Blessing of Multilinguality: A Systematic Analysis of Multilingual In-Context Learning

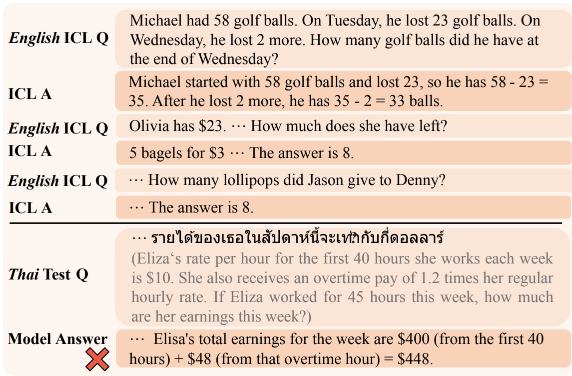
Yilei Tu[1] **Andrew Xue**[1] **Freda Shi**[1,2,3]
[1]David R. Cheriton School of Computer Science, University of Waterloo
[2]Vector Institute
[3]Canada CIFAR AI Chair
yileitu.tt@gmail.com, fhs@uwaterloo.ca

Abstract

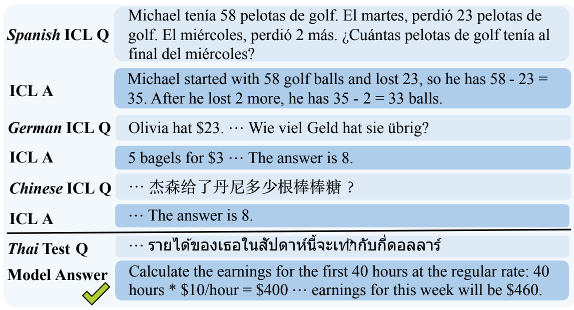
While multilingual large language models generally perform adequately, and sometimes even rival English performance on high-resource languages (HRLs), they often significantly underperform on low-resource languages (LRLs; Costa-jussà et al., 2022). Among several prompting strategies aiming at bridging the gap, multilingual in-context learning (ICL; Shi et al., 2023) has been particularly effective when demonstration in target languages is unavailable. However, there lacks a systematic understanding of when and why it works well.

In this work, we systematically analyze multilingual ICL, using demonstrations in HRLs to enhance cross-lingual transfer. We show that demonstrations in mixed HRLs consistently outperform English-only ones across the board, particularly for tasks written in LRLs. Surprisingly, our ablation study shows that the presence of irrelevant non-English sentences in the prompt yields measurable gains, suggesting the effectiveness of multilingual exposure itself. Our results highlight the potential of strategically leveraging multilingual resources to bridge the performance gap for underrepresented languages.

https://github.com/yileitu/multilingual_icl



(a) English ICL Mode



(b) Multilingual ICL Mode

Figure 1: Illustration of two ICL modes. After providing a few-shot prompt, we evaluate LLM in the same domain in various languages. In a controlled experiment, each demonstration in (a) and (b) shares the same meaning, albeit in different languages. Contents and languages of demonstrations are randomly sampled from a training set and a preset high-resource language list, respectively. We find that the `Multilingual` ICL mode (b) is more effective in helping the LLM solve tasks in different languages compared to the `English` ICL mode (a).

1 Introduction

In-context learning (ICL; Brown et al., 2020) has become a widely adopted technique in natural language processing with large language models (LLMs; Touvron et al., 2023; Chowdhery et al., 2023; Dubey et al., 2024; Yang et al., 2024a,b, *inter alia*), which enables LLMs to learn to solve problems by analogy from a few input-output examples (i.e., demonstrations) without updating model parameters. As a generic method, ICL has also been effective in improving the cross-lingual performance of multilingual LLMs (MLLMs; Winata et al., 2021; Ahuja et al., 2023; Shi et al., 2023; Razumovskaia et al., 2024; Asai et al., 2024).

Prior work has introduced two common ICL strategies for non-English languages: (i) translating the target question into English and performing English-only ICL (Shi et al., 2023; Ahuja et al., 2023; Razumovskaia et al., 2024, *inter alia*), and (ii) providing demonstrations in the target language (*in-language demonstrations;* Fu et al., 2022; Qin et al., 2023; Huang et al., 2023; Zhang et al.,

Findings of the Association for Computational Linguistics: ACL 2025, pages 6213–6248
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

2024b). Both strategies have critical shortcomings: (i) may suffer from information loss in translation due to nuanced language gap (Zhang et al., 2023; Poelman and de Lhoneux, 2024) or the unavailability of high-quality translation systems for extremely low-resource languages (LRLs), whereas (ii) may become infeasible due to data scarcity in LRLs. As alternatives, when presenting LLMs with problems in the target language, English demonstrations (Fig. 1a) lead to poor performance on LRLs, whereas demonstrations in multiple high-resource languages (HRLs; Fig. 1b) can be more effective, even when there is little alphabetical overlap between the demonstration and target languages (Shi et al., 2023); however, the underlying reasons on why it works remain unclear.

In this work, we systematically analyze multilingual ICL through a set of controlled experiments. Each test question is paired with a set of semantically equivalent demonstrations, while the demonstration languages vary according to different ICL modes (§ 3). We compare the performance differences across four ICL modes (§§ 4.1 and 4.2): English, individual-HRL, multilingual (i.e., mixed HRLs), and in-language demonstrations (Fig. 2). To disentangle the impact of demonstration language from other confounding factors (e.g., interactions between demonstration languages and in-domain demonstrations), we conduct additional control experiments by adding irrelevant sentences in various languages into English-only in-domain demonstrations (§ 4.3). We find that

- In-context demonstrations in HRLs, especially in languages with non-Latin writing systems such as Chinese and Japanese, can more effectively transfer knowledge compared to English ICL mode, leading to performance improvement in answering questions in all languages, especially in LRLs. This finding is generalizable enough across different domains and various LLMs.

- Demonstrations in mixed HRLs are more robust and effective compared to those in a single HRL, in terms of average accuracy boosting on different tasks. This strategy is favored.

- Surprisingly, simply introducing another non-English language (not necessarily in the demonstration) in the prompt could lead to performance improvement, albeit the improvement is not as significant as the aforementioned strategies.

2 Background: Multilingual ICL

In this section, we review the basic approaches of ICL with instruction-tuned LLMs (§ 2.1) and introduce the multilingual prompting modes (§ 2.2) that we evaluate in this work.

2.1 ICL for Instruction-Tuned LLM

Instruction-tuned LLMs (Ouyang et al., 2022; Mishra et al., 2022; Wang et al., 2022; Wei et al., 2022a) generally possess the capability to follow task instructions (i.e., *system prompt*), which are usually coupled with ICL to fully elicit their capability (Wei et al., 2022b, *inter alia*). Formally, denote a training set by $\mathcal{D}_{\text{train}} = \{(q_i, a_i)\}_{i=1}^{M}$ and a test set $\mathcal{D}_{\text{test}} = \{(q_j, a_j)\}_{j=1}^{N}$ in the same domain, where q_i is a task question (as model *input*). An ICL prompt for a test question $q_{\text{test}} \in \mathcal{D}_{\text{test}}$ has three core components: (1) a system prompt I_{sys} that describes the task and specifies the expected output format, (2) K sample input-output pairs (K-shot) from the training set $\{(q_k, a_k)\}_{k=1}^{K} \sim \mathcal{D}_{\text{train}}$ that provide in-context demonstrations, and (3) a verbalizer V mapping each ground truth label a_i to a textual representation, which may also include reasoning steps (i.e., chains of thoughts, or CoT in short; Wei et al., 2022b). In summary, an ICL prompt for q_{test} can be written as:

$$\begin{aligned}\text{prompt}_{q_{\text{test}}} = I_{\text{sys}} \circ q_1 \circ V(a_1) \circ q_2 \circ V(a_2) \\ \circ \cdots \circ q_K \circ V(a_K) \circ q_{\text{test}}, \qquad (1)\end{aligned}$$

where $\circ$ is the string concatenation operator with a special end-of-turn (`EOT`) token as the delimiter. The LLM with parameters $\boldsymbol{\theta}$, denoted as $p_{\boldsymbol{\theta}}$, then generates the response $\hat{a}_{\text{test}}$ given $\text{prompt}_{q_{\text{test}}}$: $\hat{a}_{\text{test}} \sim p_{\boldsymbol{\theta}}\left(\text{prompt}_{q_{\text{test}}}\right)$.

2.2 Multilingual Prompting Modes

We extend our notations as follows to adapt to the multilingual settings. A training set with L languages is denoted by $\mathcal{D}_{\text{train}} = \left\{\mathcal{D}_{\text{train}}^{\text{lang}_1}, \ldots, \mathcal{D}_{\text{train}}^{\text{lang}_L}\right\}$, where the split $\mathcal{D}_{\text{train}}^{\text{lang}_\ell}$ consists of M examples for any $\ell \in \{1, \ldots, L\}$. The same applies to the test dataset $\mathcal{D}_{\text{test}}$, with each language-specific split consisting of N examples. Without further specification, we assume that the training examples at the same index are semantically equivalent across languages.

The ground-truth labels a_i in quantitative LLM benchmarks (Ponti et al., 2020; Cobbe et al., 2021, *inter alia*) are typically language-agnostic (such as

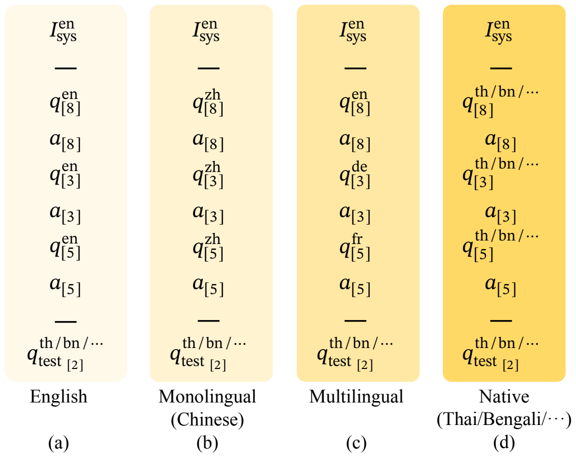


Figure 2: Illustration of ICL modes by Eq. (1). Assume $K = 3$ and $M = 10$. For the second datapoint of the test set (regardless of its language split, e.g., q_{test} could be in Thai, Bengali, etc.), we first randomly generate $K = 3$ indices from $\{1, \cdots, 10\}$, say $\{8, 3, 5\}$. Next, we determine the languages of the $K = 3$ demonstrations. For modes (a), (b), and (d), the language is uniformly specified. For mode (c), we randomly select $K = 3$ languages, say $\{\text{en}, \text{de}, \text{fr}\}$. Then $\{(8, \text{en}), (3, \text{de}), (5, \text{fr})\}$ determines each demonstration.

numbers) or represented in a single word (such as Yes/No). In such cases, the verbalizer is an identity. For answers requiring reasoning steps, $V(a_i)$ is CoT in English, as there has been strong evidence that MLLMs perform better when generating English (Shi et al., 2023; Qin et al., 2023; Huang et al., 2023, *inter alia*). For the same reasons, I_{sys} is always presented in English as well.

Following Ahuja et al. (2023) and Shi et al. (2023), we evaluate MLLMs via several different prompting strategies (*ICL modes*) in this work:

The `English` mode. The K demonstrations are always in English (Figs. 1a and 2a).

The `Monolingual` mode(s). The K demonstrations are always presented in a single non-English high-resource language such as Chinese (Fig. 2b).

The `Multilingual` mode. From a predefined list of high-resources languages $\mathcal{L}_H$, K languages are randomly selected, which, together with the sampled K indices, determine the contents and languages of the K demonstrations (Figs. 1b and 2c).

The `Native` mode. The K demonstrations are in the same language as the test question (Fig. 2d).

3 Experiment Setups

Models. We evaluate state-of-the-art instruction-tuned LLMs with about 8 billion parameters, which have officially claimed multilingual capabilities in model release: Llama3-8B-Instruct, Llama3.1-8B-Instruct (Dubey et al., 2024), Qwen2-7B-Instruct (Yang et al., 2024a), Qwen2.5-7B-Instruct (Yang et al., 2024b), Mistral-NeMo-12B-Instruct (MistralAI, 2024) and Aya-Expanse-8B (Dang et al., 2024). For additional references, we evaluate OpenAI closed-sourced commercial models, including GPT3.5-turbo (OpenAI, 2022) and GPT4o-mini (OpenAI, 2024). Detailed model cards can be found in App. A.1.

Datasets. We evaluate the models using multilingual benchmarks from 4 distinct domains: (1) MGSM (Shi et al., 2023), a benchmark of 250 grade-school math problems sampled from the English GSM8K (Cobbe et al., 2021) and translated into 10 additional languages by expert native speakers. (2) XCOPA (Ponti et al., 2020), a commonsense reasoning benchmark that extends the COPA dataset in English (Roemmele et al., 2011) to 11 additional languages. (3) XL-WiC (Raganato et al., 2020), a cross-lingual word-in-context understanding dataset spanning 13 languages, where models are expected to tell whether a polysemous word retains the same meaning in two contexts. (4) XQuAD (Artetxe et al., 2020) is designed to evaluate cross-lingual question answering performance, based on the English SQuAD (Rajpurkar et al., 2016) dataset and professionally translated into 10 languages.

MGSM, XCOPA and XQuAD are *parallel* where each corresponding datapoint across different language splits contains semantically equivalent content, allowing us to minimize semantic confounders in our experimental design. XL-WiC is language-specific and translation-variant thus naturally non-parallel. Dataset properties and examples are in Tabs. 1 and 6. Details of data curation are in App. A.2.

Languages. Languages with large-scale digitized data resources on the web are known as *high-resource languages* (HRLs; Bender, 2019), which are exemplified by English, Spanish, and Chinese, among others. In contrast, *low-resource languages* (LRL) have scarce accessible data (Costa-jussà et al., 2022). However, a universal standard for dichotomizing languages as either high- or low-resource has not been set (Bender, 2019; Joshi et al., 2020; Hedderich et al., 2021). Moreover, none of the models we evaluate has disclosed the language distributions in their training corpora. As a workaround, we define our preset HRL list as the union of the 20 most frequent languages in Llama2 (Touvron et al., 2023) and PaLM (Chowdhery et al.,

Dataset *Datapoint Example*	**Domain**	**Expected Output**	**#Languages (#HRL + #LRL)**	**Language Split Size** M/N **— Training/Test**	**En Avg Word Count$_{\pm \text{std}}$**	**Parallel**
MGSM	Mathematical Reasoning	Numerals	$11\ (7+4)$	$8/250$	$46.26_{\pm 18.29}$	✔
See Fig. 1 for examples.						
XCOPA	Commonsense Reasoning	"1" / "2"	$12\ (5+7)$	$100/500$	$26.59_{\pm 3.41}$	✔
Premise: The man turned on the faucet. What happened as a RESULT? Hypothesis 1: The toilet filled with water. Hypothesis 2: Water flowed from the spout. Answer: 2.						
XL-WiC	Word Disambiguation	"Yes" / "No"	$13\ (9+4)$	$98/390$	$35.65_{\pm 5.36}$	✖
*Sentence 1: What did you *get* at the toy store? Sentence 2: She didn't *get* his name when they met the first time. Question: Is the word "get" (marked with *) used in the same way in both sentences above? Answer: No.*						
XQuAD	Extractive QA	Text	$12\ (8+4)$	$190/1000$	$137.84_{\pm 59.35}$	✔
Passage: The Panthers defense gave up just 308 points, ranking sixth in the league, ⋯ Question: How many points did the Panthers defense surrender? Answer: 308.						

