# Beyond English: Examining the Impact of Prompt Translation Strategies in Multilingual Natural Language Tasks

**Anonymous ACL submission**


## Abstract

Despite advances in the multilingual capabilities of Large Language Models (LLMs) across diverse Natural Language Processing (NLP) tasks, English remains the dominant language for LLM research and development. This has led to the widespread practice of *pre-translation*, i.e., translating the task prompt into English before inference. *Selective pre-translation*, a more surgical approach, focuses on translating specific prompt components. However, its current use lacks a systematic research foundation. Consequently, the optimal *pre-translation* strategy for various multilingual settings and tasks remains unclear. In this work, we aim to uncover the optimal setup for *pre-translation* by systematically assessing its modes of use. Specifically, we view the prompt as a modular entity, composed of four functional parts: instruction, context, examples (zero-shot / few-shot), 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, and assessing various capabilities including Question Answering (QA), Natural Language Inference (NLI), Named Entity Recognition (NER), and Abstractive Summarization. Our experiments uncover the impact of factors as translation quality, similarity to English, and the size of pre-trained data, on the model performance with pre-translation. Finally, we suggest practical guidelines for choosing the optimal strategy in various multilingual scenarios.


## 1 Introduction

Large language models (LLMs) demonstrate impressive multilingual capability (Muller et al., 2020) across various natural language processing tasks, including machine translation (Kocmi et al., 2023), knowledge utilization, and complex reasoning (Zhao et al., 2023). The capabilities of LLMs stem from the extensive volume of training data they were trained on (Kaplan et al., 2020).

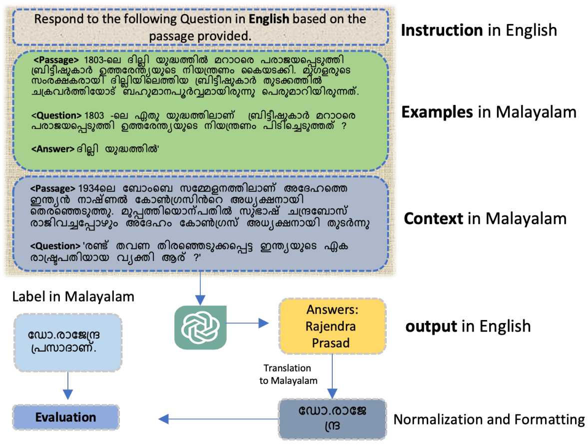



Figure 1: Our Selective Pre-Translation Approach

Current LLMs are primarily trained on English data but also include data from other languages. For example, GPT-3 was trained on 119 languages, but only 7% of the tokens are from non-English languages[1]. However, over 7,000 languages are spoken worldwide (Anderson, 2010), and as globalization accelerates and the use of various languages, the need for prompting LLMs in multilingual tasks has grown (Huang et al., 2023; Qin et al., 2023b).

One common strategy, known as the *pre-translation* approach, involves translating the prompt into English before querying the model (Ahuja et al., 2023; Shi et al., 2022), allowing to leverage the robust capabilities in the English language for multilingual tasks. Previous research has shown that using this approach, LLMs such as GPT are capable of performing a wide variety of language tasks and outperform monolingual prompts when the task is presented in English (Bareiß et al., 2024; Intrator et al., 2024; Chowdhery et al., 2023; Qin et al., 2023a; Ahuja et al., 2023). At the same time, this approach introduces complexities and risks of information loss (Nicholas and Bhatia,

[1] https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_word_count.csv

2023), impacting both efficiency and effectiveness. It's also unclear whether this approach is uniformly effective across all tasks, especially those requiring region-specific or cultural knowledge, such as Named Entity Recognition (NER). Moreover, recent studies (Intrator et al., 2024) show that direct inference of the complete prompt outperforms complete pre-translation for both discriminative and generative tasks, such as QA and summarization.

In view of the shortcomings, various studies assess *selective pre-translation*, a more nuanced method compared to the traditional *pre-translation* approach, by translating specific parts of the prompt. For example, Liu et al. (2024) evaluates instructions in both the source language and English while keeping the context in the source language. Ahuja et al. (2023) translated few-shot examples to English, and Huang et al. (2023) prompted LLMs to translate the question into English and solve the problem step-by-step in English. However, these methods lack systematic evaluation of more complex setups, such as instruction in English and output in the source language. Consequently, the optimal *pre-translation* prompt configuration for various multilingual settings and tasks remains unclear. Additionally, to our knowledge, only a few works (Jain et al., 2019) effectively use translation for tasks that have no straightforward alignment between labels in translated text such as NER. Motivated by this research gap, in our study, we examine the impact of translation in diverse tasks.

To close this research gap, in this paper we formalize the prompt as consisting of four functional parts: instruction, context, examples (zero/few-shot), and output — either of which could be *selectively* pre-translated or not (as also stated by Winata et al. (2021); Ahuja et al. (2023)). With this concept, we create a framework for exhaustively analyzing 24 configurations of cross-lingual prompt translation (using English and different source languages). Figure 1 demonstrates a schema of our *selective pre-translation* approach using an example in the Malayalam language. By formalizing the prompt as multi-functional, we decompose its constituent elements and systematically evaluate the influence of translating each segment on overall performance.

Through a large-scale evaluation encompassing 35 distinct languages, four tasks, six datasets, and three models, we demonstrate that *selective pre-translating* prompts consistently surpass both *pre-translation* of the entire prompt and *direct inference* approaches, establishing the effectiveness of *selective pre-translation*. Additionally, we analyze how various factors, including the type of task, size and family of pre-trained data, language similarity to English, impact the performance of our proposed *selective* approach. Furthermore, based on these factors, we provide guidelines for implementing optimal pre-translation strategies. Finally, we perform an additional experiment, analyzing the impact of translation on *pre-translation*.

Our findings demonstrate that in extractive tasks such as extractive 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 prioritizes the context in the source language in the case of low-resource languages. Surprisingly, we have discovered that low-resource languages yield better results when the output is in English, even in cases where there is no direct alignment between the original and translated context, such as in NER. Moreover, we find that the quality of translation significantly influences model performance, with specific configurations yielding optimal results for different languages and tasks.

## 2 Background

### 2.1 The Rise of Multilingual Large Language Models

With over 7,000 languages spoken globally (Anderson, 2010), globalization and 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 the latter approach, Multilingual LLMs, to operate in different languages (Raffel et al., 2020; Conneau et al., 2019; Chowdhery et al., 2023) comes from two sources: (1) fine-tuning on multilingual data in order to transfer knowledge and 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; Shi et al., 2022). This latter approach has gained popularity due to its efficiency and applicability to a wider range of models.

