25	0.26	0.06	0.10	0.13	0.09	0.10	0.02
S	S	E	S	0.72	0.44	0.02	0.66	0.51	0.27	0.78	0.76	0.28	0.63	0.42	0.01	0.53	0.28	0.06
S	E	S	S	0.64	0.56	0.64	0.60	0.66	0.53	0.80	0.82	0.60	0.20	0.49	0.31	0.65	0.46	0.34
E	E	E	E	0.15	0.00	0.00	0.11	0.04	0.01	0.24	0.29	0.20	0.05	0.05	0.12	0.11	0.08	0.11
S	E	E	E	0.19	0.33	0.05	0.28	0.22	0.11	0.47	0.32	0.25	0.07	0.07	0.14	0.25	0.21	0.17
E	E	S	E	0.09	0.01	0.01	0.11	0.04	0.04	0.31	0.24	0.31	0.07	0.08	0.08	0.10	0.12	0.12
S	E	E	S	0.68	0.44	0.24	0.57	0.57	0.37	0.74	0.76	0.43	0.02	0.3	0.29	0.55	0.34	0.31

Table 13: Comparing performance of different models on all languages in IndicQA. Metric: F1 Score.

Configuration			Arabic			Chinese			Greek			Hindi			Spanish			Swahili			Thai			Turkish			Urdu			Bulgarian		
P	*I*	*C*	gemini	gpt	mixtral	gemini	gpt	mixtral	gemini	gpt	mixtral	gemini	gpt	mixtral	gemini	gpt	mixtral	gemini	gpt	mixtral	gemini	gpt	mixtral	gemini	gpt	mixtral	gemini	gpt	mixtral	gemini	gpt	mixtral
E	S	E	0.72	0.68	0.51	0.63	0.58	0.55	0.75	0.53	0.61	0.59	0.49	0.59	0.56	0.32	0.65	0.54	0.54	0.52	0.71	0.33	0.45	0.45	0.0	0.66	0.52	0.41	0.49	0.54	0.46
S	S	E	0.65	0.67	0.54	0.62	0.63	0.52	0.66	0.59	0.54	1.00	0.46	0.6	0.64	0.52	0.62	0.64	0.59	0.52	0.56	1.0	0.47	0.42	0.5	N.A	0.52	N.A	0.52	0.54	0.44
S	E	Z	0.62	0.55	0.54	0.71	0.52	0.56	0.57	0.61	0.64	0.56	0.58	0.56	0.53	0.39	0.80	0.62	0.63	0.57	0.52	0.36	0.62	0.36	0.38	0.64	0.53	0.48	0.5	0.47	0.39
S	E	E	0.61	0.62	0.52	0.59	0.6	0.45	0.75	0.54	0.78	0.59	0.59	0.6	0.58	0.43	0.61	0.59	0.6	0.58	0.67	0.79	0.54	0.48	0.41	0.83	0.45	0.39	0.52	0.48	0.37
E	E	E	0.65	0.66	0.5	0.61	0.6	0.57	0.62	0.58	0.57	0.59	0.54	0.63	0.59	0.42	0.67	0.58	0.62	0.57	0.57	0.38	0.57	0.52	0.34	0.59	0.52	0.45	0.55	0.52	0.4
E	S	S	0.59	0.61	0.59	0.59	0.6	N.A	0.7	0.58	0.58	0.67	0.43	0.59	0.57	0.33	0.68	0.69	0.69	0.52	0.58	0.0	0.53	0.57	0.5	1.00	0.51	0.0	0.51	0.54	0.47
E	E	Z	0.75	0.55	N.A	0.59	0.57	0.55	0.68	0.61	0.6	0.68	0.55	0.58	0.51	0.38	0.8	0.53	0.55	0.57	0.54	0.45	0.63	0.41	0.43	0.5	0.55	0.45	0.69	0.46	0.37
S	S	Z	1.0	0.45	0.51	0.57	0.55	0.44	N.A	0.53	0.57	1.00	0.57	0.53	0.4	0.83	0.68	0.57	0.46	0.49	0.46	0.33	0.48	0.38	0.33	0.54	0.57	0.42	0.71	0.44	0.37
S	S	S	0.63	0.72	0.55	0.58	0.6	N.A	0.78	0.6	0.64	0.65	0.52	0.62	0.53	0.29	0.77	0.51	0.55	0.61	0.54	0.0	0.51	0.47	0.36	0.52	0.5	N.A	0.52	0.57	0.11
S	E	S	0.67	0.66	0.54	0.57	0.63	N.A	0.67	0.6	0.61	0.63	0.52	0.6	0.61	0.44	0.76	0.57	0.56	0.59	0.62	0.34	0.58	0.57	0.44	0.43	0.56	0.48	0.51	0.54	0.44
E	E	S	0.65	0.64	0.56	0.59	0.61	N.A	0.61	0.64	0.67	0.65	0.71	0.62	0.64	0.47	0.43	0.6	0.62	0.62	0.73	0.35	0.54	0.53	0.42	0.50	0.55	0.73	0.54	0.56	0.39
E	S	Z	0.82	0.0	0.43	0.59	0.58	0.46	0.50	0.53	0.40	0.82	0.63	0.56	0.48	0.53	0.64	0.56	0.46	0.47	0.47	0.4	N.A	0.34	0.48	0.60	0.53	0.52	0.57	0.42	0.36