Table 1: Dataset properties and examples. Each language split (for both training and test) is of the same size. In-context demonstrations are randomly drawn from the training dataset. Data source and languages are documented in Tab. 6 in App. A.2. Texts in blue represent *interfaces* acting like reserved-words in programming languages (Poelman and de Lhoneux, 2024). The language of these interfaces changes accordingly with the input source language. "En Avg Word Count$_{\pm \text{std}}$" represents the average word count and standard deviation of the demonstration questions of the English split.

2023), and classify languages out of the HRL list as LRLs. Details of the preset HRL list can be found in App. A.3.

Prompts. Following Shi et al. (2023), we use $K = 6$ examples for demonstration for any test question. For each multilingual dataset (Tab. 1), we first sample N index lists of length $K = 6$ all at once, where the index range is $\{1, 2, \cdots, M\}$. We allocate the i-th index list to the set of test questions $q_{\text{test}_i} = \left\{ q_{\text{test}_i}^{\text{lang}_1}, q_{\text{test}_i}^{\text{lang}_2}, \cdots, q_{\text{test}_i}^{\text{lang}_L} \right\}$ for the same index i across all L language splits. The training-set index list, together with the specified languages,[1] jointly determines the content and language of the demonstration for each testing example (Fig. 2). This approach both ensures linguistic diversity for multilingual prompting and, whenever applicable, mitigates confounding factors that come with semantic inconsistency across examples. All *interface* words (see Tab. 1) are in the same language as the examples rather than in English.

Inference and metrics. Throughout this work, we use greedy decoding for inference, selecting the token with the highest probability at each step. For MGSM in need of CoT, we set the maximum token length to 500; for XCOPA and XL-WiC, we set it to 10, as we expect the answers to be short; for XQuAD, it is 30. We use exact match accuracy as our evaluation metric: for MGSM, we extract the last numeral in the response; for XCOPA and XL-WiC, we extract the label (*expected output*) in the response (Tab. 1); for XQuAD, a response of an MLLM is considered correct if it contains the gold answer as a substring after normalization.

[1]For the `Multilingual` mode, we apply the same sampling procedure to generate N HRL code lists of length $K = 6$, drawing from the available HRLs in the dataset.

4 ICL Mode Evaluation

4.1 `Multilingual` Prompts Surpass `English`

Results. We first compare the 6-shot performance with `English`, `Multilingual` and `Native` ICL modes on 6 open-sourced MLLMs and 2 commercial OpenAI models (Fig. 3), with Tab. 2 presenting the detailed performance of three selected MLLMs across various LRLs. Overall, `Multilingual` mode outperforms `English` mode, both for individual LRLs and on average. In 23 out of 30 cases (Fig. 3), `Multilingual` mode achieves higher accuracy than `English` one. This phenomenon is evident even for GPT4o-mini, one of the currently strongest LLMs (Chiang et al., 2024), and for HRLs as well (see Tabs. 11 to 13 in App. B.1 for HRL accuracies). Extending the results of Shi et al. (2023) that `Multilingual` mode generally outperforms `English` mode for PaLM (Chowdhery et al., 2023) and Codex (Chen et al., 2021), our results confirm this trend is a general phenomenon across various MLLMs and datasets.

Outstanding performance of the `Native` mode and the practical unfeasibility. Admittedly, in 27 out of 30 comparisons, `Native` mode outperforms `Multilingual` mode (Fig. 3), which aligns with the machine-learning intuition that in-domain data, in terms of both genre and language, are more promising for model performance (Liu et al., 2024). However, domain-specific datasets for LRLs are often difficult to obtain due to the scarcity of na-

Acc (%)$_{\Delta\uparrow\downarrow}$	MGSM bn	sw	te	th	LRL Avg	XL-WiC bg	et	fa	hr	LRL Avg	XCOPA LRL Avg
Llama3.1-8B-Instruct											
English	52.40	67.20	39.60	69.20	57.10	51.54	44.62	29.74	51.79	44.42	55.91
Multilingual	68.00	68.80	55.60	71.60	$66.00^{***}_{8.90\uparrow}$	57.69	55.13	57.95	56.92	$57.05^{***}_{12.63\uparrow}$	$66.11^{***}_{10.20\uparrow}$
Native	67.20	72.40	58.00	76.40	$68.50^{***}_{11.40\uparrow}$	61.79	62.05	62.31	57.18	$62.88^{***}_{18.46\uparrow}$	$71.63^{***}_{15.72\uparrow}$
Qwen2-7B-Instruct											
English	57.20	20.80	23.20	73.60	43.70	28.21	56.92	63.33	54.87	48.46	62.29
Multilingual	64.80	28.40	23.60	73.20	$47.50^{**}_{3.80\uparrow}$	53.59	58.46	60.26	54.36	$56.28^{***}_{7.82\uparrow}$	$63.83^{*}_{1.54\uparrow}$
Native	72.00	32.40	40.40	76.80	$55.40^{***}_{11.70\uparrow}$	55.13	59.23	63.08	54.62	$57.76^{***}_{9.30\uparrow}$	$67.63^{***}_{5.34\uparrow}$
GPT3.5-turbo											
English	39.60	63.60	12.80	60.80	44.20	54.36	54.62	54.10	51.79	53.72	63.43
Multilingual	54.40	68.00	24.40	57.20	$51.00^{***}_{6.80\uparrow}$	52.82	60.00	54.36	56.41	$55.90^{**}_{2.18\uparrow}$	$62.71_{0.72\downarrow}$
Native	57.20	73.60	30.00	58.80	$54.90^{***}_{10.70\uparrow}$	54.62	59.49	58.46	60.26	$58.21^{***}_{4.49\uparrow}$	$67.17^{***}_{3.74\uparrow}$

Table 2: Accuracies on LRLs of `English`, `Multilingual` and `Native` modes across 3 MLLMs of 3 datasets. Please refer to Tab. 9 for language code-to-name mapping. Avg represents the average accuracy of the LRLs. Subscript indicate the performance increase↑ (or decrease↓) of `Multilingual` and `Native` compared to `English`. Superscripts are significance levels (in terms of p-value) of the same comparison — *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$. Raw evaluation accuracies and hypothesis test results for all MLLMs and all languages are in Tabs. 11 to 13 and Tabs. 15 to 17 in App. B.1, respectively.

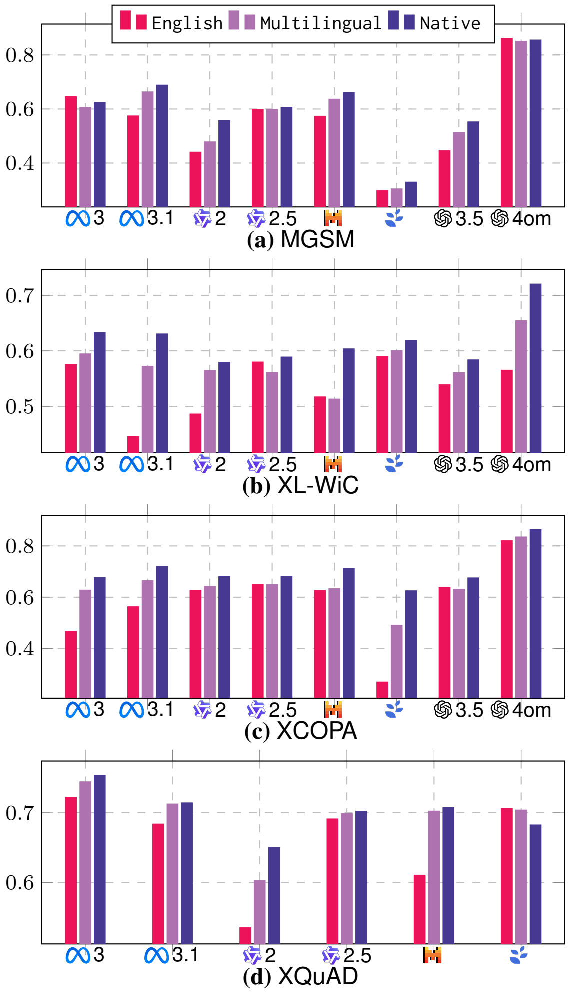


Figure 3: Average accuracies of LRLs across three ICL modes on our evaluated 4 datasets and 7 MLLMs. Raw accuracies of all language splits are in Tabs. 11 to 14 in App. B.1. For simplicity, on the x-axis, only the model logos are labeled – 3: Llama3-8B-Instruct; 3.1: Llama3.1-8B-Instruct; 2: Qwen2-7B-Instruct; 2.5: Qwen2.5-7B-Instruct; : Mistral-NeMo-12B-Instruct; : Aya-Expanse-8b; 3.5: GPT3.5-turbo; 4om: GPT4o-mini.

tive speakers or professional translators (Costa-jussà et al., 2022; NLLB-Team, 2024); therefore, in practice, it is usually challenging to provide high-quality demonstrations in the same language and domain as the test question. In contrast, annotations make the HRL-`Multilingual` mode more feasible in many scenarios.

Hypothesis Tests. To verify whether the improvement is statistically significant, we conduct McNemar's test (McNemar, 1947)—the null hypothesis means no significant accuracy difference between the baseline (`English`) and the compared mode. Let b denote the number of cases where the baseline is correct while the compared mode is incorrect, and c denote the number of cases where the baseline mode is erroneous while the compared mode is correct. We calculate the corrected version (Edwards, 1948) of the McNemar's statistic:

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c}, \tag{2}$$

which has a chi-squared distribution with one degree of freedom. Significant χ^2-test results provide strong evidence to reject the null hypothesis of no accuracy improvement. Our results results (Tab. 2 and Tabs. 15 to 18 in App. B.1) indicate that both `Multilingual` and `Native` modes significantly outperform the `English` mode.

4.2 Ablation Study: Non-English `Monolingual` Prompts Are Effective

With the success of the `Multilingual` mode, we investigate whether the improvement comes from the introduction of multiple languages or simply

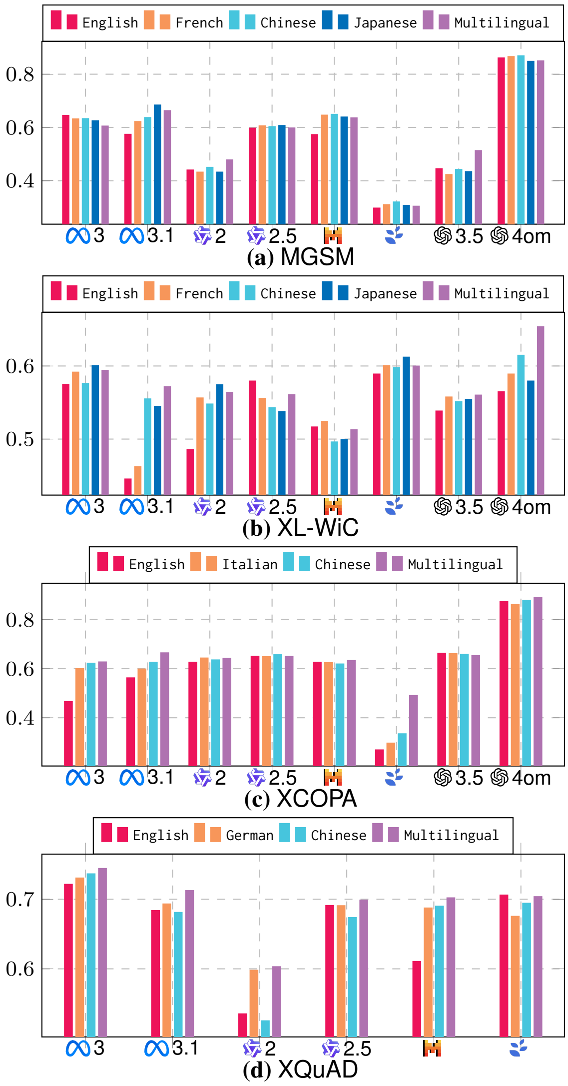


Figure 4: `Monolingual` modes vs `Multilingual` on average accuracies of LRLs. The x-axis is the same as in Fig. 3. Raw evaluation accuracies are in Tabs. 11 to 14 in App. B.1.

from a single non-English HRL. Specifically, we compare several `Monolingual` modes including `Chinese` (for all 4 datasets), `French`, `Japanese` (both for MGSM and XL-WiC), `Italian` (for XCOPA) and `German` (for XQuAD). We select French, Italian and German because they are European languages and thus share considerable subword overlap with English; in contrast, Chinese and Japanese exhibit little subword overlap with Latin-script languages, but there is a substantial overlap between the two due to their shared use of Chinese characters (or *Hanzi, Kanji*). This language selection allows us to analyze simultaneously the impact of the writing system (or *subword overlap*) on ICL performance.

Results. All HRL-`Monolingual` modes outperform `English` in a considerable number of comparisons (Fig. 4). This finding also holds for HRL evaluations (see Tabs. 11 to 14 in App. B.1 for raw accuracies of each language). Among all `Monolingual` modes, `Chinese` performs the best, with 20 out of 30 comparisons when accuracy surpasses that of `English`. `Japanese` also frequently outperforms `English`. Extending the findings of Turc et al. (2021) that non-English languages are more effective than English in pretraining and fine-tuning based cross-lingual transfer, our results suggest that non-English languages, particularly those with non-Latin scripts, may be more effective under the prompting scheme as well.

However, the `Multilingual` mode exhibits stronger robustness, outperforming `Chinese` in 23 out of 30 LRL cases. The same trend applies to HRLs. The results of the hypothesis test further confirm the robustness: for both LRL and HRL splits, the `Multilingual` mode exhibits the highest number of significant results relative to other `Monolingual` modes (Tabs. 15 to 18 in App. B.1). Intuitively, we hypothesize that `Multilingual` mode functions like an "average" of individual HRL-`Monolingual` modes, making it most robust, and thus it outperforms `English` most frequently and achieves the highest overall average accuracy.