Following these developments, benchmarks for evaluating LLMs have been proposed to measure cross-lingual transfer, including low-resource languages (Hu et al., 2020; Liang et al., 2020) and benchmarks focusing on specific language families such as Indian languages (Kakwani et al., 2020) and African languages (Ogundepo et al., 2023).

### 2.2 Multilingual Prompting Approaches

Researchers have developed various prompting methods to improve the multilingual capabilities of LLMs. Huang et al. (2023) introduced XLT, a cross-lingual prompt that directs LLMs to function as experts 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 outperformed finetuning in cross-lingual transfer. Shi et al. (2022) found that chain-of-thought (CoT) prompting leads to remarkable multilingual reasoning abilities in LLMs, even in under-represented languages. Another common strategy is *pre-transaltion* which translates the entire prompt to English (Chowdhery et al., 2023; Qin et al., 2023a; 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. However, this approach lacks a systematic research foundation. In this study, we examine this approach systematically and assess additional setups of pre-translation, to yield an empirically-grounded set of recommended modes of us.

## 3 The Proposal: A Selective Pre-Translation Prompting Approach

LLMs exhibit two remarkable qualities for effective NLP task-solving: (i) the ability to follow instructions (Wei et al., 2022), enhancing their performance in complex tasks through a series of intermediate reasoning steps, and (ii) in-context learning (Brown et al., 2020), where the model learns tasks from limited examples without weight update. The basis for the implementation of these two capabilities is the notion of the *prompt*, that acts as a general prefix for the LLM to generate a response to. Our approach formalizes the prompt accordingly, splitting it into four functional parts instruction, context, examples, and output to leverage these capabilities.

Specifically, we define them as follows. The **Instruction** ($I$) provides natural language guidance to the model, explaining the task to be performed with the task data. The **context** ($X$) represents the actual task data that the model processes. Few-shot **Examples** ($\{(x_i, y_i)\}_{i=1}^{k}$) are optional examples that are used for in-context learning and are denoted as $E$. Given these components, the final prompt is formulated as $(I\_X\_E)$ for the few-shot setting and $(I\_X)$ for the zero-shot setting, where $\_$ denotes concatenation. Finally, the prompt is processed by a model $M$ which yields an **Output** $O$, whose format and language can be controlled by the instruction. $O = M(I\_X\_E^n)$ where $n \geq 0$. Subsequently, we perform normalization and formatting on the model's output before evaluation.

## 4 Selective Pre-Translation Evaluation

In this section, we evaluate various selective pre-translation strategies and assess their impact on model performance. We systematically alter the language (English/Source) of each component: instruction, context, examples, and output. We examine 24 configurations per task, 16 ((English/Source)$^4$) for few-shot and 8 ((English/Source)$^3$) for zero-shot, except for NLI where the output is in English.

### 4.1 Experimental Setup

**Models** We conducted benchmarking experiments on several LLMs: OpenAI GPT-3.5-turbo (0125) (Ouyang et al., 2022), Mixtral-8x7B-Instruct-v0.1 (Jiang et al., 2023), and Gemini-1.0-pro (Team et al., 2023), which have context sizes of 16k, 32k, and 8k respectively. Appendix A.1 includes the platforms used for utilizing the models.

**Prompt Creation** We developed a framework for efficient multilingual prompt construction, validation, and model querying to streamline our comprehensive evaluation process. In addition, we applied different normalization phases such as converting to lowercase and removing extra spaces. Appendix A.1 provides further details on the prompt creation.

**Language Selection and Categorization** For all tasks, we selected 10-12 languages, ensuring an equal distribution across representation levels (High-, Medium-, Low-Resource). Due to the unavailability of precise distribution information

| *Affinity* | Class | *Range (% of tokens)* | *Avg. #tokens* | *STD* |
|---|---|---|---|---|
| High Resource | A | P >= 0.1% | 1240 | 1156 |
| Medium Resource | B | 0.01% < P < 0.1% | 72 | 49 |
| Low Resource | C | 0% < P < 0.01% | 5.07 | 5.41 |
| Extremely Low | D | P = 0% | 0 | 0 |

Table 1: Language Classification by GPT-3 Pre-Training Data Size.

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

Table 2: Experiment Setup: Tasks, datasets, languages.

for each model, we used the GPT-3 distribution as a proxy[2] Additionally, its extensive coverage enables us to categorize languages into classes based on their data ratios. We categorized the tested languages into four classes: High-Resource Languages (A), Medium-Resource Languages (B), Low-Resource Languages (C), and languages unrepresented in the data (D). To determine the subsets, we use the class division proposed by Lai et al. (2023) and modified class D to include only the languages that are unrepresented in the GPT-3 pre-trained data, addressing the gap of evaluating this subset (Ahuja et al., 2023). Table 1 provides a summary of this classification and basic properties. Other divisions exist; for example, Joshi et al. (2020) proposed dividing languages based on the number of speakers, However, this approach does not fully capture language diversity in LLMs, which are more influenced by the language data availability than how widely spoken it is.

**Analysis Methods** We employed three primary methods: (i) *Correlation analysis* – Assesses the relationship between the model's prediction scores and the chosen language component. This component is represented by a binary vector, where 0 indicates English and 1 indicates the source language. (ii) *Association Rule Learning (ARL) and 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. To address this limitation, we utilize *association rule learning (ARL)* with the Apriori algorithm (Piatetsky-Shapiro, 1991; Hegland, 2007). Appendix A.2.1 includes implementation details and a short recap of the algorithm. (iii) *Performance gap* – For each configuration $X_i^E$ (component $X$ in English) and $X_i^S$ (component $X$ in the source language), we calculate the difference between all the configurations with component $X$ in English to those with Source and divide by 12 (half of the configurations), which is given by the following formula: Average Gap for $X = \frac{1}{12}\sum_{i=1}^{12}\left(E(X_i^E) - E(X_i^S)\right)$.

[2]Although using the GPT-3 distribution is not optimal, we chose to use it due to its extensive multilingual coverage.

## 4.2 Tasks and Datasets

All datasets used, spanning multiple languages, are listed in Table 2. In total, these datasets encompass 35 distinct languages across 10 different language families. We test four tasks, as follows.

**Natural Language Inference (NLI)** NLI involves determining if a hypothesis is entailed by, contradicts, or is neutral to a premise. We evaluated this using the XNLI dataset (Conneau et al., 2018), which contains premise-hypothesis pairs in multiple languages. The input is a pair of sentences, and the output is a classification label: entailment, contradiction, or neutral. We measure performance using accuracy, comparing the model's predictions to the dataset's ground truth labels.

**Question Answering (QA)** We conducted extractive question answering, where the model predicts the answer to a question based on a provided context. We evaluated our model on two datasets: XQuAD (Artetxe et al., 2019) and IndicQA for Indic languages (Doddapaneni et al., 2022). The input consists of a question and a context passage, and the model's output is the predicted span of text within the context that answers the question. Performance evaluation is done by comparing the model's predicted answer spans with the ground truth answer spans provided in the dataset, using the F1 score as the evaluation metric.