Table 14: Comparing performance of different models on all languages in XNLI. Metric: Acc Score.

Configuration				Arabic			German			Greek			Romanian			Russian			Vietnamese		
P	*I*	*C*	*O*	*gemini*	*gpt*	*mixtral*	*gemini*	*gpt*	*mixtral*	*gemini*	*gpt*	*mixtral*	*gemini*	*gpt*	*mixtral*	*gemini*	*gpt*	*mixtral*	*gemini*	*gpt*	*mixtral*
S	S	Z	E	0.27	0.46	0.06	0.56	0.72	0.62	0.34	0.17	0.13	0.55	0.58	0.41	0.21	0.18	0.19	0.59	0.6	0.23
E	S	Z	E	0.28	0.12	0.05	0.56	0.48	0.49	0.34	0.19	0.16	0.56	0.62	0.31	0.21	0.20	0.20	0.55	0.13	0.26
S	S	S	E	0.78	0.53	0.33	0.76	0.67	0.58	0.50	0.54	0.53	0.45	0.64	0.48	0.57	0.69	0.26	0.67	0.55	0.44
E	E	E	S	0.36	0.28	0.10	0.62	0.85	0.35	0.43	0.47	0.20	0.61	0.54	0.41	0.39	0.26	0.14	0.52	0.45	0.30
S	E	S	E	0.84	0.72	0.30	0.73	0.65	0.46	0.74	0.65	0.37	0.54	0.67	0.51	0.74	0.61	0.33	0.72	0.58	0.47
E	S	S	S	0.62	0.19	0.37	0.71	0.50	0.43	0.61	0.38	0.30	0.68	0.51	0.37	0.47	0.34	0.14	0.62	0.48	0.44
E	S	Z	S	0.48	0.29	0.24	0.68	0.39	0.43	0.40	0.17	0.14	0.67	0.63	0.26	0.35	0.16	0.12	0.57	0.58	0.28
S	S	S	S	0.80	0.54	0.40	0.76	0.64	0.36	0.71	0.54	0.52	0.51	0.61	0.51	0.75	0.61	0.47	0.72	0.67	0.54
E	S	E	S	0.51	0.24	0.06	0.68	0.51	0.54	0.44	0.38	0.22	0.65	0.50	0.37	0.34	0.19	0.16	0.61	0.46	0.38
S	S	Z	S	0.74	0.40	0.28	0.70	0.72	0.65	0.67	0.69	0.32	0.73	0.69	0.45	0.67	0.65	0.42	0.75	0.73	0.36
E	E	S	S	0.52	0.28	0.16	0.68	0.50	0.38	0.49	0.39	0.26	0.68	0.50	0.41	0.47	0.29	0.24	0.62	0.43	0.32
E	S	E	E	0.26	0.25	0.07	0.58	0.50	0.49	0.42	0.27	0.16	0.40	0.50	0.43	0.23	0.26	0.17	0.58	0.49	0.36
E	E	Z	E	0.27	0.16	0.07	0.54	0.50	0.37	0.34	0.20	0.26	0.52	N.A	0.31	0.23	0.22	0.10	0.49	0.33	0.30
S	E	Z	E	0.55	0.48	0.15	0.61	0.62	0.61	0.56	0.38	0.50	0.62	0.66	0.50	0.48	0.46	0.23	0.67	0.66	0.35
S	E	Z	S	0.62	0.49	0.13	0.61	0.62	0.61	0.61	0.26	0.49	0.65	N.A	0.41	0.58	0.58	0.33	0.60	0.58	0.40
S	S	E	E	0.27	0.39	0.05	0.65	0.57	0.46	0.43	0.40	0.15	0.58	0.55	0.37	0.26	0.35	0.15	0.61	0.59	0.34
E	E	Z	S	0.38	0.13	0.11	0.57	0.56	0.48	0.40	0.31	0.26	0.54	N.A	0.29	0.40	0.20	0.16	0.54	0.31	0.29
E	S	S	E	0.31	0.23	0.13	0.67	0.48	0.51	0.43	0.27	0.36	0.66	0.51	0.29	0.29	0.25	0.18	0.57	0.47	0.41
S	S	E	S	0.74	0.58	0.26	0.68	0.65	0.51	0.70	0.57	0.46	0.51	0.64	0.5	0.72	0.66	0.4	0.74	0.65	0.43
S	E	S	S	0.72	0.74	0.39	0.70	0.64	0.43	0.57	0.55	0.49	0.45	0.67	0.55	0.63	0.62	0.43	0.66	0.58	0.52
E	E	E	E	0.23	0.17	0.06	0.60	0.78	0.47	0.39	0.67	0.19	0.58	0.63	0.43	0.22	0.17	0.18	0.56	0.46	0.40
S	E	E	E	0.77	0.58	0.09	0.71	0.57	0.37	0.57	0.53	0.23	0.46	0.59	0.46	0.39	0.42	0.13	0.63	0.51	0.42
E	E	S	E	0.32	0.22	0.16	0.65	0.51	0.42	0.40	0.35	0.2	0.64	0.49	0.41	0.30	0.18	0.34	0.59	0.47	0.4
S	E	E	S	0.65	0.71	0.30	0.66	0.66	0.35	0.62	0.66	0.43	0.46	0.66	0.49	0.65	0.59	0.36	0.68	0.58	0.28