Following Tang et al. (2024), we further identify ICL-mode-specific neurons and find that the neurons activated by `Multilingual` overlap most with those activated by `Native` among other modes. This further explains why `Multilingual` could achieve performance comparable to `Native`. See App. C for details.

$$I_{\text{sys}}^{\text{en}} \mid s_{[1]}^{\text{zh}} \; q_{[8]}^{\text{en}} \; a_{[8]} \mid s_{[2]}^{\text{es}} \; q_{[3]}^{\text{en}} \; a_{[3]} \mid s_{[3]}^{\text{ja}} \; q_{[5]}^{\text{en}} \; a_{[5]} \mid q_{\text{test }[2]}^{\text{th/bn/}\cdots}$$

English ICL Mode + Multilingual Context-Irrelevant Sentences (`English` + `CIS-Multi`)

Figure 5: Prepending multilingual CIS (`CIS-Multi`) $\{s_i^{\text{lang}}\}_{i=1}^K \sim \mathcal{S}^{\text{lang}}$ to demonstrations $\{(q_i^{\text{en}}, a_i)\}_{i=1}^K \sim \mathcal{D}_{\text{train}}^{\text{en}}$ of `English` ICL template illustrated in Fig. 2a.

4.3 Ablation Study: Merely Introducing New Language(s) Enhances ICL Performance

After showing that including non-English in the prompts can improve ICL performance, a natural follow-up question arises: does the gain come from the mere presence of non-English languages, or the interaction between the target language and the in-topic examples? To distinguish these two setups, we prepend a *context-irrelevant sentence* (CIS) s^{lang} before each ICL demonstration, which is unrelated to the current domain and can be in any language. Analogous to the ICL modes introduced in § 2.2, CIS resembles the construction of those modes with the same sampling strategy (§ 3). For example, based on the English mode, we could prepend a set of multilingual CIS, which augments Fig. 2a into Fig. 5. We denote such setting by "`English` + `CIS-Multi`". This naming convention applies to other settings accordingly.

We use sentences from FLORES-101 (Goyal et al., 2022) as the source of our CIS, which provides parallel Wikipedia sentences in multiple languages, a fairly distant genre to all evaluation datasets. We filter sentence-parallel sets with English word counts ranging from 10 to 15 as our sampling pool $\mathcal{S}^{\text{lang}}$, a small range compared to target datasets (Tab. 1), to mitigate the risk of introducing too much noise. Details of FLORES-101 can be found at App. A.2.

Results and analysis. We perform hypothesis testing as in § 4.1 to compare different CIS settings with `English` + `CIS-En` (Tab. 3)—`English` + `CIS-En` exhibit a small drop compared to `English` only, indicating that our filtration effectively controls the negative impact of noise to a tolerable level. We find that introducing a single non-English language generally improves ICL performance in most cases $\left(\frac{51}{72} \approx 71\%\right)$; however, the improvement is only statistically significant in $\frac{22}{51} \approx 43\%$ cases. This reveals that simply introducing a language can lead to a modest improvement in MLLM performance, but it is more pronounced when in-topic demonstrations in another language (`Monolingual` modes) are incorporated. More hypothesis test results for both LRL and HRL splits can be found in Tabs. 23 to 25 in App. B.2.

Introducing multiple languages (`CIS-Multi`, Fig. 5) is slightly more promising than CIS in a single language, with $\frac{18}{24} \approx 75\%$ of cases showing improvement, of which $\frac{11}{18} \approx 61\%$ are significant. We further conduct experiments by prepending multilingual CIS to the `Multilingual` mode (`Multilingual` + `CIS-Multi`, Tab. 3). The performance of `Multilingual` + `CIS-Multi` is not significantly lower than that of `English` + `CIS-Multi`; in $\frac{16}{24} \approx 67\%$ cases, it significantly outperforms `English` + `CIS-Multi`. This concludes that the significant improvement from `English` mode to `Multilingual` mode is attributed to the incorporation of multiple languages with in-topic demonstrations.

LRL $\overline{\text{Acc}}$	MGSM	XL-WiC	XCOPA	XQuAD
Llama3-8B-Instruct				
`English`	64.20	57.37	46.23	72.02
+ `CIS-En`	61.00	56.73	48.37	71.43
+ `CIS-Fr`	$61.60_{0.30\uparrow}$	$57.50_{0.77\uparrow}$	$58.49^{***}_{10.12\uparrow}$	$70.98_{0.45\downarrow}$
+ `CIS-Ja`	$61.30_{0.30\uparrow}$	$58.14_{1.41\uparrow}$	$58.69^{***}_{10.32\uparrow}$	$70.85_{0.58\downarrow}$
+ `CIS-Zh`	$59.30_{1.70\downarrow}$	$58.27_{1.54\uparrow}$	$58.03^{***}_{9.66\uparrow}$	$71.28_{0.15\downarrow}$
+ `CIS-`🗺	$60.50_{0.50\downarrow}$	$57.82_{1.09\uparrow}$	$61.14^{***}_{12.77\uparrow}$	$70.90_{0.53\downarrow}$
🗺 +`CIS-`🗺	$60.10_{0.40\downarrow}$	$58.33_{0.51\uparrow}$	$61.54_{0.40\uparrow}$	$74.08^{***}_{3.18\uparrow}$
Llama3.1-8B-Instruct				
`English`	57.10	44.42	55.91	68.25
+ `CIS-En`	55.90	47.88	55.46	68.45
+ `CIS-Fr`	$52.10^{*}_{3.80\downarrow}$	$52.82^{***}_{4.94\uparrow}$	$59.40^{***}_{3.94\uparrow}$	$68.95_{0.50\uparrow}$
+ `CIS-Ja`	$58.80_{2.90\uparrow}$	$55.96^{***}_{8.08\uparrow}$	$59.86^{***}_{4.40\uparrow}$	$68.90_{0.45\uparrow}$
+ `CIS-Zh`	$55.00_{0.90\downarrow}$	$54.68^{***}_{6.80\uparrow}$	$64.66^{***}_{9.20\uparrow}$	$69.10_{0.65\uparrow}$
+ `CIS-`🗺	$62.50^{***}_{6.60\uparrow}$	$56.03^{***}_{8.15\uparrow}$	$64.74^{***}_{9.28\uparrow}$	$69.60^{***}_{1.15\uparrow}$
🗺 +`CIS-`🗺	$68.60^{***}_{6.10\uparrow}$	$57.24_{1.21\uparrow}$	$66.20^{*}_{1.46\uparrow}$	$71.13^{***}_{1.53\uparrow}$
Qwen2-7B-Instruct				
`English`	43.70	48.46	62.29	53.40
+ `CIS-En`	43.00	50.90	62.31	58.55
+ `CIS-Fr`	$43.50_{0.50\uparrow}$	$53.65^{***}_{2.75\uparrow}$	$62.31_{0.00-}$	$60.90^{***}_{2.35\uparrow}$
+ `CIS-Ja`	$43.80_{0.80\uparrow}$	$56.22^{**}_{5.32\uparrow}$	$63.09_{0.78\uparrow}$	$60.58^{***}_{2.03\uparrow}$
+ `CIS-Zh`	$42.60_{0.40\downarrow}$	$56.79^{***}_{5.89\uparrow}$	$62.49_{0.18\uparrow}$	$58.95_{0.40\uparrow}$
+ `CIS-`🗺	$42.70_{0.30\downarrow}$	$54.94^{***}_{4.04\uparrow}$	$62.51_{0.20\uparrow}$	$60.33^{***}_{1.77\uparrow}$
🗺 +`CIS-`🗺	$47.30^{***}_{4.70\uparrow}$	$55.83_{0.89\uparrow}$	$63.51_{1.00\uparrow}$	$61.93^{***}_{1.60\uparrow}$
Qwen2.5-7B-Instruct				
`English`	59.40	57.82	64.69	68.98
+ `CIS-En`	59.40	59.04	64.20	69.05
+ `CIS-Fr`	$59.40_{0.00-}$	$59.04_{0.00-}$	$64.11_{0.09\downarrow}$	$69.13_{0.08\uparrow}$
+ `CIS-Ja`	$60.10_{0.70\uparrow}$	$59.36_{0.32\uparrow}$	$64.20_{0.00-}$	$68.65_{0.40\downarrow}$
+ `CIS-Zh`	$60.90_{1.50\uparrow}$	$59.23_{0.19\uparrow}$	$65.20^{*}_{1.00\uparrow}$	$68.25^{*}_{0.80\downarrow}$
+ `CIS-`🗺	$59.50_{0.10\uparrow}$	$59.36_{0.32\uparrow}$	$65.06_{0.86\uparrow}$	$68.63_{0.42\downarrow}$
🗺 +`CIS-`🗺	$59.20_{0.30\downarrow}$	$56.79^{*}_{2.57\downarrow}$	$64.14_{0.92\downarrow}$	$69.63^{**}_{1.00\uparrow}$
NeMo-12B-Instruct				
`English`	57.00	51.54	62.26	60.93
+ `CIS-En`	60.70	49.62	61.00	56.48
+ `CIS-Fr`	$61.40_{0.70\uparrow}$	$50.58^{*}_{0.96\uparrow}$	$61.23_{0.23\uparrow}$	$51.15^{***}_{5.33\downarrow}$
+ `CIS-Ja`	$60.20_{0.50\downarrow}$	$49.87_{0.25\uparrow}$	$61.80_{0.80\uparrow}$	$49.90^{***}_{6.58\downarrow}$
+ `CIS-Zh`	$60.50_{0.20\downarrow}$	$50.06_{0.44\uparrow}$	$61.14_{0.14\uparrow}$	$50.05^{***}_{6.43\downarrow}$
+ `CIS-`🗺	$64.90^{***}_{4.20\uparrow}$	$50.19_{0.57\uparrow}$	$62.23^{*}_{1.23\uparrow}$	$51.85^{***}_{4.63\downarrow}$
🗺 +`CIS-`🗺	$62.90_{2.00\downarrow}$	$52.76^{**}_{2.57\uparrow}$	$61.97_{0.26\downarrow}$	$69.38^{***}_{17.53\uparrow}$
Aya-Expanse-8B				
`English`	29.40	58.78	26.54	70.48
+ `CIS-En`	27.80	57.82	23.20	70.30
+ `CIS-Fr`	$28.70_{0.90\uparrow}$	$61.54^{***}_{3.72\uparrow}$	$28.71^{***}_{5.51\uparrow}$	$69.68^{*}_{0.62\downarrow}$
+ `CIS-Ja`	$29.30_{1.50\uparrow}$	$61.03^{**}_{3.21\uparrow}$	$33.06^{***}_{9.86\uparrow}$	$70.00_{0.30\downarrow}$
+ `CIS-Zh`	$28.60_{0.80\uparrow}$	$60.58^{**}_{2.76\uparrow}$	$26.94^{***}_{3.74\uparrow}$	$70.25_{0.05\downarrow}$
+ `CIS-`🗺	$27.90_{0.10\uparrow}$	$62.24^{***}_{4.42\uparrow}$	$33.60^{***}_{10.40\uparrow}$	$70.10_{0.20\downarrow}$
🗺 +`CIS-`🗺	$31.50^{**}_{3.60\uparrow}$	$61.09_{1.15\downarrow}$	$45.91^{***}_{12.31\uparrow}$	$69.70_{0.40\downarrow}$

Table 3: Average accuracies (%) on LRLs after prepending English, monolingual or multilingual CIS to the original `English` mode. Subscripts denote the accuracy delta between the current value and that of `English` + `CIS-En`. Except for `Multilingual` + `CIS-Multilingual` (i.e., 🗺 +`CIS-`🗺 where 🗺 denotes `Multilingual`), subscripts represent the difference of the current value and `English` + `CIS-`🗺. The asterisk superscript indicates the significance level, which we compare in the same way as the accuracy delta . Raw evaluation accuracies for all languages are in Tabs. 19 to 22 in App. B.2. Raw hypothesis test results for CIS mode comparisons are in Tabs. 23 to 26 in App. B.2.

Avg Acc (%)	MGSM		XCOPA	
	LRL	HRL	LRL	HRL
Llama3.1-8B-Instruct				
`Multilingual`	66.00	79.20	66.11	89.64
`Native`	68.50	79.43	71.63	90.80
`Transl-Lang→En`	60.00	74.80	76.74	89.68
`Transl-En→Lang`	68.60	80.63	70.49	90.04
Qwen2.5-7B-Instruct				
`Multilingual`	59.50	86.97	64.63	91.84
`Native`	60.30	87.31	67.69	92.64
`Transl-Lang→En`	63.40	78.63	80.49	91.52
`Transl-En→Lang`	59.90	86.97	66.23	92.48
NeMo-12B-Instruct				
`Multilingual`	63.30	81.09	62.94	88.16
`Native`	65.80	81.26	70.91	90.28
`Transl-Lang→En`	61.60	76.00	77.83	89.32
`Transl-En→Lang`	66.60	82.00	68.46	90.52

Table 4: Average accuracies (%) on MGSM and XCOPA datasets of translation strategies. The comparison includes two translation strategies, the `Native` mode and our proposed `Multilingual` mode, evaluated across low- and high-resource language splits. Raw accuracies of individual languages and more models are recorded in Tabs. 11 and 12 in App. B.1.

4.4 Translation-Based Performance

Mirroring the translation-training (Hu et al., 2020, *inter alia*) setup, existing work suggest that translating from LRLs into English and prompting with the translation results may yield better results (Ahuja et al., 2023, *inter alia*). While translation is not the main focus of our work, we conduct experiment to compare the performance of translation-based strategies for reference.

Strategies. We test two translation strategies for baseline comparison. (1) `Transl-Lang→En`: Translating test questions in other languages in the `English` ICL mode (Fig. 2a) into English. (2) `Transl-En→Lang`: Translating demonstrations in `English` ICL mode (Fig. 2a) into the source language of the current test question, which mirrors the `Native` mode (Fig. 2d). We use the Google Cloud Translation API for translation.[2]

Analysis. For LRLs, `Transl-Lang→En` sometimes outperform the `Native` mode, while `Transl-En→Lang` performs comparably to the `Multilingual` mode, but falls short of the `Native` mode performance (Tab. 4). These results resonate with the phenomena that the translation quality of LRL→En is generally higher than that of the reversed direction (Fan et al., 2021; Goyal et al., 2022; Costa-jussà et al., 2022).

[2]https://cloud.google.com/translate/

For HRLs, however, `Transl-Lang→En` underperforms to `Native` and, sometimes, even `Multilingual` (e.g., on MGSM), suggestting that if a language is sufficiently well-trained, generating responses directly in that language is more effective than translating into English before inference. These two findings highlight that MLLMs approach the ideal of being equally capable in HRLs (Liu et al., 2024), but are still undertrained on LRLs, making translation into English a favored strategy.

In agreement with Poelman and de Lhoneux (2024), we would like to note that even if the task performance of translation-based strategies are the best, the ultimate goal of multilingualism is not just about optimizing task-specific performances. A universal language model should be able to understand and generate text in all languages, instead of relying on specific language(s) as an intermediary. On the other hand, due to the loss of semantic nuances, grammatical structures and cultural context, translation-based strategies may not be the best choice for tasks heavily reliant on language-specific nuances (Liu et al., 2024).