**Named Entity Recognition (NER)** In this task, the model is instructed to identify and classify named entities within a given sentence. We employed two datasets for evaluation: WikiANN (Pan et al., 2017) and MasakhaNER (Adelani et al., 2021). WikiANN is a widely used NER dataset containing Wikipedia sentences annotated with LOC (Location), PER (Person), and ORG (organization) tags which supports 176 languages. MasakhaNER is specifically designed for African languages.

| Question Answering (QA) | | | | | Summarization | | | | | Named Entity Recognition (NER) | | | | | Natural Language Inference (NLI) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Lang* | *F1* | ↑ *Src. (%)* | ↑ *Eng. (%)* | *Class* | *Lang* | *ROUGE1* | ↑ *Src. (%)* | ↑ *Eng. (%)* | *Class* | *Lang* | *F1* | ↑ *Src. (%)* | ↑ *Eng. (%)* | *Class* | *Lang* | *Acc.* | ↑ *Src. (%)* | ↑ *Eng. (%)* | *Class* |
| en | 0.77 | N.A | N.A | A+ | en | 30.23 | N.A | N.A | A+ | en | 0.65 | N.A | N.A | A+ | en | 0.69 | N.A | N.A | A+ |
| de | 0.85 | 18.05 | 9.00 | A | fr | 35.12 | 16.45 | 10.41 | A | sr | 0.77 | 52.62 | 265.50 | B | sw | 0.73 | 58.89 | 28.33 | C |
| hi | 0.82 | 32.25 | 182.76 | C | ja | 32.47 | 17.48 | 14.83 | A | it | 0.75 | 9.33 | 41.47 | A | bg | 0.72 | 57.84 | 8.36 | C |
| ar | 0.74 | 84.99 | 138.70 | B | fa | 29.34 | 21.35 | 0.00 | C | sk | 0.72 | 15.37 | 36.86 | B | el | 0.71 | 24.35 | 30.28 | B |
| vi | 0.73 | 0.00 | 58.69 | B | es | 28.28 | 10.55 | 3.43 | A | po | 0.72 | 18.46 | 20.67 | A | es | 0.69 | 20.78 | 18.14 | A |
| ro | 0.69 | 0.00 | 9.52 | A | po | 27.40 | 8.26 | 0.00 | A | fr | 0.72 | 23.60 | 24.09 | A | ar | 0.67 | 28.57 | 23.08 | B |
| ru | 0.69 | 6.15 | 305.80 | A | tr | 20.87 | 18.11 | 0.00 | B | hau | 0.70 | 62.11 | 51.55 | C | hi | 0.64 | 59.28 | 8.48 | C |
| el | 0.69 | 0.00 | 2.98 | B | ne | 19.58 | 31.36 | 28.69 | C | ee | 0.68 | 46.03 | 81.47 | D | de | 0.64 | 19.46 | 8.99 | A |
| bn | 0.68 | 44.68 | 423.07 | D | as | 15.79 | 17.92 | 7.07 | D | sv | 0.68 | 12.07 | 9.25 | A | zh | 0.63 | 16.17 | 4.65 | B |
| as | 0.56 | 138.46 | 450.00 | D | uz | 15.72 | 58.56 | 24.87 | C | zh | 0.63 | 90.00 | 121.22 | B | th | 0.57 | 49.75 | 10.61 | B |
| te | 0.53 | 231.25 | 253.30 | C | ko | 11.84 | 36.99 | 11.78 | B | bam | 0.33 | 33.25 | 80.24 | D | ur | 0.57 | 29.96 | 9.08 | C |
| ml | 0.49 | 104.30 | 600.00 | C | | | | | | yor | 0.32 | 66.02 | 49.19 | D | tr | 0.57 | 0.00 | 8.14 | B |

Table 3: Highest-Performing Prompt Translation Configuration: Improvement (%) over *direct inference* (*Src.*), and over few-shot pre-translation (*Eng.*). Generated with GPT-3.5-Turbo.

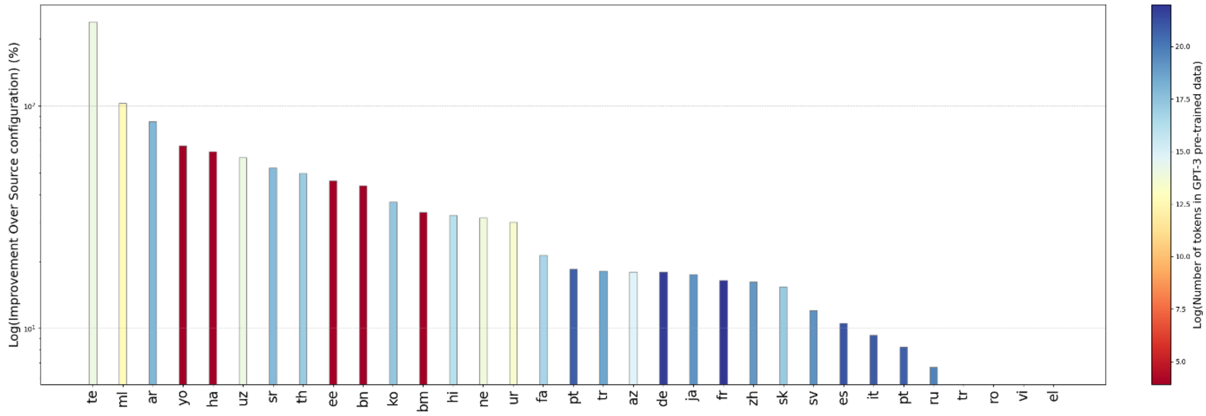



Figure 2: *Selective Pre-Translation* (best configuration) improvement (%) over *Direct Inference*.

Both datasets utilize the BIOSE scheme (B=Begin, I=Inside, O=Outside, S=Singleton, E=End) to mark entity boundaries. However, for our experiments, we rephrased the task as a generative task, instructing the model to directly output entity spans without requiring strict adherence to the BIOSE tags. We evaluated the model using the F1 metric.

**Abstractive Text Summarization** Abstractive summarization is the task where the model generates concise and informative summaries from longer texts by generating new text, unlike extractive summarization, which selects existing sentences. We evaluated our model using the XL-Sum dataset (Hasan et al., 2021), which provides summaries of news articles in 44 languages, making it ideal for evaluating multilingual summarization models. We used the ROUGE metric for evaluation, which measures the overlap between generated summaries and reference summaries.

## 4.3 Results

In this section, we present the results of the *selective pre-translation* approach, demonstrating its advantage over two common strategies: *direct inference* (source language only) and *pre-translation* (English only). Later, we display the optimal configurations for each task and analyze how each configuration part relates to overall performance, emphasizing considerations in prompt selection.