Table 15: Comparing performance of different models on all languages in XQuAD. Metric: F1 Score.

Configuration				**Chinese**			French			**Italian**			**Portuguese**			**Serbian**			**Slovak**			Swedish		
P	*I*	*C*	*O*	*gemini*	*gpt*	*mixtral*	*gemini*	*gpt*	*mixtral*	*gemini*	*gpt*	*mixtral*	*gemini*	*gpt*	*mixtral*	*gemini*	*gpt*	*mixtral*	*gemini*	*gpt*	*mixtral*	*gemini*	*gpt*	*mixtral*
E	E	E	E	0.00	0.01	0.23	0.53	0.58	0.15	0.38	0.53	0.23	0.56	0.60	0.41	0.20	0.21	0.00	0.48	0.53	0.19	0.59	0.62	0.41
E	E	E	S	0.00	0.00	0.18	0.37	0.63	0.17	0.31	0.52	0.23	0.55	0.60	0.42	0.22	0.26	0.01	0.48	0.50	0.21	0.59	0.58	0.38
E	E	S	E	0.00	0.01	0.21	0.54	0.6	0.25	0.66	0.66	0.52	0.56	0.58	0.29	0.17	0.18	0.01	0.45	0.51	0.31	0.59	0.59	0.42
E	E	S	S	0.00	0.06	0.22	0.55	0.61	0.00	0.59	0.68	0.48	0.55	0.61	0.38	0.13	0.13	0.02	0.46	0.48	0.24	0.57	0.56	0.21
E	E	Z	E	0.00	0.00	0.22	0.45	0.54	N.A	0.51	0.63	0.49	0.45	0.57	0.34	0.13	0.12	0.02	0.47	0.47	0.28	0.51	0.56	0.32
E	E	S	S	0.00	0.00	0.20	0.44	0.53	0.0	0.44	0.63	0.50	0.43	0.54	0.35	0.09	0.11	0.03	0.38	0.49	0.29	0.44	0.59	0.36
E	S	E	E	0.00	0.00	0.23	0.55	0.60	0.25	0.34	0.48	0.26	0.55	0.60	0.37	0.21	0.22	0.01	0.39	0.50	0.18	0.60	0.60	0.27
E	S	E	S	0.00	0.00	0.28	0.53	0.59	0.21	0.30	0.48	0.26	0.55	0.61	0.40	0.17	0.21	0.01	0.47	0.52	0.30	0.58	0.64	0.26
E	S	S	E	0.01	0.02	0.22	0.42	0.59	0.26	0.65	0.63	0.52	0.58	0.61	0.40	0.14	0.17	0.01	0.46	0.49	0.25	0.57	0.61	0.22
E	S	S	S	0.01	0.03	0.24	0.45	0.59	0.28	0.61	0.67	0.49	0.56	0.60	0.37	0.16	0.16	0.01	0.46	0.36	N.A	0.55	0.55	0.22
E	S	Z	E	0.00	0.00	0.24	0.48	0.55	0.36	0.53	0.64	0.50	0.47	0.55	0.35	0.12	0.12	0.02	0.46	0.48	0.29	0.52	0.56	0.33
E	S	Z	S	0.01	0.00	0.20	0.45	0.55	0.38	0.52	0.62	0.48	0.46	0.53	0.33	0.11	0.12	0.03	0.41	0.50	0.29	0.50	0.57	0.36
S	E	E	E	0.06	0.02	0.09	0.61	0.69	0.32	0.40	0.53	0.33	0.60	0.68	0.49	0.57	0.20	0.05	0.66	0.62	0.26	0.63	0.64	0.14
S	E	E	S	0.11	0.05	0.07	0.68	0.71	0.31	0.36	0.53	0.29	0.64	0.66	0.54	0.68	0.64	0.3	0.61	0.61	0.31	0.60	0.64	0.24
S	E	S	E	0.61	0.61	0.00	0.64	0.72	0.32	0.69	0.74	0.64	0.71	0.69	0.53	0.72	0.75	0.47	0.73	0.71	0.39	0.69	0.67	0.34
S	E	S	S	0.59	0.63	0.00	0.63	0.66	0.35	0.69	0.75	0.62	0.76	0.70	0.48	0.77	0.68	0.46	0.69	0.72	0.54	0.66	0.68	0.35
S	E	Z	E	0.07	0.07	0.02	0.48	0.59	0.48	0.54	0.66	0.56	0.50	0.63	0.48	0.22	0.42	0.25	0.57	0.65	0.49	0.50	0.57	0.37
S	E	Z	S	0.16	0.05	0.00	0.55	0.60	0.48	0.48	0.67	0.57	0.48	0.62	0.48	0.47	0.5	0.32	0.51	0.62	0.45	0.50	0.62	0.39
S	S	E	E	0.17	0.02	0.01	0.64	0.68	0.29	0.38	0.54	0.32	0.66	0.66	0.48	0.53	0.17	0.05	0.67	0.29	0.27	0.61	0.63	0.18
S	S	E	S	0.14	0.02	0.07	0.59	0.68	0.31	0.35	0.59	0.32	0.69	0.65	0.47	0.63	0.61	0.29	0.65	0.38	0.52	0.62	0.63	0.25
S	S	S	E	0.60	0.61	0.00	0.63	0.70	0.45	0.70	0.74	0.59	0.70	0.72	0.51	0.72	0.77	0.45	0.72	0.58	0.40	0.67	0.67	0.53
S	S	S	S	0.58	0.62	0.01	0.64	0.69	0.31	0.69	0.71	0.61	0.68	0.72	0.48	0.77	0.72	0.44	0.69	0.56	0.40	0.65	0.66	0.53
S	S	Z	E	0.12	0.08	0.02	0.57	0.57	0.45	0.57	0.67	0.56	0.51	0.63	0.48	0.28	0.46	0.29	0.55	0.65	0.47	0.54	0.58	0.37
S	S	Z	S	0.13	0.04	0.00	0.49	0.58	0.48	0.55	0.68	0.58	0.48	0.61	0.48	0.42	0.50	0.33	0.52	0.63	0.45	0.53	0.60	0.39

Table 16: Comparing performance of different models on all languages in WikiANN. Metric: F1 Score.