5 Related Work

Prompt engineering. Instruction tuning aligns LLMs more closely with human instructions (Ouyang et al., 2022; Mishra et al., 2022; Wei et al., 2022a; Askell et al., 2021; Wang et al., 2022, 2023). Concurrently, numerous prompting strategies have been developed (Liu et al., 2023) and shown to consistently enhance the performance of instruction-tuned LLMs, such as in-context learning (Brown et al., 2020; Min et al., 2022) and chain-of-thought (Wei et al., 2022b; Kojima et al., 2022). These prompting strategies are proven effective in multilingual tasks as well (Winata et al., 2021; Lin et al., 2022; Shi et al., 2023).

Multilingual ICL. For languages of templates, demonstrations and sample questions in native languages are conventionally inserted into a predefined English template (Lin et al., 2022; Fu et al., 2022; Qin et al., 2023; Huang et al., 2023; Ahuja et al., 2023; Zhang et al., 2024b). Poelman and de Lhoneux (2024) critiques this widespread misuse of English as the *interface* language. Qin et al. (2023); Huang et al. (2023); Zhang et al. (2024b) guide models to “think” and generate CoT in English, regardless of the input language, leading to improved performance for generation tasks compared to “thinking” in other language(s). Sclar et al. (2024); Zhang et al. (2024a) highlight that models are sensitive to those templates. For languages of demonstrations and test questions, Shi et al. (2023); Ahuja et al. (2023) conclude that in-language demonstrations outperform English demonstrations. Etxaniz et al. (2024); Liu et al. (2024) suggest translating questions from LRLs into English can improve performance.

6 Conclusion and Discussion

This work systematically analyzes multiple ICL strategies for MLLMs, and confirms that the presence of multiple languages is an effective strategy across multiple MLLMs. This improvement is partially due to the inclusion of non-English languages in the prompting, and partially due to the in-topic demonstrations in non-English languages, which together strengthen the models' cross-lingual transfer capabilities, particularly the capability to process LRLs. Our work echoes with Turc et al. (2021)—who suggest that HRLs other than English excel in the pretraining-finetuning framework—in the in-context learning framework, highlighting the importance of language inclusivity.

We are in agreement with Costa-jussà et al. (2022) and Liu et al. (2024) that an ideal language-universal LLM should be equally capable in all languages. Beyond this belief, we found that non-English languages may better elicit the potential of MLLMs. Although these observations remain in using HRLs for LRL processing, our results strongly support the call for greater research investment in enhancing MLLM capabilities for a broader range of languages.

Limitations

This paper treats multilingual LLMs as black-box models, drawing the findings and conclusions based solely on their input-output behavior. Hence, we have not interpreted the internal mechanism of how multilingualism could affect MLLM's "thinking" process and its manifested performance. We have briefly touched on the impact of demonstrations in different languages on the MLLM performance. However, we do not conduct a thorough empirical analysis to identify which specific linguistic characteristics (e.g., writing systems, grammatical structures, or linguistic relatedness) contribute to the observed performance differences.

References

Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *CoRR*, abs/1803.08375.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Uttama Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 4232–4267. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Akari Asai, Sneha Kudugunta, Xinyan Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2024. BUFFET: Benchmarking large language models for few-shot cross-lingual transfer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1771–1800, Mexico City, Mexico. Association for Computational Linguistics.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Benjamin Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A general language assistant as a laboratory for alignment. *CoRR*, abs/2112.00861.

Emily Bender. 2019. The #benderrule: On naming the languages we study and why it matters. *The Gradient*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen

Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *CoRR*, abs/2107.03374.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

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

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan N. Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *CoRR*, abs/2412.04261.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Allen L. Edwards. 1948. Note on the "correction for continuity" in testing the significance of the difference between correlated proportions. *Psychometrika*, 13(3):185–187.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in English? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564, Mexico City, Mexico. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48.

The International Organization for Standardization. 2023. Iso 639:2023(en)code for individual languages and language groups.

Jinlan Fu, See-Kiong Ng, and Pengfei Liu. 2022. Polyglot prompt: Multilingual multitask prompt training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9919–9935, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

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 *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, 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. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6919–6971, Mexico City, Mexico. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9019–9052. Association for Computational Linguistics.

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. *CoRR*, abs/2403.10258.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods* *in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 11048–11064. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3470–3487. Association for Computational Linguistics.

MistralAI. 2024. Mistral nemo.

Yongyu Mu, Peinan Feng, Zhiquan Cao, Yuzhang Wu, Bei Li, Chenglong Wang, Tong Xiao, Kai Song, Tongran Liu, Chunliang Zhang, and JingBo Zhu. 2024. Revealing the parallel multilingual learning within large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6976–6997, Miami, Florida, USA. Association for Computational Linguistics.

NLLB-Team. 2024. Scaling neural machine translation to 200 languages. *Nat.*, 630(8018):841–846.

OpenAI. 2022. Introducing chatGPT.

OpenAI. 2024. GPT-4o mini: advancing cost-efficient intelligence.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Wessel Poelman and Miryam de Lhoneux. 2024. The roles of english in evaluating multilingual language models. *CoRR*, abs/2412.08392.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709, Singapore. Association for Computational Linguistics.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Evgeniia Razumovskaia, Ivan Vulic, and Anna Korhonen. 2024. Analyzing and adapting large language models for few-shot multilingual NLU: are we there yet? *CoRR*, abs/2403.01929.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Logical Formalizations of Commonsense Reasoning, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-06, Stanford, California, USA, March 21-23, 2011*. AAAI.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Noam Shazeer. 2020. GLU variants improve transformer. *CoRR*, abs/2002.05202.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Shaomu Tan, Di Wu, and Christof Monz. 2024. Neuron specialization: Leveraging intrinsic task modularity for multilingual machine translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6506–6527, Miami, Florida, USA. Association for Computational Linguistics.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the primacy of english in zero-shot cross-lingual transfer. *CoRR*, abs/2106.16171.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13484–13508. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. 2024. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. *CoRR*, abs/2411.04986.

Haoyun Xu, Runzhe Zhan, Derek F. Wong, and Lidia S. Chao. 2024. Let's focus on neuron: Neuron-level supervised fine-tuning for large language model. *CoRR*, abs/2403.11621.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024a. Qwen2 technical report. *CoRR*, abs/2407.10671.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian

Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024b. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Miaoran Zhang, Vagrant Gautam, Mingyang Wang, Jesujoba Alabi, Xiaoyu Shen, Dietrich Klakow, and Marius Mosbach. 2024a. The impact of demonstrations on multilingual in-context learning: A multidimensional analysis. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7342–7371, Bangkok, Thailand. Association for Computational Linguistics.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, Singapore. Association for Computational Linguistics.

Zhihan Zhang, Dong-Ho Lee, Yuwei Fang, Wenhao Yu, Mengzhao Jia, Meng Jiang, and Francesco Barbieri. 2024b. PLUG: Leveraging pivot language in cross-lingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7025–7046, Bangkok, Thailand. Association for Computational Linguistics.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? In *Advances in Neural Information Processing Systems*, volume 37, pages 15296–15319. Curran Associates, Inc.

A Experiment Setup

A.1 Model

All model checkpoints we use and their properties are listed in Tab. 5.

A.2 Dataset

After preprocessing, datapoints for each language split are stored in a single JSON file. Tab. 6 summarizes the supported languages for each dataset and the sources from which they are obtained.

MGSM The original dataset consists of parallel datapoints across all language splits and training/test splits of the same size. Example datapoint and the Chat Template for few-shot demonstrations can be found in Fig. 6.

XCOPA The 100 datapoints in the training split of XCOPA are parallel to the last 100 datapoints in the English COPA development split. Therefore, we exclude the first 400 datapoints from the development split of English COPA. The test splits of both datasets are parallel and contain the same number of datapoints. We then merge both into our XCOPA dataset. In XCOPA, there are two types of questions: "cause" and "effect", each corresponding to a distinct template, as shown in Figs. 7a and 7b.

```
{
  "question": "Question: Leah had 32 chocolatesand her sister had 42. If they ate 35, how many pieces do they have left in total?",
  "answer": "Step-by-Step Answer: Leah had 32 chocolates and Leah's sister had 42. That means there were originally 32 + 42 = 74 chocolates. 35 have been eaten. So in total they still have 74 - 35 = 39 chocolates. The answer is 39.",
  "answer_number": 39,
  "language": "en",
},
```

ICL Q	Question: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?
ICL A	Step-by-Step Answer: Leah had 32 chocolates and Leah's sister had 42. That means there were originally 32 + 42 = 74 chocolates. 35 have been eaten. So in total they still have 74 - 35 = 39 chocolates. The answer is 39.

Figure 6: An example of an English datapoint from MGSM training set. When calling Chat Template API, `user` role message is the "question" value, while `assistant` role message is the "answer" value. Note that in the test set, the answer is null without exemplar CoT response. The correct numerical answer is stored in "answer_number".

```
{
  "premise": "My eyes became red and puffy.",
  "choice1": "I was sobbing.",
  "choice2": "I was laughing.",
  "question": "cause",
  "label": 1,
  "language": "en"
},
```

ICL Q	Premise: My eyes became red and puffy. What was the CAUSE of this? Hypothesis 1: I was sobbing. Hypothesis 2: I was laughing.
ICL A	1.

(a) An example of "cause" datapoint.

```
{
  "premise": "The man turned on the faucet.",
  "choice1": "The toilet filled with water.",
  "choice2": "Water flowed from the spout.",
  "question": "effect",
  "label": 2,
  "language": "en",
}
```

ICL Q	Premise: The man turned on the faucet. What happened as a RESULT? Hypothesis 1: The toilet filled with water. Hypothesis 2: Water flowed from the spout.
ICL A	2.

(b) An example of "effect" datapoint.

Figure 7: Examples of English datapoints from XCOPA training set. First, based on whether "question" is "cause" or "effect", we fill the "premise", "choice1", and "choice2" values into one of the two predefined templates. The template's language is changeable as per the language split of the datapoint. Then we call the Chat Template API, `user` role message is the filled template, while `assistant` role message is the "label" value.

XL-WiC This benchmark is designed to determine whether a specific word in a given language has the same meaning in two different sentences. As a result, the dataset is inherently non-parallel. Among all language splits, Estonian (et) contains the fewest datapoints, with 98 in the training split and 390 in the test split. For all other languages, we randomly subsample to match the size of the Estonian split to satisfy the demonstration sampling requirements outlined in §2.2. To leverage the attention mechanism of transformers (Vaswani et al., 2017), we add asterisks around the target word in both sentences to indicate that the LLM needs to disambiguate the meaning of that specific word. An example datapoint is shown in Fig. 8.

COMBINED Identifying specific neurons depends on the nature of the input corpus. Since language is inherently conjugate with the task, and

Model Name	Scale	Instruct?	Open-Source?	Checkpoint
Llama3-8B-Instruct	8B	✔	✔	meta-llama/Meta-Llama-3-8B-Instruct
Llama3.1-8B-Instruct	8B	✔	✔	meta-llama/Meta-Llama-3.1-8B-Instruct
Qwen2-7B-Instruct	7B	✔	✔	Qwen/Qwen2-7B-Instruct
Qwen2.5-7B-Instruct	7B	✔	✔	Qwen/Qwen2.5-7B-Instruct
NeMo-12B-Instruct	12B	✔	✔	mistralai/Mistral-Nemo-Instruct-2407
Aya-Expanse-8B	8B	✔	✔	CohereForAI/aya-expanse-8b
GPT3.5-turbo	NA	✔	✖	gpt-3.5-turbo-0125
GPT4o-mini	NA	✔	✖	gpt-4o-mini-2024-07-18

Table 5: Model details. Checkpoints are either from Hugging Face or OpenAI API.

Dataset	HRL	LRL	Source
MGSM	de, en, es, fr, ja, ru, zh	bn, sw, te, th	juletxara/mgsm
XCOPA	en, id, it, tr, zh	et, ht, qu, sw, ta, th, vi	English COPA & cambridgeltl/xcopa
XL-WiC	da, de, en, fr, it, ja, ko, nl, zh	bg, et, fa, hr	pilehvar.github.io/xlwic/
XQuAD	ar, de, en, es, ru, tr, vi, zh	el, hi, ro, th	google/xquad

Table 6: Additional information for the three datasets we evaluate. The correspondence between language codes and names can be found in Tab. 9. "Source" indicates where to download the dataset.


```
{
  "target_word": "get",
  "example_1": "What did you *get* at the toy store?",
  "example_2": "She didn't *get* his name when they met the first time.",
  "label": 0,
  "language": "en",
},
```


ICL Q Sentence 1: What did you *get* at the toy store?
Sentence 2: She didn't *get* his name when they met the first time.
Question: Is the word "get" (marked with *) used in the same way in both sentences above?

ICL A No.

Figure 8: An example of English datapoint from XL-WiC training set. We first fill the "example_1", "example_2", and "target_word" values into the predefined templates. Asterisks * are surrounded around "target_word" to draw the LLM's attention. The template's language is changeable as per the language split of the datapoint. Then we call the Chat Template API, `user` role message is the filled template, while `assistant` role message is "Yes" or "No" ("label" is 1 or 0).

our focus is on language-specific neurons rather than task-specific neurons, it is necessary to input all three datasets into the LLM to eliminate the confounding factor of the task domain. To balance the three datasets, we subsample their test splits to only 250 datapoints each. To balance the number of language splits across datasets, we excluded two HRL splits, Korean (ko) and Dutch (nl), from the XL-WiC dataset. This ensures that all three (sub-)datasets have 11 languages for combination. Additionally, for MGSM, we limit the answers to only include the final numeric result without the CoT reasoning. This approach ensures that the total number of datapoints and the overall token count are roughly the same across the original three datasets.

XQuAD The dataset consists of parallel question-answering examples across 12 languages, derived from the same subset of the English SQuAD (Rajpurkar et al., 2016) v1.1 development set. For each language, the context passages, questions, and answers are professionally translated while preserving semantic alignment across languages. Each example includes a context paragraph, a question, and a span-based answer. An example and the input prompt format used in our experiments are illustrated in Fig. 9.