### 4.3.1 Selective Pre-Translation Advantage

Table 3 shows each language's highest-performing configuration score among all distinct configurations (24 per language).[3] Additionally, we display the improvement (%) of this configuration over *direct inference* and *pre-translation*.

**Improvement Over Pre-Translation** 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, low-resource languages demonstrate a greater improvement of 65% on average compared to high-resource languages.

**Improvement Over Direct-Inference** The results reveal that 90% of the languages show improvement over the basic *direct inference* configuration. Similar to the pre-translation approach, low-resource languages like Telugu and Assamese

[3] Appendix A.4.3 displays all language/task configurations.

**QA**

| *lang* | *instruction* | *context* | *examples* | *output* | *class* |
|---|---|---|---|---|---|
| ru | -0.08** | **0.35**** | 0.12** | 0.09** | A |
| de | -0.03** | **0.30**** | 0.08 | 0.03* | A |
| ro | -0.03 | 0.12** | 0.04 | 0.02 | A |
| vi | 0.04 | **0.40**** | 0.10** | 0.10 | B |
| ar | -0.07** | **0.20**** | 0.13** | 0.04* | B |
| el | -0.06 | **0.48**** | 0.03 | 0.07* | B |
| bn | -0.10** | **0.38**** | 0.03 | 0.03 | C |
| ma | -0.14** | **0.30**** | 0.01 | 0.03 | C |
| te | -0.10** | **0.38**** | 0.03 | 0.03 | C |
| hi | -0.07** | **0.30**** | 0.05 | 0.01 | C |
| as | -0.04** | **0.30**** | 0.06 | 0.06 | D |

**Summarization**

| *lang* | *instruction* | *context* | *examples* | *output* | *class* |
|---|---|---|---|---|---|
| ja | **-0.33**** | -0.08 | -0.02* | 0.00 | A |
| fr | 0.01 | 0.020 | -0.04 | 0.06 | A |
| po | -0.08* | 0.05* | -0.03* | 0.10* | A |
| es | -0.09* | 0.03* | -0.03 | 0.05 | A |
| tr | -0.14** | 0.10 | -0.1 | -0.03* | B |
| ko | -0.10** | 0.13 | 0.01 | 0.05 | B |
| uz | **-0.42**** | 0.14 | 0.03 | -0.12* | C |
| fa | **-0.37**** | 0.05 | -0.07** | -0.04 | C |
| ne | **-0.35**** | -0.09 | 0.07** | -0.14 | C |
| az | **-0.30**** | 0.04 | -0.00 | -0.05 | C |

**NER**

| *lang* | *instruction* | *context* | *examples* | *output* | *class* |
|---|---|---|---|---|---|
| fr | -0.11* | 0.10* | -0.01 | 0.01 | A |
| it | 0.02 | 0.04 | -0.04 | 0.01 | A |
| po | -0.15 | 0.09* | 0.1 | 0.01 | A |
| sv | -0.11* | 0.06* | -0.03** | 0.01 | A |
| zh | -0.26** | **0.44**** | 0.00 | 0.07 | B |
| sr | -0.26** | **0.44**** | 0.09** | 0.05 | B |
| sk | -0.11** | **0.30**** | -0.1* | 0.01 | B |
| bam | 0.03 | **0.44**** | -0.11 | 0.02 | D |
| ewe | -0.01 | **0.38**** | -0.12** | 0.01 | D |
| yo | -0.01 | **0.36**** | 0.01* | 0.03 | D |
| hau | -0.04 | **0.30**** | 0.08* | 0.02 | D |

**NLI**

| *lang* | *instruction* | *context* | *examples* | *class* |
|---|---|---|---|---|
| de | -0.03 | -0.02 | -0.01 | A |
| es | -0.03 | 0.02 | -0.03* | A |
| el | -0.04 | 0.01 | 0.07 | B |
| zh | 0.01 | -0.06 | -0.06 | B |
| ar | 0.00 | -0.02 | -0.04 | B |
| th | -0.03 | 0.03 | -0.14* | B |
| tr | -0.02 | 0.00 | 0.02 | B |
| ur | 0.01 | 0.01 | -0.08* | C |
| bg | 0.01 | 0.05 | -0.13* | C |
| sw | 0.12 | -0.06 | -0.09 | C |
| hi | -0.03 | -0.09 | -0.09** | C |

Table 4: Correlation ($\tau$) between GPT-3.5 Turbo performance and each prompt component. *p < 0.05, **p < 0.01. Positive $|\tau|$ indicates correlation with source language, negative $|\tau|$ indicates correlation with English language.

**High Resource Languages**

| *Task* | *Law* | *Support* | *Confidence* | *lift* |
|---|---|---|---|---|
| *NER* | (Context: S) + (Context: S) =>Percentile 70 | 0.173 | 1.0 | 2.01 |
| *QA* | (Context: S) + (Output: S) =>Percentile 70 | 0.25 | 1.0 | 2.03 |
| Summarization | (Instruction: S) + (Context: S) + (Examples: E) =>Percentile 70 | 0.1 | 0.88 | 1.57 |
| | (Instruction: S) + (Context: E) + (Examples: S) =>Percentile 70 | 0.1 | 0.88 | 1.57 |
| | (Instruction: E) + (Examples: S) + (Output: E) =>Percentile 30 | 0.1 | 1.0 | 2.0 |
| *NLI* | (Instruction: S) + (Context: S) + (Examples: E) =>Percentile 70 | 0.09 | 1.0 | 2.03 |
| Cross-tasks | (Context: S) + (Output: S) =>Percentile 70 | 0.05 | 0.24 | 0.6 |

***Low Resource Languages***

| *Task* | *Law* | *Support* | *Confidence* | *lift* |
|---|---|---|---|---|
| *NER* | (Context: S) + (Examples: S) + (Output: E) =>Percentile 70 | 0.07 | 0.75 | 0.05 |
| *QA* | (Instruction: S) + (Context: S) + (Output: S) =>Percentile 70 | 0.25 | 1.00 | 2.00 |
| *Summarization* | (Instruction: S) + (Context: S) + (Examples: E) =>Percentile 70 | 0.07 | 0.84 | 1.67 |
| | (Instruction : E) + (Examples: S) + (Output: S) =>Percentile 30 | 0.09 | 1.00 | 1.96 |
| *NLI* | (Instruction: E) + (Context: E) + (Examples: S) =>Percentile 70 | 0.17 | 1.00 | 2.00 |
| Cross-tasks | (Context, S) =>Percentile 70 | 0.06 | 0.12 | 1.8 |

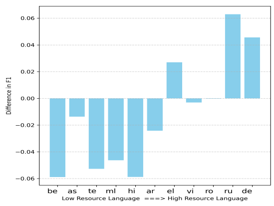
Table 5: Apriori-Based Association Rules. (prompt part: English/Source) => Top X% scores over configurations.

demonstrate a relative improvement of > 100%. High-resource languages also show impressive improvement; for example, French and Portuguese exhibit an improvement of > 20% in NER tasks. Figure 2 displays the percentage improvement of choosing the highest configuration score compared to the *direct inference* score, highlighting that languages with lower representation achieve better improvement than those with higher representation.