Configuration				**Bambara**			**Ewe**			**Hausa**			**Yoruba**		
P	I	C	O	*gemini*	*gpt*	*mixtral*	*gemini*	*gpt*	*mixtral*	*gemini*	*gpt*	*mixtral*	*gemini*	*gpt*	*mixtral*
E	E	E	E	0.16	0.18	0.10	0.42	0.47	0.22	0.45	0.24	0.26	0.06	0.09	0.03
E	E	E	S	0.17	0.17	0.07	0.43	0.48	0.23	0.46	0.61	0.31	0.07	0.09	0.05
E	E	S	E	0.15	0.17	0.12	0.44	0.52	0.22	0.45	0.52	0.20	0.07	0.08	0.04
E	E	S	S	0.14	0.15	0.07	0.39	0.5	0.17	0.43	0.39	0.10	0.06	0.08	0.03
E	E	Z	E	0.13	0.09	N.A	0.41	0.44	0.23	0.52	0.59	0.26	0.05	0.08	0.04
E	E	S	S	0.13	0.16	0.06	0.41	0.47	0.21	0.45	0.60	0.29	0.04	0.08	0.03
E	S	E	E	0.16	0.18	0.10	0.44	0.47	0.22	0.48	0.59	0.32	0.07	0.08	0.03
E	S	E	S	0.16	0.18	0.08	0.40	0.47	0.15	0.48	0.58	0.34	0.06	0.08	0.05
E	S	S	E	0.15	0.15	0.06	0.45	0.50	0.28	0.50	0.54	0.22	0.08	0.10	0.05
E	S	S	S	0.11	0.15	0.06	0.39	0.49	0.15	0.45	0.42	0.11	0.06	0.07	0.06
E	S	Z	E	0.15	0.28	0.19	0.40	0.46	0.23	0.54	0.62	0.27	0.08	0.08	0.05
E	S	Z	S	0.17	0.13	0.05	0.38	0.44	0.25	0.47	0.28	0.29	0.06	0.09	0.06
S	E	E	E	0.09	0.24	0.00	0.31	0.47	0.01	0.46	0.43	0.00	0.08	0.20	0.05
S	E	E	S	0.27	0.27	0.06	0.40	0.52	0.04	0.61	0.58	0.10	0.09	0.23	0.10
S	E	S	E	0.28	0.32	0.14	0.56	0.68	0.47	0.56	0.70	0.30	0.26	0.32	0.15
S	E	S	S	0.26	0.32	0.15	0.54	0.66	0.51	0.55	0.70	0.32	0.23	0.29	0.20
S	E	Z	E	0.07	0.09	0.06	0.31	0.39	0.00	0.48	0.70	0.00	0.20	0.20	0.01
S	E	Z	S	0.21	0.20	0.17	0.43	0.42	0.42	0.57	0.69	0.07	0.17	0.26	0.07
S	S	E	E	0.10	0.25	0.00	0.27	0.39	0.00	0.45	0.43	0.00	0.08	0.21	0.04
S	S	E	S	0.23	0.27	0.05	0.40	0.51	0.08	0.54	0.60	0.09	0.13	0.22	0.05
S	S	S	E	0.29	0.33	0.21	0.61	0.67	0.43	0.58	0.67	0.34	0.27	0.30	0.19
S	S	S	S	0.27	0.32	0.21	0.61	0.63	0.45	0.59	0.69	0.41	0.26	0.31	N.A
S	S	Z	E	0.08	0.18	0.00	0.31	0.37	0.00	0.46	0.46	0.01	0.03	0.22	0.03
S	S	Z	S	0.23	0.28	0.04	0.39	0.46	0.06	0.06	0.64	0.05	0.07	0.26	0.07

Table 17: Comparing performance of different models on all languages in MasakhaNER. Metric: F1 Score.