FLORES-101 This machine translation benchmark comprises 3,001 sentences extracted from English Wikipedia, spanning diverse topics and domains. These sentences were translated into 101 languages by professional translators via a carefully controlled process. The Wikipedia domain is largely unrelated to the domains of the three datasets we evaluate (math, commonsense reasoning, and word disambiguation). Therefore, we choose FLORES-101 as our source pool of irrelevant sentences. Note that in this benchmark, each datapoint carries the same semantic meaning across all 101 language splits. We do not want irrelevant sentences to affect the LLM's understanding of the original task excessively, thus we select sentences with word counts between 10 and 15 in the English split, introducing limited noise. CIS are drawn form the filtered FLORES-101 dataset. An

```
{
  "context": "The Panthers defense gave up just 308
points, ranking sixth in the league, while also leading
the NFL in interceptions with 24 and boasting four
Pro Bowl selections. Pro Bowl defensive tackle
Kawann Short led the team in sacks with 11, while
also forcing three fumbles and recovering two.
Fellow lineman Mario Addison added 6½ sacks. The
Panthers line also featured veteran defensive end
Jared Allen, a 5-time pro bowler who was the NFL's
active career sack leader with 136, along with
defensive end Kony Ealy, who had 5 sacks in just 9
starts. Behind them, two of the Panthers three
starting linebackers were also selected to play in the
Pro Bowl: Thomas Davis and Luke Kuechly. Davis
compiled 5½ sacks, four forced fumbles, and four
interceptions, while Kuechly led the team in tackles
(118) forced two fumbles, and intercepted four
passes of his own. Carolina's secondary featured Pro
Bowl safety Kurt Coleman, who led the team with a
career high seven interceptions, while also racking
up 88 tackles and Pro Bowl cornerback Josh
Norman, who developed into a shutdown corner
during the season and had four interceptions, two of
which were returned for touchdowns.",
  "question": "How many balls did Josh Norman
intercept?",
  "answers": {
    "text": [
      "four"
    ],
    "answer_start": [
      140
    ]
  },
  "language": "en"
},
```

ICL Q Passage: The Panthers defense gave up …
Question: How many balls did Josh Norman intercept?

ICL A 140.

Figure 9: An example of English datapoint from XQuAD validation set. We fill the "context" and "question" values into the predefined templates. The template's language is changeable as per the language split of the datapoint. Then we call the Chat Template API, `user` role message is the filled template, while `assistant` role message is "answers['text']".

example is provided in Fig. 10.

A.3 Language

A.3.1 High-Resource Language List

To the best of our knowledge, we find two multilingual LLMs—Llama2 (Touvron et al., 2023) and PaLM (Chowdhery et al., 2023)—that publicly report the language distribution used during pretraining. The top 20 languages and their percentages for each model are listed in Tab. 7 and Tab. 8, respectively. We take the union of these languages to form what we consider a *high-resource language*

Code	Language	Percentage
en	English	89.70%
de	German	0.17%
fr	French	0.16%
sv	Swedish	0.15%
zh	Chinese	0.13%
es	Spanish	0.13%
ru	Russian	0.13%
nl	Dutch	0.12%
it	Italian	0.11%
ja	Japanese	0.10%
pl	Polish	0.09%
pt	Portuguese	0.09%
vi	Vietnamese	0.08%
uk	Ukrainian	0.07%
ko	Korean	0.06%
ca	Catalan	0.04%
sr	Serbian	0.04%
id	Indonesian	0.03%
cs	Czech	0.03%
fi	Finnish	0.03%

Table 7: Top 20 language distribution of the training data for Llama2, excluding code and unknown data. Adopted from Table 10 in Touvron et al. (2023).

Code	Language	Tokens (B)	Percentage
en	English	578.064	77.984%
de	German	25.954	3.501%
fr	French	24.094	3.250%
es	Spanish	15.654	2.112%
pl	Polish	10.764	1.452%
it	Italian	9.699	1.308%
nl	Dutch	7.690	1.037%
sv	Swedish	5.218	0.704%
tr	Turkish	4.855	0.655%
pt	Portuguese	4.701	0.634%
ru	Russian	3.932	0.530%
fi	Finnish	3.101	0.418%
cs	Czech	2.991	0.404%
zh	Chinese	2.977	0.402%
ja	Japanese	2.832	0.382%
no	Norwegian	2.695	0.364%
ko	Korean	1.444	0.195%
da	Danish	1.387	0.187%
id	Indonesian	1.175	0.159%
ar	Arabic	1.091	0.147%

Table 8: Top 20 language distribution of the training data for PaLM, excluding code and unknown data. Adopted from Table 28 in Chowdhery et al. (2023).

```
{
"da": "Vidal har, siden han flyttede til den catalanske hovedstad, spillet 49 kampe for klubben.",
"de": "Seit seinem Umzug in die katalanische Hauptstadt hatte Vidal für den Verein 49 Partien gespielt.",
"en": "Since moving to the Catalan-capital, Vidal had played 49 games for the club.",
"es": "Desde su mudanza a la capital catalana, Vidal jugó 49 partidos para el club.",
"fr": "Depuis son arrivée dans la capitale catalane, Vidal a joué 49 matchs pour le club.",
"id": "Sejak menetap di ibu kota Catalan, Vidal sudah bermain di 49 laga untuk klub ini.",
"it": "Dal suo arrivo nella capitale catalana, Vidal ha giocato per il club 49 partite.",
"ja": "カタルーニャの州都に移って以来、ビダルはクラブで49試合に出場しました。",
"ko": "바르셀로나로 이적한 후 비달은 클럽을 위해 49경기를 뛰었습니다.",
"nl": "Sinds hij verhuisde naar de Catalaanse hoofdstad heeft hij 49 wedstrijden gespeeld voor de club.",
"ru": "После переезда в столицу Каталонии  Видаль провел за клуб 49 матчей.",
"tr": "Katalan başkentine taşındığından beri Vidal kulüp adına 49 maça çıktı.",
"zh": "自从转会到加泰罗尼亚的首府球队，维达尔已经为俱乐部踢了 49 场比赛。",
}
```


Figure 10: A datapoint example from FLORES-101 of semantic-equivalent context-irrelevant sentences in all high resource languages we study in this work.

list (in terms of the richness in the LLM pretraining dataset), including the following 24 languages:

$$\mathcal{L}_H = \{\text{ar, ca, cs, da, de, en, es, fi, fr, id, it, ja, ko, nl, no, pl, pt, ru, sr, sv, tr, uk, vi, zh}\}. \quad (3)$$

The high overlap between the top languages of the two LLMs further supports the rationale for applying this list to other LLMs.

A.3.2 Languages We Evaluate

Tab. 9 presents the union of languages supported by the datasets we evaluated (§ 3). Based on whether a language appears in the high-resource language list, we categorized the 29 languages into HRL and LRL groups, with 14 classified as HRL and 15 as LRL. Among them, only English and Chinese are present in all three datasets.

It is worth highlighting that the 11 LRLs span 7 distinct writing systems and 6 language families. This diversity suggests that when tokenizing inputs in these LRLs, they are unlikely to share common tokens with HRLs (9 out of 13 use the Latin script). Consequently, this limits the MLLMs' ability to leverage shared subwords or similar syntax structures for cross-lingual transfer across LRLs.

B Experiment Raw Results

B.1 ICL Modes Evaluation

The raw data for vanilla evaluation is recorded in Tabs. 11 to 14. McNemar's test results for ICL modes are in Tabs. 15 to 18. Significance (sig.) levels – *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

B.2 Context-Irrelevant Sentence

The raw data for CIS is recorded in Tabs. 19 to 22. McNemar's test results for CIS modes are in Tabs. 23 to 26.

C ICL Behavioral Analysis

C.1 Specilized Neuron

Inspired by the universal concept space (Wendler et al., 2024; Zhao et al., 2024; Wu et al., 2024), we hypothesize that MLLMs could activate more cross-lingual capabilities by aligning different linguistic representations. To validate the above hypothesis, we seek to find patterns in neuron behavior between ICL modes. Following Tan et al. (2024); Tang et al. (2024); Kojima et al. (2024); Mu et al. (2024); Zhao et al. (2024); Xu et al. (2024), we look at the activations of neurons in the multilayer perceptron (MLP, or *Feedforward Network, FFN*) modules of the MLLMs.

C.1.1 Identifying Top-Activated Neurons

Each neuron in every MLP layer of the model is assigned a dedicated counter, initialized to 0. During vanilla evaluation, we monitor the activation of every neuron in the forward pass. Since LLMs typically employ ReLU-like (Agarap, 2018) activation functions (e.g., SwiGLU (Shazeer, 2020) for Llama series), a positive activation value can be interpreted as the neuron being "activated". If a neuron is "activated", we increment the corresponding counter by 1, otherwise no action. To ensure balanced input across our three datasets, we curated a COMBINED dataset, see App. A.2 for details. After processing the inputs of a single ICL mode, each neuron accumulates an "activated" count. The neurons with the highest counts are identified as specialized neurons attributed to this ICL mode.

We employ two methods for selecting the most activated neurons. Top-k selects neurons in the top k percentile (Tang et al., 2024), while top-p selects neurons progressively until their cumulative activation counts reach p (%) of the sum of all activation values (Tan et al., 2024).

Code	Name in English	HRL/LRL	In Which Dataset(s)	Writing System	Family
ar	Arabic	H	XQuAD	Arabic	Afro-Asiatic
bg	Bulgarian	L	XL-WiC	Cyrillic	Indo-European
bn	Bengali	L	MGSM	Bengali–Assamese	Indo-European
da	Danish	H	XL-WiC	Latin	Indo-European
de	German	H	MGSM, XL-WiC, XQuAD	Latin	Indo-European
el	Greek	L	XQuAD	Greek	Indo-European
en	English	H	MGSM, XCOPA, XL-WiC XQuAD	Latin	Indo-European
es	Spanish	H	MGSM, XQuAD	Latin	Indo-European
et	Estonian	L	XCOPA	Latin	Indo-European
fa	Persian	L	XL-WiC	Arabic	Indo-European
fr	French	H	MGSM, XL-WiC	Latin	Indo-European
hi	Hindi	L	XQuAD	Devanagari	Indo-European
hr	Croatian	L	XL-WiC	Latin	Indo-European
ht	Haitian	L	XCOPA	Latin	French Creole
id	Indonesian	H	XCOPA	Latin	Austronesian
it	Italian	H	XL-WiC, XCOPA	Latin	Indo-European
ja	Japanese	H	MGSM, XL-WiC	Kana & Hanzi	Japonic
ko	Korean	H	XL-WiC	Hangul	Koreanic
nl	Dutch	H	XL-WiC	Latin	Indo-European
qu	Quechua	L	XCOPA	Latin	Quechumaran
ro	Romanian	L	XQuAD	Latin	Indo-European
ru	Russian	H	MGSM, XQuAD	Cyrillic	Indo-European
sw	Swahili	L	MGSM	Latin	Niger–Congo
ta	Tamil	L	XCOPA	Tamil	Dravidian
te	Telegu	L	MGSM	Telegu	Dravidian
th	Thai	L	MGSM, XCOPA, XQuAD	Thai	Kra–Dai
tr	Turkish	H	XCOPA, XQuAD	Latin	Turkic
vi	Vietnamese	L	XCOPA, XQuAD	Latin	Austroasiatic
zh	Chinese	H	MGSM, XCOPA, XL-WiC XQuAD	Hanzi	Sino-Tibetan

Table 9: List of 29 languages we study and their properties, in ascending order of ISO 639-1 codes (for Standardization, 2023).

C.2 `Multilingual`-specific Neurons Overlap Most with `Native`-specific Neurons

IoU (%)	All Langs	LRL	HRL
`English– Multilingual– Native`			
Llama3.1-8B-Instruct			
`Native- English`	61.82	56.81	68.50
`Native- Multilingual`	78.84	66.19	85.10
`English- Multilingual`	65.84	66.89	64.60
Qwen2-7B-Instruct			
`Native- English`	70.61	66.05	81.22
`Native- Multilingual`	85.13	77.40	91.31
`English- Multilingual`	78.24	79.83	77.46
`English– Chinese– Native`			
Llama3.1-8B-Instruct			
`Native- English`	61.82	56.81	68.50
`Native- Chinese`	60.58	55.35	65.73
`English- Chinese`	56.52	57.94	55.72
Qwen2-7B-Instruct			
`Native- English`	70.61	66.05	81.22
`Native- Chinese`	73.41	69.39	82.18
`English- Chinese`	77.07	78.34	76.51

Table 10: The IoU score of most-activated neurons between every pair of ICL modes in triplets `English– Multilingual– Native` and `English– Chinese– Native`. Neurons were selected by first filtering out neurons outside the first 8 and last 8 MLP layers, and applying top-k method with $k = 0.7$.

We examined the overlaps among the most-activated neurons (*specialized neurons*) across different ICL modes by calculating the Intersection over Union (IoU) scores. For ICL modes M_1 and M_2, with specialized neurons denoted as sets S_1 and S_2, their overlap is quantified by Eq. (4):

$$\text{IoU}\,(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}. \tag{4}$$

Prior research (Tang et al., 2024; Zhao et al., 2024) has discovered that *language-specific neurons* are located primarily in the models' top and bottom layers. Because we want to explain the multilingual reasoning capabilities of a model, we only consider neurons that belong to a certain prefix or suffix of the models' layers in an effort to restrict our selected neuron sets to be mainly language-specific neurons.

The similarity in performance between `Multilingual` and `Native` can be explained by the high overlap in the sets of most-activated neurons. On the other hand, the poorer performance of `English` can be explained by the low overlap between `English` most-activated neurons and other ICL modes .

The same experiment was also performed but with `Multilingual` replaced by HRL `Chinese`. In this case, the patterns were less obvious and model-specific. This result aligns with our findings in vanilla evaluation that `English` $\leq$ Non-`English` HRL `Monolingual` $\leq$ `Multilingual` (§4.2).

We further split this neuron experiment to either use only LRL or HRL ICL modes when recording neuron activations. We observed that HRL tends to activate similar neurons between `Native` and `Multilingual`, whereas LRL tends to activate similar neurons between `English` and `Multilingual`.