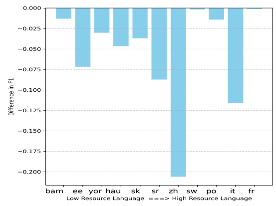
### 4.4 The Optimal Configuration

In this section, we analyze how individual prompt components influence the performance of GPT-3.5-Turbo. Table 4 displays the correlation between model performance and different configuration components. Table 5 summarizes the optimal configurations for achieving the best results.

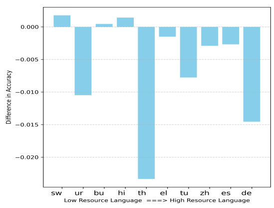
**Effects of Context Language Selection** Table 4 indicates that extractive tasks, such as QA and NER, benefit from including source-language context. This effect is particularly pronounced for low-resource languages (Class C/D), exhibiting a 70% higher correlation coefficient with source-language context compared to high-resource languages. Conversely, tasks like abstractive summarization and NLI seem to be agnostic to context translation. Our rule-association analysis in Table 5 further highlights the model's preference for source-language context in extractive tasks (NER, QA).



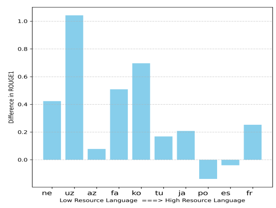
(a) QA



(b) NER



(c) NLI



(d) Summarization.

Figure 3: Few-Shot Translation Performance Gap (English - Source).

**Use of Examples** The rule association analysis in Table 5 shows that the optimal configuration requires examples (few-shot). In fact, there is no high percentile rule that includes zero-shot examples as part of the rule items. Appendix A.4 strengthens the argument for incorporating examples in prompts, especially for high-resource languages.

**Translation of Examples** Analysis of rule associations in Table 5 reveals that extractive tasks (NER, QA) benefit from incorporating source-language examples. We hypothesize that because NER is a region-specific task that requires the model's pre-trained knowledge, examples in the source language help the model augment its knowledge base in the specific language. Figure 3, aligns with these findings, depicting a similar trend in the gap between few-shot translation performance using English examples and source-language examples. Interestingly, the figure shows a preference for English examples in the summarization task. Specifically, for high-resource languages, the model performs well regardless of the examples' language.

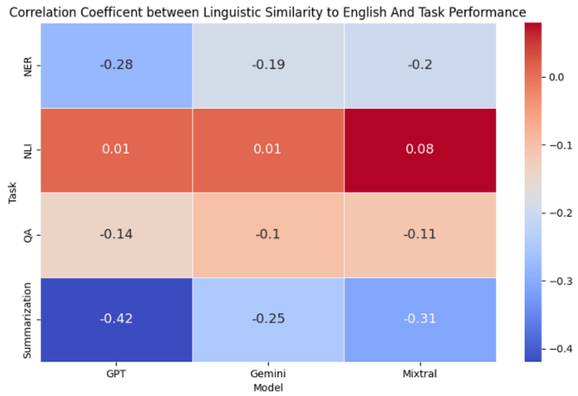



Figure 4: Person correlation between linguistic syntactic similarity to English and task performance for GPT, Mixtral, and Gemini. * p < 0.05, ** p < 0.01.

However, for low-resource languages, English examples result in better performance, as the model finds it easier to generate text.

**Output Selection Effects** Unlike context and examples, the output relates to the model's ability to produce coherent and grammatical text. The rule association analysis in Table 5 shows that for extractive tasks like extractive QA and NER the output should be in the source language for all languages. However, for generative tasks like summarization, English output demonstrates higher performance. This might be attributed to limitations in the model's ability to generate fluent and informative text in languages other than English. Surprisingly, NER in low-resource languages benefits from English output, despite context (Source)-output (English) mismatch.

### 4.5 Factors Explaining Performance

**Pre-Training Data Size Impact** Table 3 results show that for QA, NER and summarization tasks, languages in classes A and B generally achieve better results than those in classes C and D in, indicating that more pre-trained data yields better performance. However, notable exceptions exist. Hausa and Ewe (Class C/D) achieve better results on the NER task compared to Swedish and Chinese (Class A/B). Interestingly, this dependence on pre-training data volume appears attenuated for the NLI task. Here, we observe several Class C/D languages outperforming Class A/B languages. This finding suggests that NLI tasks might leverage distinct aspects of the model's learned language representation, potentially rendering them less susceptible to limitations in pre-training data quantity.

**Linguistic Similarity To English** We employed the pre-computed distances from the URIEL dataset (Littell et al., 2017), using phonological and syntactic distances from English. Figure 4 reveals a strong positive correlation between model performance and syntactic features, particularly for the summarization task, indicating that syntactic similarity to English significantly enhances performance in this task. Also, NER exhibits significant correlations, which suggests that models can better identify and classify entities in texts when they share syntactic features with English.

## 5 The Impact of Translation Quality on Pre-Translation

We aim to provide the factors influencing translation quality and its impact on performance. Understanding these factors is essential since translation is a fundamental component in the *selective pre-translation* approach.

### 5.1 Experimental Setup

**Models** To evaluate the translation quality we used two translator engines: *Google Translate API*[4] and *Bing Translator API* over Azure.

**Data** To evaluate translation quality we focus our experimentation on languages featured in the validation set of the FLORES-200 (Guzmán et al., 2019) benchmark, which offers extensive support for numerous low-resource languages. Specifically, we selected translation directions from 91 languages to English, aligning with the scenario of translating prompts to English. To outline our language selection process, we referenced the lists of supported languages provided by both Google Translate[5] and Bing Translator[6]. Each language subset comprised 997 sentences, along with human translations into English.

**Evaluation** We evaluate translation quality by translating each sentence using both Google and Bing translation engines. The resulting machine translations are then compared to human-generated references. We employed five established metrics: (1) n-gram matching metrics – Meteor (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), and BLEU

[4]We used the open-source library https://pypi.org/project/easygoogletranslate/.