Configuration				Azerbaijani			French			Japanese			Korean			Nepali			Persian			Portuguese			Spanish			Turkish			Uzbek		
P	*I*	*C*	*O*	*gpt*	*mixtral*	*mixtral*	*gemini*	*gpt*	*mixtral*	*gemini*	*gpt*	*mixtral*	*gemini*	*gpt*	*mixtral*	*gemini*	*gpt*	*mixtral*	*gemini*	*gpt*	*mixtral*	*gemini*	*gpt*	*mixtral*	*gemini*	*gpt*	*mixtral*	*gemini*	*gpt*	*mixtral*	*gemini*	*gpt*	*mixtral*
E	E	S	E	10.48	6.12	-	19.51	22.94	20.58	16.2	17.14	16.9	6.63	7.03	6.94	10.86	9.69	10.53	13.79	16.77	14.28	14.20	18.22	15.50	13.9	18.41	16.99	13.23	13.21	10.49	6.02	4.02	5.33
S	S	S	S	11.0	1.07	-	15.75	21.06	19.67	18.86	18.87	16.45	4.96	7.67	6.11	7.22	11.22	9.70	13.29	16.86	12.31	16.25	17.46	15.98	13.47	17.72	16.73	7.93	14.04	9.90	0.0	3.93	5.40
S	S	S	E	11.51	5.34	-	18.27	22.13	19.79	12.52	18.26	15.12	3.40	5.16	6.10	0.00	7.44	8.27	30.15	15.76	13.49	13.79	18.79	16.60	13.57	18.64	16.60	9.38	16.09	9.71	3.93	4.33	4.74
S	E	E	S	10.16	9.38		15.63	18.88	20.42	19.83	17.53	15.67	5.80	8.20	7.90	6.33	10.29	8.24	14.07	16.72	13.81	14.38	17.15	15.93	14.31	17.61	17.09	8.99	12.51	10.54	2.5	4.53	5.39
E	E	S	S	11.13		-	21.61	20.98	20.04	21.98	21.49	16.85	10.64	7.72	8.55	11.46	10.86	10.64	17.68	17.81	17.71	18.32	17.56	17.83	18.15	17.93	17.64	14.23	15.22	12.95	2.75	6.75	9.41
S	E	Z	E	11.31	9.92	-	22.08	22.11	20.72	20.58	18.89	16.41	9.23	8.89	6.54	0.00	12.44	7.21	14.14	18.78	14.44	18.19	18.25	16.87	17.90	18.32	17.97	13.34	15.4	13.21	10.40	6.39	5.61
E	E	Z	E	9.89	10.32	-	16.92	21.51	17.96	15.26	17.79	18.39	5.06	7.41	8.32	9.89	8.3	9.12	14.43	15.04	14.71	15.64	16.73	15.09	14.05	18.04	17.37	9.52	11.99	11.04	10.23	6.19	5.87
S	E	S	E	9.85	7.43	-	17.19	21.75	19.12	16.18	18.32	17.87	5.25	7.76	3.24	9.03	10.24	8.4	12.65	18.3	N.A	14.46	18.24	15.51	9.49	17.94	16.87	13.39	14.12	N.A	4.33	7.03	4.51
E	S	Z	S	12.26	10.42	-	21.77	21.64	20.09	20.62	18.93	18.52	8.77	8.71	8.29	13.98	11.74	11.51	17.93	18.04	17.31	18.51	16.94	16.26	17.66	17.02	17.49	14.0	14.75	13.1	5.66	8.03	8.41
S	S	Z	S	10.44	7.49	-	21.34	21.6	19.87	19.62	17.69	13.92	8.10	7.29	6.32	7.09	11.62	7.04	0.0	16.45	15.16	16.03	16.78	16.98	16.96	17.59	18.54	12.83	14.38	10.72	7.84	7.88	6.85
E	S	E	E	9.32	6.62	-	18.8	20.69	20.58	12.51	18.24	15.09	3.82	7.21	7.66	8.04	9.09	9.51	16.31	0.0	16.00	13.64	17.68	16.63	12.79	16.99	18.38	7.85	13.02	9.87	5.11	9.93	4.13