MGSM	bn	de	en	es	fr	ja	ru	sw	te	th	zh	LRL Avg	HRL Avg	ALL Avg
Llama3-8B-Instruct														
English	66.40	76.40	86.40	78.80	77.60	66.40	77.20	59.20	60.00	71.20	75.60	64.20	76.91	72.29
French	67.60	77.60	86.00	77.60	75.20	70.00	79.60	56.40	59.60	68.00	70.40	$62.90_{1.30\downarrow}$	$76.63_{0.28\downarrow}$	$71.64_{0.65\downarrow}$
Chinese	66.80	77.20	82.80	79.60	74.80	68.00	76.40	55.60	59.20	70.40	72.80	$63.00_{1.20\downarrow}$	$75.94_{0.97\downarrow}$	$71.24_{1.05\downarrow}$
Japanese	66.80	78.80	83.20	76.00	74.80	70.80	76.40	56.40	56.80	68.80	68.00	$62.20_{2.00\downarrow}$	$75.43_{1.48\downarrow}$	$70.62_{1.67\downarrow}$
Multilingual	65.20	79.60	84.40	77.60	74.00	69.60	76.80	52.40	53.60	69.60	70.00	$60.20_{4.00\downarrow}$	$76.00_{0.91\downarrow}$	$70.25_{2.04\downarrow}$
Native	64.80	76.80	86.40	80.00	75.20	70.80	76.80	58.40	54.80	70.40	72.80	$62.10_{2.10\downarrow}$	$76.97_{0.06\uparrow}$	$71.56_{0.73\downarrow}$
Transl-Lang→En	63.60	73.20	86.40	78.00	74.00	60.40	77.60	69.60	60.00	46.40	58.80	$59.90_{4.30\downarrow}$	$72.63_{4.28\downarrow}$	$68.00_{4.29\downarrow}$
Transl-En→Lang	65.20	78.80	86.40	79.60	76.00	68.40	76.40	58.80	52.80	69.20	73.60	$61.50_{2.70\downarrow}$	$77.03_{0.12\uparrow}$	$71.38_{0.91\downarrow}$
Llama3.1-8B-Instruct														
English	52.40	69.20	87.20	63.60	76.80	68.00	78.80	67.20	39.60	69.20	75.20	57.10	74.11	67.93
French	62.80	74.40	89.20	82.40	82.40	69.60	79.20	69.60	45.20	70.00	76.80	$61.90_{4.80\uparrow}$	$79.14_{5.03\uparrow}$	$72.87_{4.94\uparrow}$
Chinese	66.00	76.80	87.20	83.20	78.00	67.60	79.60	70.00	48.40	69.20	77.60	$63.40_{6.30\uparrow}$	$78.57_{4.46\uparrow}$	$73.05_{5.12\uparrow}$
Japanese	66.40	76.80	86.00	80.80	78.00	67.20	79.60	72.80	58.00	75.20	77.60	$68.10_{11.00\uparrow}$	$78.00_{3.89\uparrow}$	$74.40_{6.47\uparrow}$
Multilingual	68.00	76.40	88.00	84.80	80.40	69.20	80.40	68.80	55.60	71.60	75.20	$66.00_{8.90\uparrow}$	$79.20_{5.09\uparrow}$	$74.40_{6.47\uparrow}$
Native	67.20	80.40	87.20	81.20	82.40	67.20	80.00	72.40	58.00	76.40	77.60	$68.50_{11.40\uparrow}$	$79.43_{5.32\uparrow}$	$75.45_{7.52\uparrow}$
Transl-Lang→En	62.80	74.80	87.20	82.00	77.60	62.80	81.60	68.00	62.00	47.20	57.60	$60.00_{2.90\uparrow}$	$74.80_{0.69\uparrow}$	$69.42_{1.49\uparrow}$
Transl-En→Lang	67.60	80.00	87.20	82.80	83.60	70.80	80.80	70.80	59.60	76.40	79.20	$68.60_{11.50\uparrow}$	$80.63_{6.52\uparrow}$	$76.25_{8.32\uparrow}$
Qwen2-7B-Instruct														
English	57.20	74.00	90.40	82.00	80.00	67.60	80.40	20.80	23.20	73.60	79.20	43.70	79.09	66.22
French	54.80	74.00	92.00	82.00	79.60	66.80	79.20	22.40	20.80	73.60	80.40	$42.90_{0.80\downarrow}$	$79.14_{0.05\uparrow}$	$65.96_{0.26\downarrow}$
Chinese	56.00	76.40	92.00	81.20	77.60	71.20	79.60	28.80	20.40	73.60	85.20	$44.70_{1.00\uparrow}$	$80.46_{1.37\uparrow}$	$67.45_{1.23\uparrow}$
Japanese	51.20	72.80	88.80	78.80	76.00	75.20	79.20	27.20	20.80	72.40	80.80	$42.90_{0.80\downarrow}$	$78.80_{0.29\downarrow}$	$65.75_{0.47\downarrow}$
Multilingual	64.80	77.20	89.20	82.80	81.60	70.40	82.00	28.40	23.60	73.20	79.60	$47.50_{3.80\uparrow}$	$80.40_{1.31\uparrow}$	$68.44_{2.22\uparrow}$
Native	72.00	83.60	90.40	82.80	79.60	75.20	83.20	32.40	40.40	76.80	85.20	$55.40_{11.70\uparrow}$	$82.86_{3.77\uparrow}$	$72.87_{6.65\uparrow}$
Transl-Lang→En	66.00	76.00	90.40	79.60	79.20	64.00	77.60	72.40	61.20	47.60	58.40	$61.80_{18.10\uparrow}$	$75.03_{4.06\downarrow}$	$70.22_{4.00\uparrow}$
Transl-En→Lang	72.40	81.20	90.40	82.40	80.00	69.20	81.20	32.80	43.20	74.80	80.80	$55.80_{12.10\uparrow}$	$80.74_{1.65\uparrow}$	$71.67_{5.45\uparrow}$
Qwen2.5-7B-Instruct														
English	75.20	82.40	94.40	88.80	87.60	76.80	88.00	30.80	48.80	82.80	86.80	59.40	86.40	76.58
French	74.00	85.20	92.40	90.00	85.60	78.80	86.80	33.60	51.20	82.40	84.00	$60.30_{0.90\uparrow}$	$86.11_{0.29\downarrow}$	$76.73_{0.15\uparrow}$
Chinese	77.20	85.20	94.00	88.80	86.00	81.60	88.00	31.20	50.40	81.20	86.40	$60.00_{0.60\uparrow}$	$87.14_{0.74\uparrow}$	$77.27_{0.69\uparrow}$
Japanese	74.80	84.80	94.80	90.80	88.00	81.20	88.40	32.80	50.80	83.20	85.60	$60.40_{1.00\uparrow}$	$87.66_{1.26\uparrow}$	$77.75_{1.17\uparrow}$
Multilingual	76.40	86.80	92.80	89.60	88.00	78.40	87.20	30.40	50.00	81.20	86.00	$59.50_{0.10\uparrow}$	$86.97_{0.57\uparrow}$	$76.98_{0.40\uparrow}$
Native	75.20	86.80	94.40	89.20	85.60	81.20	87.60	30.40	46.00	84.40	86.40	$60.30_{0.90\uparrow}$	$87.31_{0.91\uparrow}$	$77.49_{0.91\uparrow}$
Transl-Lang→En	70.40	79.60	94.40	83.60	82.40	64.00	81.60	72.80	62.40	48.00	64.80	$63.40_{4.60\uparrow}$	$78.63_{7.77\downarrow}$	$73.09_{3.49\downarrow}$
Transl-En→Lang	74.80	85.60	94.40	88.80	84.40	80.40	87.20	35.20	46.80	82.80	88.00	$59.90_{0.50\uparrow}$	$86.97_{0.57\uparrow}$	$77.13_{0.55\uparrow}$
NeMo-12B-Instruct														
English	66.80	75.60	90.80	81.20	80.00	64.40	83.60	42.00	54.00	65.20	76.80	57.00	78.91	70.95
French	66.00	81.20	92.40	82.00	81.60	74.80	83.20	46.40	70.40	74.40	76.40	$64.30_{7.30\uparrow}$	$81.66_{2.75\uparrow}$	$75.35_{4.40\uparrow}$
Chinese	70.40	81.60	91.60	81.60	82.00	73.20	84.80	49.20	69.60	69.20	77.20	$64.60_{7.60\uparrow}$	$81.71_{2.80\uparrow}$	$75.49_{4.54\uparrow}$
Japanese	67.20	84.40	90.00	85.60	83.20	74.40	84.80	45.60	69.60	72.00	76.00	$63.60_{6.60\uparrow}$	$82.63_{3.72\uparrow}$	$75.71_{4.76\uparrow}$
Multilingual	67.20	82.00	90.40	82.80	77.60	71.60	83.20	50.80	66.40	68.80	80.00	$63.30_{6.30\uparrow}$	$81.09_{2.18\uparrow}$	$74.62_{3.67\uparrow}$
Native	66.40	81.20	90.80	82.00	81.60	74.40	81.60	58.00	67.60	71.20	77.20	$65.80_{8.80\uparrow}$	$81.26_{2.35\uparrow}$	$75.64_{4.69\uparrow}$
Transl-Lang→En	65.20	76.80	90.80	82.00	78.80	62.00	80.00	71.60	62.40	47.20	61.60	$61.60_{4.60\uparrow}$	$76.00_{2.91\downarrow}$	$70.76_{0.19\downarrow}$
Transl-En→Lang	68.80	83.60	90.80	80.40	82.00	74.00	85.60	57.60	66.80	73.20	77.60	$66.60_{9.60\uparrow}$	$82.00_{3.09\uparrow}$	$76.40_{5.45\uparrow}$
Aya-Expanse-8B														
English	39.60	75.60	82.40	81.60	72.40	67.20	77.20	20.00	17.20	40.80	71.20	29.40	75.37	58.65
French	43.60	74.40	81.60	80.00	75.20	67.60	77.60	19.60	23.20	36.40	74.00	$30.70_{1.30\uparrow}$	$75.77_{0.40\uparrow}$	$59.38_{0.73\uparrow}$
Chinese	44.00	74.40	80.40	77.60	73.20	66.00	77.20	20.00	22.40	40.40	76.40	$31.70_{2.30\uparrow}$	$75.03_{0.34\downarrow}$	$59.27_{0.62\uparrow}$
Japanese	42.80	72.40	83.20	79.20	72.80	66.80	77.20	20.40	20.00	38.40	72.00	$30.40_{1.00\uparrow}$	$74.80_{0.57\downarrow}$	$58.65_{0.00-}$
Multilingual	46.80	74.00	83.60	78.00	72.80	67.60	79.60	19.20	17.60	36.80	72.00	$30.10_{0.70\uparrow}$	$75.37_{0.00-}$	$58.91_{0.26\uparrow}$
Native	44.80	77.60	82.40	79.60	75.20	66.80	79.20	19.60	22.00	44.00	76.40	$32.60_{3.20\uparrow}$	$76.74_{1.37\uparrow}$	$60.69_{2.04\uparrow}$
Transl-Lang→En	64.40	71.20	82.40	76.40	74.40	57.60	69.20	65.20	62.80	48.00	59.60	$60.10_{30.70\uparrow}$	$70.11_{5.26\downarrow}$	$66.47_{7.82\uparrow}$
Transl-En→Lang	46.40	74.80	82.40	80.00	74.80	66.00	78.80	19.20	20.80	42.40	76.00	$32.20_{2.80\uparrow}$	$76.11_{0.74\uparrow}$	$60.15_{1.50\uparrow}$
GPT3.5-turbo														
English	39.60	75.20	86.00	76.80	62.80	59.20	66.40	63.60	12.80	60.80	67.20	44.20	70.51	60.95
French	34.00	74.80	86.00	84.80	79.20	66.00	67.60	63.20	15.20	55.60	74.40	$42.00_{2.20\downarrow}$	$76.11_{5.60\uparrow}$	$63.71_{2.76\uparrow}$
Chinese	30.40	78.80	83.20	78.40	77.20	69.20	77.60	67.20	15.60	62.40	73.20	$43.90_{0.30\downarrow}$	$76.80_{6.29\uparrow}$	$64.84_{3.89\uparrow}$
Japanese	27.60	78.80	85.20	82.00	72.40	73.20	74.40	69.60	14.00	61.20	76.00	$43.10_{1.10\downarrow}$	$77.43_{6.92\uparrow}$	$64.95_{4.00\uparrow}$
Multilingual	54.40	79.60	83.60	79.20	73.60	68.00	75.20	68.00	24.40	57.20	74.40	$51.00_{6.80\uparrow}$	$76.23_{5.72\uparrow}$	$67.05_{6.10\uparrow}$
Native	57.20	79.60	86.00	80.40	81.60	75.20	77.60	73.60	30.00	58.80	74.00	$54.90_{10.70\uparrow}$	$79.26_{8.75\uparrow}$	$70.40_{9.45\uparrow}$
GPT4o-mini														
English	87.20	90.80	94.80	92.00	87.20	84.80	92.00	83.20	84.00	88.80	90.00	85.80	90.23	88.62
French	87.20	90.80	94.40	92.80	89.20	82.40	90.80	85.20	84.00	88.80	87.60	$86.30_{0.50\uparrow}$	$89.71_{0.52\downarrow}$	$88.47_{0.15\downarrow}$
Chinese	86.80	90.80	93.60	93.20	90.00	83.20	91.20	85.20	84.40	90.00	90.80	$86.60_{0.80\uparrow}$	$90.40_{0.17\uparrow}$	$89.02_{0.40\uparrow}$
Japanese	85.20	89.60	94.80	92.80	86.80	86.00	92.00	82.80	81.60	88.40	87.60	$84.50_{1.30\downarrow}$	$89.94_{0.29\downarrow}$	$87.96_{0.66\downarrow}$
Multilingual	86.40	90.00	92.80	94.00	89.20	84.00	92.40	84.00	80.40	88.00	89.20	$84.70_{1.10\downarrow}$	$90.23_{0.00-}$	$88.22_{0.40\downarrow}$
Native	85.20	88.00	94.80	91.20	89.20	85.20	90.40	85.60	81.60	88.40	90.00	$85.20_{0.60\downarrow}$	$89.66_{0.57\downarrow}$	$88.04_{0.58\downarrow}$

Table 11: Accuracies (%) of English, Multilingual, Native, all Monolingual ICL modes (French, Chinese and Japanese) and two translation strategies (Transl-Lang→En, Transl-En→Lang) across 11 languages of the MGSM dataset. AVG represents the average accuracy of the language set (LRLs, HRLs or All languages). The underlined languages in the table header are LRLs, otherwise HRLs. The subscript indicates the performance increase↑ (or decrease↓) of all other modes compared to the English ICL mode.