[5]https://cloud.google.com/translate/docs/languages

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

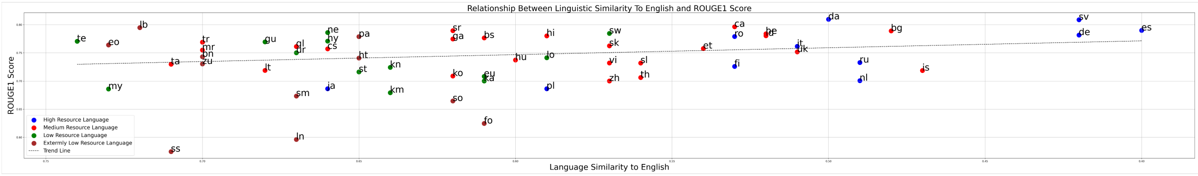



Figure 5: Scatter plot showing the relationship between syntactic similarity to English and translation performance for four subsets of languages - High, Medium, Low, and Extremely Low. The upward Trend is represented by a positive linear regression.

(Papineni et al., 2002), and (2) neural network-based evaluation metrics – BertScore (Zhang et al., 2019) and Comet (Rei et al., 2020).

### 5.2 Results

**Machine Translation Engines Comparison** Our evaluation demonstrates that Google Translate consistently yields the highest overall performance compared to Bing Translator for most of the chosen languages. On average, Google Translate achieved a 15% improvement in BLEU score over Bing Translator when evaluated across 70 languages, highlighting it's superior performance. Consequently, we chose Google Translate as our main machine translation engine. Appendix C provides additional analysis and a full breakdown of the comparative findings.

**Pre-Trained Data Size Impact** To determine the impact of the size of a language in the pre-trained data, we used the data ratio of the language in the GPT-3 unlabeled pre-training data. We found no significant correlation between a language's token proportion and its translation quality. Subsequently, we analyzed the influence of linguistic similarity to English on translation quality. We employed the pre-computed distances from the URIEL dataset (Littell et al., 2017) which is based on typological information. We calculated the correlation between the linguistic similarity scores (genetic, geographical, inventory, phonology, and syntactic) and the corresponding translation scores for each subset of resource languages. Figure 5 visualizes the relationship between a language's syntactic similarity to English and translation performance, measured by ROUGE-1. It reveals a significant positive correlation (coefficient=0.33, p-value=0.01), indicating that higher syntactic similarity to English corresponds with better translation quality, particularly for high-resource languages (coefficient=0.73, p-value=0.004). Appendix C provides the full results of the correlation.

**Translation Quality Impact On Downstream Tasks** To investigate this question, we utilized the XQuAD dataset for the QA task, which includes parallel splits for English and multiple other languages. Each language subset consisted of 200 sentences. For each sentence, we calculated the translation score by translating the context into English and then computed the F1 score. Our analysis revealed a weak positive Pearson correlation of 0.233 (p-value<0.001) for the GPT-3.5 Turbo model. However, further research is needed to validate these findings on additional datasets and tasks.

## 6 Conclusion

This research conducts a comprehensive assessment of translation-based prompting strategies in multilingual LLMs across 4 tasks, 6 datasets, 3, models, and 35 distinct languages. These tasks include traditional NLP challenges as well as generative and region-related tasks. Our evaluation is the first to assess all existing prompting configurations for translation systematically. We demonstrate that *selective pre-translation* prompts consistently suppress both *pre-translation* of the entire prompt and *diret-inference* - all in source approach, establishing the effectiveness of *selective pre-translation* in both high and low resource languages. Additionally, we found various factors, including the type of task, size and family of pre-trained data, language similarity to English, and translation quality impact the model performance. We hope that the practical guidance we provide will assist people in effectively prompting LLMs in multilingual scenarios.

## Limitations

This study aims to systematically assess the effectiveness of various prompting strategies across different tasks and LLMs. Due to limitations in computer resources, it was not possible to evaluate more advanced models such as GPT-4. However, we endeavored to cover several LLMs representing different architectures. In our evaluation, we attempted to influence the output by instructing the model to generate in a specific language. We acknowledge that sometimes the model may not follow our instructions and produce output in another language, which could affect the results. Appendix B provides error analysis on the different problems we dealt with. 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.

## 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.

Patrick Bareiß, Roman Klinger, and Jeremy Barnes. 2024. English prompts are better for nli-based zero-shot emotion classification than target-language prompts. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1318–1326.

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*.

Markus Hegland. 2007. The apriori algorithm–a tutorial. *Mathematics and computation in imaging science and information processing*, pages 209–262.

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

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*.

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*.

Alankar Jain, Bhargavi Paranjape, and Zachary C Lipton. 2019. Entity projection via machine translation for cross-lingual ner. *arXiv preprint arXiv:1909.05356*.

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*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

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*.

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.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv preprint arXiv:2004.01401*.

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*.

Benjamin Muller, Antonis Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2020. When being unseen from mbert is just the beginning: Handling new languages with multilingual language models. *arXiv preprint arXiv:2010.12858*.

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

Odunayo Ogundepo, Tajuddeen R Gwadabe, Clara E Rivera, Jonathan H Clark, Sebastian Ruder, David Ifeoluwa Adelani, Bonaventure FP Dossou, Abdou Aziz

Diop, Claytone Sikasote, Gilles Hacheme, et al. 2023. Afriqa: Cross-lingual open-retrieval question answering for african languages. *arXiv preprint arXiv:2305.06897*.

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. 2023a. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023b. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. *arXiv preprint arXiv:2310.14799*.

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*.

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*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

# A Modular Multilingual Translation Prompting Approach

## A.1 Experimental Setup

**Models** To query GPT-3.5-turbo (0125), we used the Azure platform via the API[7]. For Mixtral-8x7B-287 Instruct-v0.1, we utilized the API platform provided by Together.ai[8]. Lastly, for Gemini-1.0-pro, we accessed the API through Google AI Studio[9].

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

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

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

| *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 д IV 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] | Insuffcent 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 6: Error analysis of unexpected model outputs and observed in various tasks/languages.

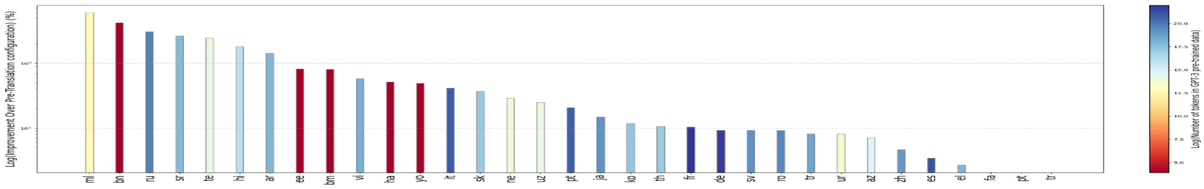


Figure 6: Percentage of improvement over *pre-translation* configuration, when using the highest configuration for each languages/task For GPT-3.5-Turbo. The bars are color-coded based on the norm of the log of the number of tokens in the pre-trained data of GPT-3, as elaborated in Table 7

**Prompt Creation** For constructing the prompts we used the LangChain library[10] 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.