S	E	Z	S	10.79	5.35	-	19.68	21.2	19.91	20.54	17.99	19.71	6.52	8.55	7.37	10.53	11.31	11.34	13.47	18.61	15.42	16.52	17.35	16.06	15.50	18.27	17.05	9.97	14.75	10.52	4.58	8.27	6.38
E	S	E	S	9.79	8.11	-	17.99	21.21	15.85	10.75	19.69	17.98	3.17	8.55	2.67	8.98	10.71	8.15	14.61	17.29	13.88	15.09	17.39	17.99	14.41	18.86	17.41	9.61	13.74	7.96	3.44	8.06	5.66
E	E	E	S	12.44	10.9	-	23.06	21.66	28.57	20.81	19.47	19.64	9.79	7.88	9.04	12.61	11.08	12.33	17.49	18.49	18.43	17.29	16.97	16.59	16.96	17.33	17.77	14.53	15.07	13.83	0.00	9.2	9.68
S	E	E	E	11.16	11.54	-	19.67	20.64	21.02	15.06	19.51	19.3	6.52	8.95	7.85	9.19	10.97	12.07	12.77	18.32	16.80	14.75	16.32	15.86	14.68	17.19	16.38	8.98	13.5	12.52	6.46	10.14	9.18
S	S	Z	E	10.48		-	21.58	21.82	19.82	20.97	21.12	19.41	9.35	8.96	8.16	13.05	12.69	12.28	17.19	19.66	17.51	17.69	18.78	15.83	17.02	17.44	16.67	15.48	16.01	13.6	11.44	10.09	9.09
S	E	S	S	11.07	7.14	19.13	17.22	21.32	21.3	14.14	18.95	18.11	4.98	6.90	6.12	10.23	11.92	8.81	15.32	18.85	12.94	14.74	17.93	17.77	14.58	17.58	14.55	12.46	13.26	11.14	2.88	10.62	3.95
S	S	E	E	11.38	8.07	-	20.45	22.75	21.44	11.54	18.26	16.85	4.54	9.53	3.24	7.51	14.11	8.02	13.19	18.84	12.45	15.75	18.08	17.67	15.79	17.82	17.51	9.77	15.35	9.94	5.97	9.15	6.07
E	E	E	E	12.65	10.63	-	20.2	22.09	21.47	16.71	19.94	19.66	6.35	9.08	8.13	9.76	12.54	11.43	14.1	21.08	16.41	15.89	19.39	16.27	15.85	18.81	16.69	12.15	17.37	13.6	6.65	10.29	9.08
S	S	E	S	12.46	2.89	-	16.04	22.52	19.34	17.26	21.86	17.34	3.74	8.9	7.71	11.59	13.77	9.55	14.6	19.36	15.95	14.03	17.09	14.85	13.40	17.76	13.98	5.75	15.0	11.7	4.49	10.88	5.41
E	S	Z	E	12.1	6.25	-	16.07	22.22	20.77	15.62	21.02	16.2	3.89	8.96	4.5	6.72	13.97	7.87	15.05	18.95	11.90	15.65	17.18	15.08	14.86	18.42	13.94	9.83	15.65	12.56	10.62	11.30	4.41
E	S	S	S	11.19		-	21.07	15.96	22.19	19.41	18.39	-	9.44	9.11	8.41	11.46	12.26	12.20	17.66	18.24	16.70	17.3	16.4	17.38	17.45	17.62	17.28	13.62	13.3	13.96	2.28	10.26	8.22
E	S	Z	S	4.65	9.56	-	21.32	20.57	20.27	21.78	18.03	17.15	9.44	8.69	8.26	5.99	11.89	11.01	17.84	17.67	18.30	17.29	16.12	17.45	18.40	17.27	17.52	15.38	13.29	12.24	6.51	10.60	8.21
E	S	S	E	12.19	8.65	-	14.83	21.26	18.81	10.31	20.6	12.81	4.33	8.59	6.40	7.77	13.86	6.64	0.00	18.01	13.31	14.92	17.11	16.15	12.34	19.00	11.16	6.8	14.3	9.65	4.65	12.72	6.67