XCOPA	en	<u>et</u>	<u>ht</u>	id	it	<u>qu</u>	<u>sw</u>	<u>ta</u>	<u>th</u>	tr	<u>vi</u>	zh	<u>LRL AVG</u>	HRL AVG	ALL AVG
Llama3-8B-Instruct															
English	95.20	55.80	10.20	79.60	86.60	7.60	40.00	59.40	69.60	72.00	81.00	87.00	46.23	84.08	62.00
Italian	93.80	59.80	46.80	81.00	89.60	41.20	57.40	58.80	73.40	76.00	80.20	88.20	59.66$_{13.43\uparrow}$	85.72$_{1.64\uparrow}$	70.52$_{8.52\uparrow}$
Chinese	93.80	59.80	52.80	82.40	85.60	47.60	55.40	59.00	79.00	74.60	79.60	90.80	61.89$_{15.66\uparrow}$	85.44$_{1.36\uparrow}$	71.70$_{9.70\uparrow}$
Multilingual	94.60	57.80	51.80	81.60	87.80	46.40	60.40	60.80	79.20	78.80	80.40	88.80	62.40$_{16.17\uparrow}$	86.32$_{2.24\uparrow}$	72.37$_{10.37\uparrow}$
Native	95.20	68.60	61.60	85.00	89.60	50.40	62.00	65.80	77.80	79.80	84.80	90.80	67.29$_{21.06\uparrow}$	88.08$_{4.00\uparrow}$	75.95$_{13.95\uparrow}$
Transl-Lang→En	95.20	84.00	69.80	83.00	88.40	61.00	69.00	70.40	66.00	86.20	85.20	85.20	72.20$_{25.97\uparrow}$	87.60$_{3.52\uparrow}$	78.62$_{16.62\uparrow}$
Transl-En→Lang	95.20	69.60	62.00	87.40	89.20	52.00	62.00	63.80	78.60	79.80	84.00	89.20	67.43$_{21.20\uparrow}$	88.16$_{4.08\uparrow}$	76.07$_{14.07\uparrow}$
Llama3.1-8B-Instruct															
English	95.60	65.20	27.00	84.20	88.40	26.00	52.60	62.40	73.80	76.80	84.40	89.80	55.91	86.96	68.85
Italian	95.00	62.20	28.40	86.00	92.80	32.60	61.60	70.20	76.60	78.20	85.00	90.80	59.51$_{3.60\uparrow}$	88.56$_{1.60\uparrow}$	71.62$_{2.77\uparrow}$
Chinese	95.00	65.20	48.40	83.80	88.20	32.20	58.80	68.80	76.40	78.60	86.00	91.40	62.26$_{6.35\uparrow}$	87.40$_{0.44\uparrow}$	72.73$_{3.88\uparrow}$
Multilingual	96.00	60.40	57.20	87.60	90.80	46.80	61.60	70.80	78.60	83.00	87.40	90.80	66.11$_{10.20\uparrow}$	89.64$_{2.68\uparrow}$	75.92$_{7.07\uparrow}$
Native	95.60	72.40	66.20	89.80	92.80	52.60	66.60	75.20	80.80	84.40	87.60	91.40	71.63$_{15.72\uparrow}$	90.80$_{3.84\uparrow}$	79.62$_{10.77\uparrow}$
Transl-Lang→En	95.60	87.60	74.60	88.20	89.60	68.20	73.20	74.20	72.60	88.00	86.80	87.00	76.74$_{20.83\uparrow}$	89.68$_{2.72\uparrow}$	82.13$_{13.28\uparrow}$
Transl-En→Lang	95.60	74.40	63.40	89.00	90.00	52.20	65.20	74.00	78.40	83.60	85.80	92.00	70.49$_{14.58\uparrow}$	90.04$_{3.08\uparrow}$	78.63$_{9.78\uparrow}$
Qwen2-7B-Instruct															
English	97.00	61.80	50.60	88.00	90.20	49.80	53.20	58.40	77.80	75.40	84.40	91.00	62.29	88.32	73.13
Italian	92.20	66.40	51.60	88.40	95.40	52.20	54.40	59.80	79.80	78.20	83.80	87.20	64.00$_{1.71\uparrow}$	88.28$_{0.04\downarrow}$	74.12$_{0.99\uparrow}$
Chinese	88.60	65.00	51.80	81.20	86.00	50.20	53.00	60.00	79.00	77.60	83.60	93.60	63.23$_{0.94\uparrow}$	85.40$_{2.92\downarrow}$	72.47$_{0.66\downarrow}$
Multilingual	96.00	63.20	53.80	90.60	94.00	53.00	53.40	59.60	80.40	77.60	83.40	93.00	63.83$_{1.54\uparrow}$	90.24$_{1.92\uparrow}$	74.83$_{1.70\uparrow}$
Native	97.00	71.20	54.80	93.00	95.40	51.40	60.20	63.20	83.60	81.20	89.00	93.60	67.63$_{5.34\uparrow}$	92.04$_{3.72\uparrow}$	77.80$_{4.67\uparrow}$
Transl-Lang→En	97.00	88.40	80.00	91.00	92.00	75.60	79.80	82.80	76.20	90.20	88.40	89.40	81.60$_{19.31\uparrow}$	91.92$_{3.60\uparrow}$	85.90$_{12.77\uparrow}$
Transl-En→Lang	97.00	67.20	54.60	91.20	94.80	51.00	56.80	61.60	83.00	80.20	86.40	91.80	65.80$_{3.51\uparrow}$	91.00$_{2.68\uparrow}$	76.30$_{3.17\uparrow}$
Qwen2.5-7B-Instruct															
English	97.40	62.20	56.40	89.40	93.20	50.80	53.40	58.20	83.40	80.20	88.40	93.60	64.69	90.76	75.55
Italian	96.40	65.20	57.80	89.00	95.40	49.00	52.20	58.40	82.00	83.60	87.20	92.80	64.54$_{0.15\downarrow}$	91.44$_{0.68\uparrow}$	75.75$_{0.20\uparrow}$
Chinese	95.80	65.00	58.40	89.00	92.40	50.80	51.40	58.60	83.40	82.00	89.80	94.60	65.34$_{0.65\uparrow}$	90.76$_{0.00-}$	75.93$_{0.38\uparrow}$
Multilingual	97.00	65.40	56.80	91.80	94.00	49.20	49.60	59.20	83.60	84.00	88.60	92.40	64.63$_{0.06\downarrow}$	91.84$_{1.08\uparrow}$	75.97$_{0.42\uparrow}$
Native	97.40	69.60	62.80	91.40	95.40	50.60	54.00	61.40	85.40	84.40	90.00	94.60	67.69$_{3.00\uparrow}$	92.64$_{1.88\uparrow}$	78.08$_{2.53\uparrow}$
Transl-Lang→En	97.40	87.80	80.60	91.20	91.20	75.00	77.40	80.60	74.20	89.80	87.80	88.00	80.49$_{15.80\uparrow}$	91.52$_{0.76\uparrow}$	85.08$_{9.53\uparrow}$
Transl-En→Lang	97.40	67.40	61.20	91.00	96.00	46.80	53.60	60.40	85.40	82.80	88.80	95.20	66.23$_{1.54\uparrow}$	92.48$_{1.72\uparrow}$	77.17$_{1.62\uparrow}$
NeMo-12B-Instruct															
English	96.60	57.60	58.00	83.00	93.20	50.60	51.60	73.40	64.60	73.60	80.00	91.20	62.26	87.52	72.78
Italian	96.00	57.40	55.40	82.00	95.80	48.60	53.80	73.00	64.80	73.00	81.60	90.00	62.09$_{0.17\downarrow}$	87.36$_{0.16\downarrow}$	72.62$_{0.16\downarrow}$
Chinese	96.40	58.20	52.00	83.00	92.60	49.00	53.80	70.20	66.80	71.00	81.00	91.60	61.57$_{0.69\downarrow}$	86.92$_{0.60\downarrow}$	72.13$_{0.65\downarrow}$
Multilingual	95.80	60.20	54.00	85.80	94.40	48.80	56.00	76.20	62.80	72.60	82.60	92.20	62.94$_{0.68\uparrow}$	88.16$_{0.64\uparrow}$	73.45$_{0.67\uparrow}$
Native	96.60	74.00	64.00	87.40	95.80	48.80	62.20	82.20	79.00	80.00	86.20	91.60	70.91$_{8.65\uparrow}$	90.28$_{2.76\uparrow}$	78.98$_{6.20\uparrow}$
Transl-Lang→En	96.60	88.00	75.40	87.60	89.40	70.00	74.20	77.40	74.00	88.20	85.80	84.80	77.83$_{15.57\uparrow}$	89.32$_{1.80\uparrow}$	82.62$_{9.84\uparrow}$
Transl-En→Lang	96.60	71.40	63.00	88.00	94.40	49.20	60.60	76.40	73.20	80.60	85.40	93.00	68.46$_{6.20\uparrow}$	90.52$_{3.00\uparrow}$	77.65$_{4.87\uparrow}$
Aya-Expanse-8B															
English	95.40	28.20	17.60	83.80	84.20	0.00	11.20	45.20	10.00	80.40	73.60	77.60	26.54	84.28	50.60
Italian	92.00	15.40	18.40	86.40	91.40	0.20	11.20	45.80	34.60	82.20	79.60	85.60	29.31$_{2.77\uparrow}$	87.52$_{3.24\uparrow}$	53.57$_{2.97\uparrow}$
Chinese	90.60	39.40	34.80	81.00	82.80	0.00	17.60	38.60	23.00	79.00	78.40	92.20	33.11$_{6.57\uparrow}$	85.12$_{0.84\uparrow}$	54.78$_{4.18\uparrow}$
Multilingual	93.40	52.20	52.00	88.00	90.00	5.60	46.00	58.20	43.60	84.00	83.20	90.20	48.69$_{22.15\uparrow}$	89.12$_{4.84\uparrow}$	65.53$_{14.93\uparrow}$
Native	95.40	54.00	56.20	88.80	91.40	53.40	53.40	69.40	62.40	85.60	86.20	92.20	62.14$_{35.60\uparrow}$	90.68$_{6.40\uparrow}$	74.03$_{23.43\uparrow}$
Transl-Lang→En	95.40	84.40	75.20	86.60	86.80	67.40	74.40	74.60	67.80	86.40	84.20	85.60	75.43$_{48.89\uparrow}$	88.16$_{3.88\uparrow}$	80.73$_{30.13\uparrow}$
Transl-En→Lang	95.40	54.20	54.40	89.40	90.80	45.80	51.00	65.80	60.00	83.80	86.00	90.60	59.60$_{33.06\uparrow}$	90.00$_{5.72\uparrow}$	72.27$_{21.67\uparrow}$
GPT3.5-turbo															
English	96.00	77.20	56.80	83.00	88.80	48.40	70.80	52.00	64.80	76.20	74.00	83.80	63.43	85.56	72.65
Italian	95.00	78.60	57.20	83.20	91.20	49.20	71.60	49.40	60.40	79.40	77.40	85.60	63.40$_{0.03\downarrow}$	86.88$_{1.32\uparrow}$	73.18$_{0.53\uparrow}$
Chinese	94.40	77.20	58.60	84.40	86.80	48.80	69.80	50.00	62.00	76.80	75.20	86.80	63.09$_{0.34\downarrow}$	85.84$_{0.28\uparrow}$	72.57$_{0.08\downarrow}$
Multilingual	93.80	75.00	56.80	87.40	89.60	49.20	71.40	48.00	62.80	85.80	75.80	86.80	62.71$_{0.72\downarrow}$	88.68$_{3.12\uparrow}$	73.53$_{0.88\uparrow}$
Native	96.00	85.20	67.60	87.20	90.80	48.00	77.20	53.20	61.40	87.60	77.60	87.60	67.17$_{3.74\uparrow}$	89.80$_{4.24\uparrow}$	76.60$_{3.95\uparrow}$
GPT4o-mini															
English	98.60	93.20	80.00	94.20	97.60	49.80	84.20	83.40	88.20	95.20	92.80	95.60	81.66	96.24	87.73
Italian	98.80	91.60	74.40	95.00	98.20	49.20	82.60	84.40	89.00	95.60	92.80	96.40	80.57$_{1.09\downarrow}$	96.80$_{0.56\uparrow}$	87.33$_{0.40\downarrow}$
Chinese	98.00	93.40	81.00	94.20	97.80	49.80	84.00	84.40	88.80	94.80	93.40	95.00	82.11$_{0.45\uparrow}$	95.96$_{0.28\downarrow}$	87.88$_{0.15\uparrow}$
Multilingual	98.60	93.80	80.00	96.00	98.20	50.20	85.00	86.20	92.40	96.20	94.40	95.20	83.14$_{1.48\uparrow}$	96.84$_{0.60\uparrow}$	88.85$_{1.12\uparrow}$
Native	98.60	94.80	89.60	95.20	98.20	52.20	87.60	88.80	93.80	95.20	95.00	95.20	85.97$_{4.31\uparrow}$	96.48$_{0.24\uparrow}$	90.35$_{2.62\uparrow}$

Table 12: Accuracies (%) of English, Multilingual, Native, both Monolingual ICL modes (Italian and Chinese) and two translation strategies (Transl-Lang→En, Transl-En→Lang) across 12 languages of the XCOPA dataset. AVG represents the average accuracy of the language set (LRLs, HRLs or All languages). The <u>underlined languages</u> in the table header are <u>LRLs</u>, otherwise HRLs. The subscript indicates the performance increase↑ (or decrease↓) of all other modes compared to the English ICL mode.