**Python Libraries In Use** For evaluation of the different models, we used the most common ROUGE package for non-English papers[11]. for loading and processing the data, we used NumPy[12] For help with writing the code, we used assistance from ChatGPT.

**Normalization And Formatting** Before evaluation, we normalized the model's output, with each task having its unique normalization process. For the QA task, for instance, 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, with each tuple in the format (Tag, Entity). After the normalization phase, we conducted further formatting if necessary. For example, in the NER task, we transformed the normalized output into a list in the BIOSE format. This involved 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').

[10] `https://pypi.org/project/langchain/`

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

[12] `https://pypi.org/project/numpy/`

## A.2 Methods

### A.2.1 Rule Association

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

(iii) **Lift**: $\text{lift}(X \to Y) = \frac{c(X \to Y)}{s(Y)}$, measures how much more likely itemset $Y$ is to occur when itemset $X$ is present compared to when $X$ is absent.

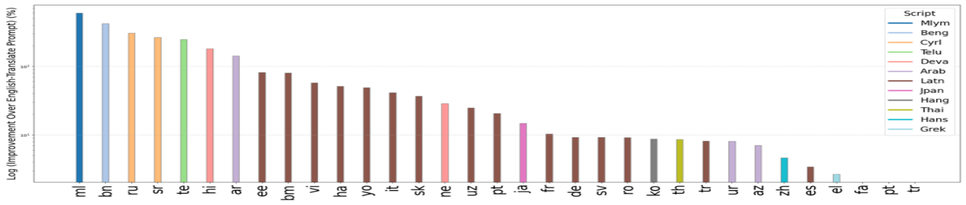
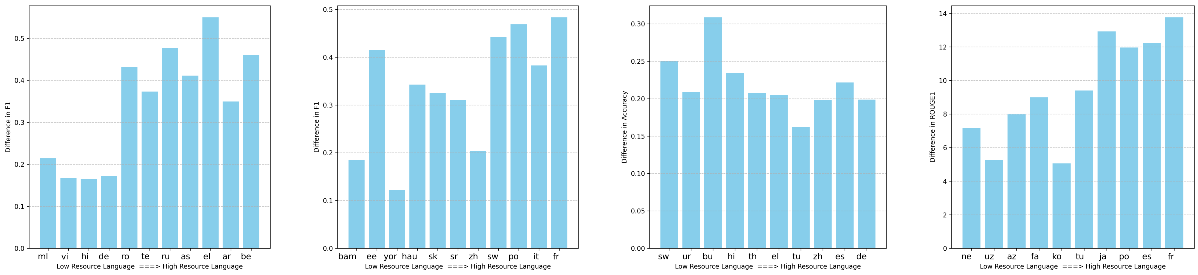



Figure 7: Percentage improvement over *pre-translation* approach, when using the highest configuration for each task For GPT-3.5-Turbo. The bars are color-coded based on the language family script.



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

Figure 8: Few-Shot and Zero-Shot Performance Gap (Few-Shot - Zero-Shot) for each task/language.

**Implementation Details** To implement the Rule Association algorithm, we created a DataFrame for each task's results using pandas DataFrames[13]. 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[14] library, which produces two outputs - itemsets and rules. Later, we filtered weak rules (support > 0.05 & confidence > 0.75).

### A.3 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*.

**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.

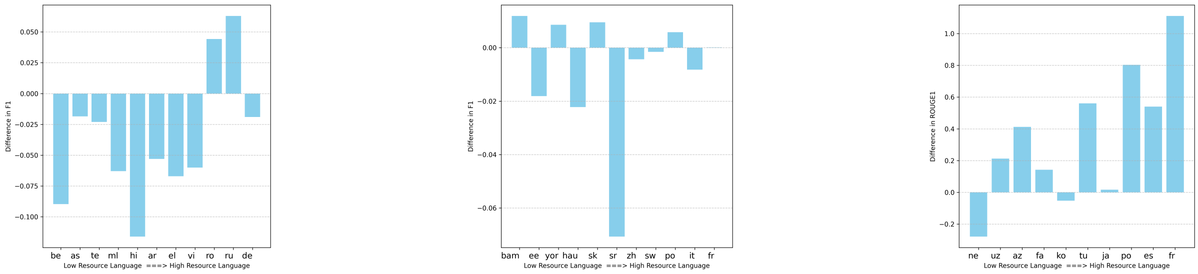
**Natural Language Inference** You are an NLP assistant whose purpose is to solve Natural Language Inference (NLI) problems. NLI is the task of determining the inference relation between two texts: entailment, contradiction, or neutral. Your answer should be one word from the following: entailment, contradiction,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

### A.4 Results

#### A.4.1 The Optimal Configuration

**Use of Examples** Figure 8 demonstrates that for all tasks, using a few-shot setting over a zero-shot

[13]https://pypi.org/project/pandas/
[14]https://pypi.org/project/efficient-apriori/




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

Figure 9: Output Performance Gap (English - Source) for each task/language

| 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 7: 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.

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.

| Model | QA | NER | Summarizatin |
|---|---|---|---|
| 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

**Output Selection Effects** Figure 9 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

**Language Family Impact** Figure 7 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.

### A.4.3 Detailed Results

The results across all tasks, languages and models are included in our benchmarking exercise are provided in Table 11 (for XQuAD), 12 (for indciQA), 14 (for WikiANN), 15 (for MasakhNER), 16 (for XL-Sum), 13 (for XNLI). The result of the correlation for Gemini are included in Table 17 and for Mixtral in Table 18.

| 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 9: Correlation between syntactic similarity to English and the ROUGE score (by language subset).

# 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 6 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) 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 The Effect of Translation Quality

**Machine Translation Engines Comparison** The results in Table 10 and Figure 10 demonstrate that Google Translate API outperformed Bing Translator in all the evaluated metrics, highlighting its high performance. Interestingly, the languages that achieved the highest scores are Welsh and Maltese which are both considered low-resource languages.