Table 18: Comparing performance of different models on all languages in XlSUM. Metric: ROUGE1 Score.

Question Answering					Summarization					Named Entity Recognition					NLI		
code	instruction	context	Examples	output	code	instruction	context	Examples	output	code	instruction	context	Examples	output		instruction	context	Examples
ar	-0.02	**0.30****	-0.10**	**0.24****	az	-0.14**	-0.01	0.04	**-0.34****	zh	-0.07**	**0.44****	**-0.26****	0.00	ar	-0.02	-0.01(	-0.05
as	-0.05	**0.35****	-0.00	**0.30****	fr	-0.07*	0.03	**0.19****	-0.04*	fr	0.01	0.10	-0.11*	-0.01	bu	-0.00	-0.02	0.03
be	-0.10*	**0.39****	-0.00	**0.30****	ja	0.15**	-0.02	**0.34****	-0.07	it	0.01	0.04	0.02	-0.04	zh	0.01	-0.01	-0.01
ge	0.03	0.07*	-0.09**	0.06	ko	0.04	-0.01	**0.25****	-0.06	po	0.01	0.09*	-0.15**	0.02	ge	0.01	0.05	-0.07
gr	-0.02	**0.19****	-0.09**	0.12**	ne	0.00	0.02	0.13*	-0.25**	sr	0.05	**0.44****	**-0.26****	0.09	gr	0.02	-0.02	0.04
hi	0.04	**0.31****	-0.13**	**0.30****	fa	-0.08	0.07	**0.21****	-0.05	sk	-0.01	0.21**	-0.11	-0.04	hi	-0.02	-0.01	-0.04
ma	0.10**	**0.31****	0.05	-0.01	po	0.03	-0.02	**0.25****	-0.02	sw	0.01	0.06*	-0.11	-0.03	es	-0.01	0.00	0.02
ro	0.03	-0.07*	0.05	0.05	es	0.02	0.03	**0.22****	-0.03	bam	-0.00	0.07	-0.05	0.05	sw	-0.05	0.01	-0.02
ru	-0.03	**0.27****	-0.09**	**0.22**	tr	0.04	0.12*	**0.19****	-0.06	ewe	-0.00	0.02	-0.07	0.03	th	-0.07	-0.01	-0.01
te	-0.07	**0.40****	0.06	**0.20**	uz	0.00	-0.00	**0.28****	**-0.21****	hau	-0.09**	-0.08	-0.17	-0.14	tu	-0.01	-0.04	-0.03
vi	0.04	**0.13****	-0.04	0.04*						yo	0.00	0.12*	-0.13	0.00	ur	-0.02	-0.01	0.03

Table 19: Point-biserial correlation of Gemini for each Language (denoted by ISO 639 code) nd each of the 4 prompt components - Instruction, Context, Examples, and Output. The p-value is given in the parentheses