**Linguistic Similarity To English** The results in Table 9 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 correlation was observed in languages belonging to the high-resource category, and this correlation decreases as the class of the language becomes low-resource.

| | 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 10: Comparsion between Google Translate API and Bing Translator.

| 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 11: Comparing performance of different models on all languages in XQuAD. Metric: F1 Score.

| 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 12: 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 | 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 13: Comparing performance of different models on all languages in XNLI. Metric: Acc 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 14: 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 15: 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 | E | 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 |

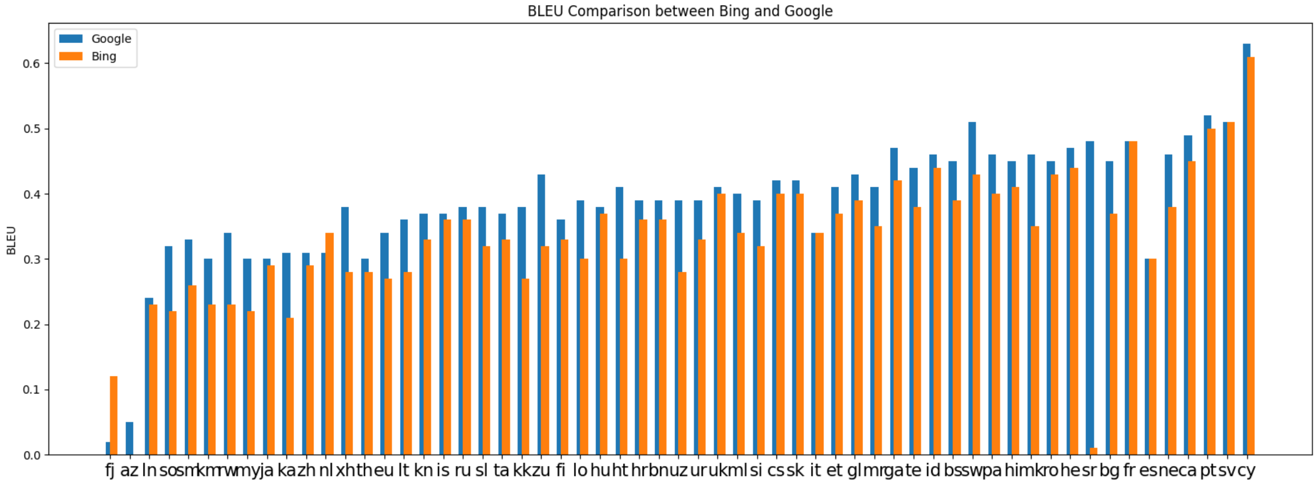
Table 16: Comparing performance of different models on all languages in XlSUM. Metric: ROUGE1 Score.

| 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 |

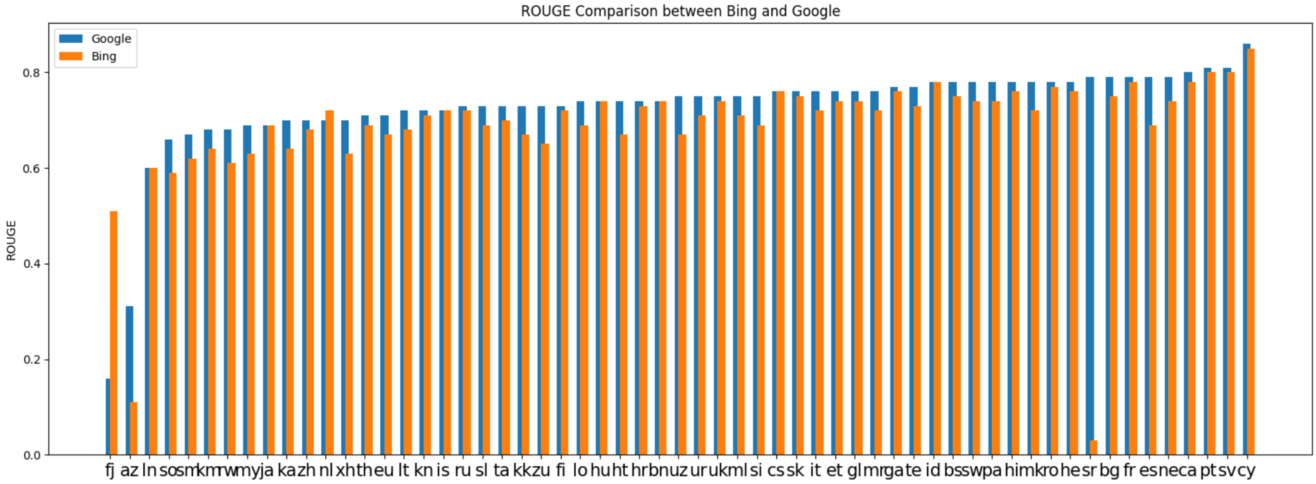
Table 17: Point-biserial correlation of Gemini for each Language (denoted by ISO 639 code) nd each of the 4 prompt components - Instruction, Context, Examples, and Output. The p-value is given in the parentheses

| 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 |

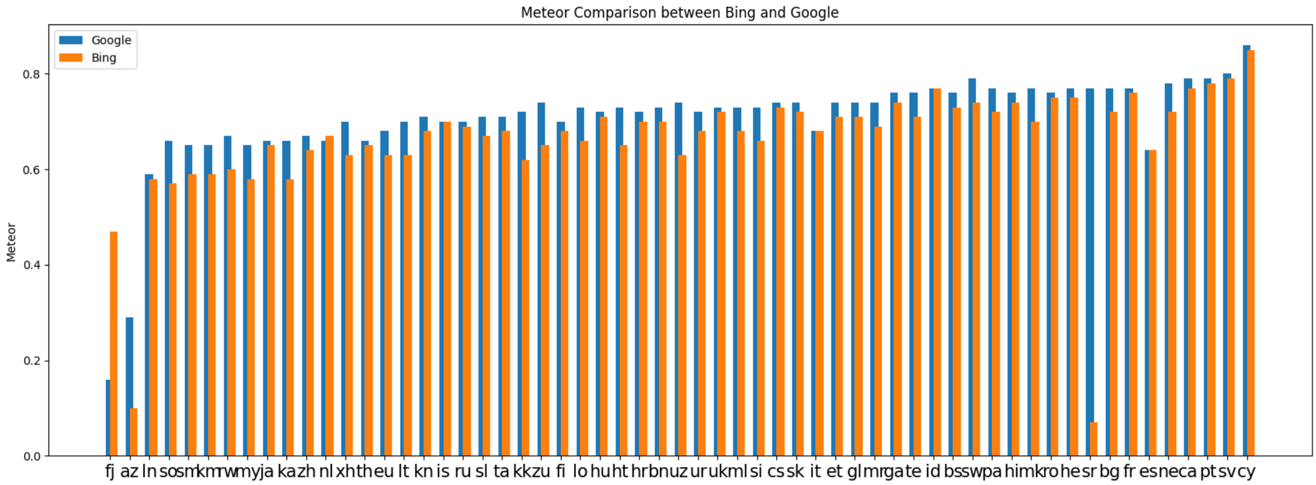
Table 18: Point-biserial correlation of Mixtral for each Language (denoted by ISO 639 code) nd each of the 4 prompt components - Instruction, context, Examples, and Output. The p-value is given in the parentheses




(a) BLEU




(b) ROUGE




(c) Meteor

Figure 10: Google Translate API vs Bing Translator Comparsion