Question Answering					Summarization					Named Entity Recognition					NLI			
code	instruction	context	examples	output	code	instruction	context	examples	output	code	instruction	context	examples	output	code	instruction	context	examples
ar	0.04	0.07	-0.07**	0.09**	az	-0.09*	-0.09	0.03	-0.10	zh	0.16	**0.34****	**-0.29****	-0.02	ar	-0.05	0.01	0.04
as	-0.16	0.02	-0.01**	0.04	fr	-0.07*	0.00	0.05	-0.02	fr	0.04	0.09	0.15**	-0.04	bu	0.01	0.01	-0.03
be	-0.07	**0.31****	-0.07*	0.15**	ja	-0.03	0.04	0.13*	-0.18	it	-0.00	0.04	0.14**	-0.01	zh	-0.03	0.00	-0.01
ge	0.04	0.07	0.13**	-0.05	ko	0.02	-0.04	0.08	-0.04	po	-0.01	0.08	-0.02	-0.01	ge	0.02	0.02	-0.01
gr	-0.06**	0.17	0.03	0.03	ne	-0.16	0.02	0.10	-0.10	sr	0.02	**0.36****	0.09**	0.08	gr	-0.06	0.01(	-0.04
hi	-0.07**	0.17	-0.05	0.10	fa	-0.09	-0.00	0.14	0.05	sk	0.05	0.11	0.12	0.06	hi	-0.01	-0.01	0.01
ma	**-0.23****	0.09	-0.05	-0.00	po	-0.12**	-0.02	0.06	0.04	sw	-0.03	0.01	0.13	-0.03	es	-0.01	0.00	0.03
ro	-0.02	0.16	-0.14**	-0.00	es	-0.06	0.02	0.02	0.03	bam	0.04	0.00	-0.00	0.01	sw	0.01	-0.02	-0.03
ru	-0.00	**0.22****	-0.10**	0.07*	tr	-0.03	-0.04	0.09	-0.10**	ewe	-0.03	0.09	-0.05	0.01	th	-0.01	-0.01	-0.07
te	-0.07	**0.19****	-0.07	0.06*	uz	**-0.31**	-0.13**	0.10*	-0.14*	hau	0.05	**-0.27****	-0.15	0.03	tu	-0.03	-0.04	-0.09
vi	-0.00	0.10	-0.16*	0.02						yo	-0.02	0.10	-0.08**	0.03	ur	0.03	-0.01	-0.02

Table 20: Point-biserial correlation of Mixtral for each Language (denoted by ISO 639 code) nd each of the 4 prompt components - Instruction, context, Examples, and Output. The p-value is given in the parentheses

Question Answering					Summarization					NER					NLI			
code	instruction	context	Examples	output	code	instruction	context	Examples	output	code	instruction	context	Examples	output	code	instruction	context	output
ar	0.01	**0.43****	N.A	0.02	fr	-0.06*	0.12**	N.A	-0.02	zh	0.03	0.14	**-0.16****	0.00	ar	0.04**	0.05	-0.07*
as	0.05	**0.47****	N.A	-0.03	ja	**-0.17****	**-0.32****	N.A	0.05	fr	-0.01	0.06	**-0.35****	-0.01	zh	0.02	-0.02	-0.02
be	0.001*	**0.51****	N.A	0.04	ne	-0.05	0.04	N.A	0.01	it	0.04	0.02	**-0.28****	0.01	ge	0.06	0.04	0.02
ge	0.05*	-0.11**	N.A	0.02	po	-0.01	0.04	N.A	-0.02	po	0.02	0.04	**-0.31****	0.01	hi	-0.06*	0.01	-0.07
gr	0.05*	**-0.24****	N.A	,0.11	es	0.03	-0.06	N.A	0.01	sr	0.05	0.01	**-0.20****	0.021	es	0.02	-0.01*	-0.01
hi	,-0.00	**0.44**	N.A	0.01	pe	**-0.24***	**-0.25****	N.A	0.06	sk	0.01	0.02	**-0.30****	-0.02	sw	0.07	-0.01	-0.03
ma	-0.04*	**0.38****	N.A	-0.02						sw	0.02	-0.05	**-0.33****	0.02	th	-0.03	0.03	-0.01
ro	0.03	-0.14**	N.A	-0.04						bam	0.00	0.02	-0.03	-0.01	ur	0.04	0.02	0.05
ru	0.05	**0.17****	N.A	0.04						ewe	-0.03	0.00	-0.01	0.03				
te	,0.14**	**0.45*****	N.A	,-0.02						hau	,0.03	-0.01	0.00	0.01				
vi	-0.08	**0.20****	N.A	-0.02						yor	,0.01	0.06	-0.03	0.01				

Table 21: Point-biserial correlation of Bloomz for each Language (denoted by ISO 639 code) nd each of the 4 prompt components - Instruction, context, Examples, and Output. The p-value is given in the parentheses