XL-WiC	<u>bg</u>	da	de	en	<u>et</u>	<u>fa</u>	fr	<u>hr</u>	it	ja	ko	nl	zh	<u>LRL AVG</u>	HRL AVG	ALL AVG
Llama3-8B-Instruct																
English	55.13	66.15	59.49	67.44	55.38	63.33	59.49	55.64	53.85	54.62	56.67	55.90	64.10	57.37	59.74	59.01
French	57.95	63.08	66.67	64.62	52.82	68.21	66.15	57.18	56.15	55.13	51.79	62.82	65.38	$59.04_{1.67\uparrow}$	$61.31_{1.57\uparrow}$	$60.61_{1.60\uparrow}$
Chinese	56.67	64.10	64.87	63.59	55.38	63.08	58.97	54.87	62.31	57.18	48.21	63.85	63.33	$57.50_{0.13\uparrow}$	$60.71_{0.97\uparrow}$	$59.72_{0.71\uparrow}$
Japanese	59.74	61.79	66.15	65.38	52.56	70.77	60.00	56.67	59.23	58.97	55.13	64.87	63.33	$59.94_{2.57\uparrow}$	$61.65_{1.91\uparrow}$	$61.12_{2.11\uparrow}$
Multilingual	57.18	60.00	62.82	64.10	53.85	69.23	60.00	56.92	58.97	58.46	57.44	62.82	58.72	$59.29_{1.92\uparrow}$	$60.37_{0.63\uparrow}$	$60.04_{1.03\uparrow}$
Native	63.59	55.90	69.23	67.44	62.05	68.72	66.15	58.21	59.49	58.97	66.67	66.41	63.33	$63.14_{5.77\uparrow}$	$63.73_{3.99\uparrow}$	$63.55_{4.54\uparrow}$
Llama3.1-8B-Instruct																
English	51.54	50.77	61.28	66.92	44.62	29.74	56.67	51.79	15.90	48.46	33.85	51.28	57.44	44.42	49.17	47.71
French	54.62	60.00	66.41	64.62	49.74	22.56	62.31	57.44	52.56	49.49	53.33	65.13	54.10	$46.09_{1.67\uparrow}$	$58.66_{9.49\uparrow}$	$54.79_{7.08\uparrow}$
Chinese	54.87	62.82	62.31	61.79	53.08	55.64	56.15	57.95	56.67	49.74	48.21	62.31	62.05	$55.38_{10.96\uparrow}$	$58.01_{8.84\uparrow}$	$57.20_{9.49\uparrow}$
Japanese	54.36	60.00	63.59	60.00	47.69	63.33	58.72	52.05	58.21	53.59	48.97	58.21	56.41	$54.36_{9.94\uparrow}$	$57.52_{8.35\uparrow}$	$56.55_{8.84\uparrow}$
Multilingual	57.69	60.00	65.38	62.56	55.13	58.46	57.95	56.92	55.38	54.36	52.31	64.62	58.46	$57.05_{12.63\uparrow}$	$59.00_{9.83\uparrow}$	$58.40_{10.69\uparrow}$
Native	61.79	64.10	68.46	66.92	62.05	70.51	62.31	57.18	56.15	53.59	63.08	69.49	62.05	$62.88_{18.46\uparrow}$	$62.91_{13.74\uparrow}$	$62.90_{15.19\uparrow}$
Qwen2-7B-Instruct																
English	28.21	38.46	66.67	68.46	56.92	53.85	63.33	54.87	39.23	1.28	15.90	64.36	0.77	48.46	39.83	42.49
French	53.59	51.54	64.87	66.92	53.85	58.21	63.08	56.41	49.74	25.90	33.08	62.82	14.10	$55.51_{7.05\uparrow}$	$48.01_{8.18\uparrow}$	$50.32_{7.83\uparrow}$
Chinese	47.18	27.44	68.46	67.44	57.44	58.97	61.79	55.13	40.51	56.41	37.69	60.00	68.72	$54.68_{6.22\uparrow}$	$54.27_{14.44\uparrow}$	$54.40_{11.91\uparrow}$
Japanese	56.15	30.26	66.41	66.92	56.15	60.26	61.79	56.67	47.95	60.77	59.49	66.41	40.26	$57.31_{8.85\uparrow}$	$55.58_{15.75\uparrow}$	$56.11_{13.62\uparrow}$
Multilingual	53.59	60.00	65.64	65.64	58.46	58.72	60.26	54.36	56.67	58.72	66.92	67.18	57.95	$56.28_{7.82\uparrow}$	$62.11_{22.28\uparrow}$	$60.32_{17.83\uparrow}$
Native	55.13	62.05	69.49	68.46	59.23	62.05	63.08	54.62	60.51	60.77	68.21	64.87	68.72	$57.76_{9.30\uparrow}$	$65.13_{25.30\uparrow}$	$62.86_{20.37\uparrow}$
Qwen2.5-7B-Instruct																
English	58.97	61.28	73.33	70.51	55.64	53.59	65.90	63.08	57.95	55.64	64.10	67.69	59.74	57.82	64.02	62.11
French	55.13	57.18	68.97	65.13	55.38	54.10	64.87	57.18	62.05	57.95	62.82	65.64	62.56	$55.45_{2.37\downarrow}$	$63.02_{1.00\downarrow}$	$60.69_{1.42\downarrow}$
Chinese	53.33	54.87	67.18	67.18	54.62	51.28	63.85	57.44	57.95	64.87	62.82	61.79	66.67	$54.17_{3.65\downarrow}$	$63.02_{1.00\downarrow}$	$60.30_{1.81\downarrow}$
Japanese	53.59	54.62	67.44	67.18	54.10	49.74	63.08	57.18	56.41	64.10	63.33	65.13	60.00	$53.65_{4.17\downarrow}$	$62.36_{1.66\downarrow}$	$59.68_{2.43\downarrow}$
Multilingual	57.44	60.51	69.23	68.21	54.62	55.38	61.28	56.41	57.18	64.36	66.67	67.69	64.62	$55.96_{1.86\downarrow}$	$64.42_{0.40\uparrow}$	$61.81_{0.30\downarrow}$
Native	60.00	63.33	72.56	70.51	55.90	63.33	64.87	55.64	62.31	64.10	70.77	72.82	66.67	$58.72_{0.90\uparrow}$	$67.55_{3.53\uparrow}$	$64.83_{2.72\uparrow}$
NeMo-12B-Instruct																
English	52.56	63.85	70.51	65.64	51.79	49.23	61.03	52.56	38.97	23.08	53.59	67.44	54.10	51.54	55.36	54.18
French	52.05	57.69	66.41	64.36	51.54	48.46	60.77	57.18	57.18	56.67	55.13	59.74	54.10	$52.31_{0.77\uparrow}$	$59.12_{3.76\uparrow}$	$57.02_{2.84\uparrow}$
Chinese	47.95	57.95	63.33	63.85	50.51	47.18	59.49	52.31	56.41	57.18	54.62	61.28	62.82	$49.49_{2.05\downarrow}$	$59.66_{4.30\uparrow}$	$56.53_{2.35\uparrow}$
Japanese	48.21	53.08	60.77	64.10	51.28	47.44	58.46	52.31	56.92	64.36	56.67	58.72	53.59	$49.81_{1.73\downarrow}$	$58.52_{3.16\uparrow}$	$55.84_{1.66\uparrow}$
Multilingual	51.54	56.67	67.18	61.28	52.05	49.74	61.28	51.28	54.36	57.69	56.67	58.72	59.49	$51.15_{0.39\downarrow}$	$59.26_{3.90\uparrow}$	$56.77_{2.59\uparrow}$
Native	57.44	57.69	70.00	65.64	57.69	66.92	60.77	58.72	55.13	64.36	64.87	65.90	62.82	$60.19_{8.65\uparrow}$	$63.02_{7.66\uparrow}$	$62.15_{7.97\uparrow}$
Aya-Expanse-8B																
English	53.08	57.44	60.51	66.41	58.46	66.41	57.44	57.18	26.41	44.62	59.74	61.79	60.51	58.78	54.99	56.15
French	55.64	59.23	71.79	63.85	57.69	72.31	64.87	54.10	61.79	65.38	65.64	70.77	65.64	$59.94_{1.16\uparrow}$	$65.44_{10.45\uparrow}$	$63.75_{7.60\uparrow}$
Chinese	55.90	64.62	68.97	63.85	56.15	73.59	61.28	53.08	58.46	59.49	67.69	66.41	61.79	$59.68_{0.90\uparrow}$	$63.62_{8.63\uparrow}$	$62.41_{6.26\uparrow}$
Japanese	59.49	58.72	70.00	66.67	58.21	71.54	63.59	55.13	54.62	61.03	68.46	70.00	64.36	$61.09_{2.31\uparrow}$	$64.16_{9.17\uparrow}$	$63.21_{7.06\uparrow}$
Multilingual	58.46	58.46	69.74	63.33	56.41	69.23	66.15	55.38	61.28	65.64	68.21	71.28	63.08	$59.87_{1.09\uparrow}$	$65.24_{10.25\uparrow}$	$63.59_{7.44\uparrow}$
Native	51.54	63.08	71.54	66.41	56.92	78.97	64.87	59.49	61.03	61.03	67.95	70.51	61.79	$61.73_{2.95\uparrow}$	$65.36_{10.37\uparrow}$	$64.24_{8.09\uparrow}$
GPT3.5-turbo																
English	54.36	50.77	62.31	63.59	54.62	54.10	58.46	51.79	32.82	30.77	56.67	65.13	21.03	53.72	49.06	50.49
French	55.13	56.67	64.62	62.56	58.97	52.05	58.72	56.41	56.41	58.46	59.23	61.79	58.97	$55.64_{1.92\uparrow}$	$59.72_{10.66\uparrow}$	$58.46_{7.97\uparrow}$
Chinese	53.59	61.28	62.56	59.74	55.13	54.36	56.41	56.92	55.90	55.64	56.41	59.23	53.85	$55.00_{1.28\uparrow}$	$57.89_{8.83\uparrow}$	$57.00_{6.51\uparrow}$
Japanese	53.33	55.38	65.90	63.08	58.97	53.08	58.97	55.90	56.15	56.41	55.90	64.36	54.87	$55.32_{1.60\uparrow}$	$59.00_{9.94\uparrow}$	$57.87_{7.38\uparrow}$
Multilingual	52.82	60.00	66.92	59.23	60.00	54.36	60.77	56.41	53.59	56.67	61.54	63.33	59.49	$55.90_{2.18\uparrow}$	$60.17_{11.11\uparrow}$	$58.86_{8.37\uparrow}$
Native	54.62	62.56	64.87	63.59	59.49	58.46	58.21	60.26	54.36	57.95	59.74	64.87	53.85	$58.21_{4.49\uparrow}$	$60.11_{11.05\uparrow}$	$59.53_{9.04\uparrow}$
GPT4o-mini																
English	68.72	27.95	74.36	73.33	62.56	28.46	71.79	65.64	38.21	4.10	53.59	75.64	5.38	56.35	47.15	49.98
French	66.41	58.97	72.82	67.95	60.51	39.23	71.28	68.97	61.28	63.08	72.05	71.03	70.00	$58.78_{2.43\uparrow}$	$67.61_{20.46\uparrow}$	$64.89_{14.91\uparrow}$
Chinese	67.44	58.21	73.08	69.23	58.72	50.26	67.69	68.97	58.72	66.15	70.77	71.79	73.59	$61.35_{5.00\uparrow}$	$67.69_{20.54\uparrow}$	$65.74_{15.76\uparrow}$
Japanese	66.92	52.82	72.56	67.69	60.00	39.23	66.67	65.13	57.95	66.67	68.46	73.85	67.95	$57.82_{1.47\uparrow}$	$66.07_{18.92\uparrow}$	$63.53_{13.55\uparrow}$
Multilingual	66.41	65.13	72.31	68.72	62.05	64.62	68.97	67.95	65.64	64.10	71.28	73.85	67.69	$65.26_{8.91\uparrow}$	$68.63_{21.48\uparrow}$	$67.59_{17.61\uparrow}$
Native	71.54	70.26	76.15	73.33	65.13	82.82	71.79	67.95	67.95	66.15	76.41	75.64	72.05	$71.86_{15.51\uparrow}$	$72.19_{25.04\uparrow}$	$72.09_{22.11\uparrow}$

Table 13: Accuracies (%) of `English`, `Multilingual`, `Native` all `Monolingual` ICL modes (`French`, `Chinese` and `Japanese`) across 13 languages of the XL-WiC dataset. AVG represents the average accuracy of the language set (LRLs, HRLs or All languages). The <u>underlined languages</u> in the table header are <u>LRLs</u>, otherwise HRLs. The subscript indicates the performance increase↑ (or decrease↓) of all other modes compared to the `English` ICL mode.

XQuAD	ar	de	<u>el</u>	en	es	hi	<u>ro</u>	ru	<u>th</u>	tr	<u>vi</u>	zh	LRL Avg	HRL Avg	ALL Avg
Llama3-8B-Instruct															
`English`	60.90	75.80	69.10	86.80	79.10	69.90	79.20	62.20	69.90	72.20	73.90	77.40	72.02	73.54	73.03
`Chinese`	65.80	76.30	68.40	85.50	79.80	71.00	79.40	65.60	75.30	74.30	75.90	81.30	$73.53_{1.51\uparrow}$	$75.56_{2.02\uparrow}$	$74.88_{1.85\uparrow}$
`German`	63.50	75.10	69.40	87.10	79.20	70.80	79.00	62.70	72.50	73.30	76.70	77.90	$72.93_{0.91\uparrow}$	$74.44_{0.90\uparrow}$	$73.93_{0.90\uparrow}$
`Multilingual`	65.70	75.80	71.90	87.40	80.60	71.30	80.10	65.90	73.90	74.70	76.70	81.00	$74.30_{2.28\uparrow}$	$75.98_{2.44\uparrow}$	$75.42_{2.39\uparrow}$
`Native`	66.70	75.10	72.00	86.80	79.40	73.60	79.90	63.90	75.40	75.60	78.60	81.30	$75.22_{3.20\uparrow}$	$75.92_{2.38\uparrow}$	$75.69_{2.66\uparrow}$
Llama3.1-8B-Instruct															
`English`	60.80	73.20	65.40	85.80	77.80	66.50	73.70	53.60	67.40	66.20	71.90	73.80	68.25	70.39	69.68
`Chinese`	60.00	72.50	62.70	75.70	76.20	67.60	73.20	54.50	68.40	65.30	70.70	75.00	$67.98_{0.27\downarrow}$	$68.74_{1.65\downarrow}$	$68.48_{1.20\downarrow}$
`German`	59.60	72.80	66.60	66.80	78.30	67.40	75.10	55.70	67.70	66.90	73.40	74.50	$69.20_{0.95\uparrow}$	$68.50_{1.89\downarrow}$	$68.73_{0.95\downarrow}$
`Multilingual`	60.80	73.50	68.10	83.50	79.70	69.40	76.00	58.20	71.00	66.60	73.00	75.10	$71.12_{2.87\uparrow}$	$71.30_{0.91\uparrow}$	$71.24_{1.56\uparrow}$
`Native`	61.60	72.80	67.30	85.80	77.40	69.30	76.90	59.00	71.60	67.70	75.80	75.00	$71.28_{3.03\uparrow}$	$71.89_{1.50\uparrow}$	$71.68_{2.00\uparrow}$
Qwen2-7B-Instruct															
`English`	60.40	69.60	39.10	80.70	77.20	50.80	69.60	54.90	54.10	61.00	74.40	80.10	53.40	69.79	64.32
`Chinese`	58.00	68.40	36.80	83.10	74.60	48.80	67.40	53.80	56.60	61.40	72.80	83.70	$52.40_{1.00\downarrow}$	$69.48_{0.31\downarrow}$	$63.78_{0.54\downarrow}$
`German`	61.00	70.40	49.20	82.40	77.10	54.10	72.30	59.60	63.10	65.80	73.60	82.20	$59.67_{6.27\uparrow}$	$71.51_{1.72\uparrow}$	$67.57_{3.25\uparrow}$
`Multilingual`	60.60	71.50	48.50	83.10	78.10	55.00	72.80	60.80	64.40	64.90	74.60	82.10	$60.18_{6.78\uparrow}$	$71.96_{2.17\uparrow}$	$68.03_{3.71\uparrow}$
`Native`	62.80	70.40	55.70	80.70	77.20	58.60	72.90	65.20	72.40	65.70	75.70	83.70	$64.90_{11.50\uparrow}$	$72.68_{2.89\uparrow}$	$70.08_{5.76\uparrow}$
Qwen2.5-7B-Instruct															
`English`	63.70	74.30	57.00	85.50	80.60	62.20	80.20	67.00	76.50	69.90	75.10	85.10	68.98	75.15	73.09
`Chinese`	57.80	67.30	54.10	69.60	67.40	60.90	77.10	61.30	76.90	70.10	65.30	87.20	$67.25_{1.73\downarrow}$	$68.25_{6.90\downarrow}$	$67.92_{5.17\downarrow}$
`German`	63.00	75.90	57.30	82.70	80.50	62.50	79.80	67.80	76.20	71.60	76.00	85.20	$68.95_{0.03\downarrow}$	$75.34_{0.19\uparrow}$	$73.21_{0.12\uparrow}$
`Multilingual`	63.70	75.50	58.40	81.20	80.20	62.30	80.40	67.90	77.90	70.80	76.30	86.20	$69.75_{0.77\uparrow}$	$75.23_{0.08\uparrow}$	$73.40_{0.31\uparrow}$
`Native`	67.90	75.90	61.00	85.50	79.80	64.90	79.50	69.30	78.40	73.10	78.10	87.20	$70.95_{1.97\uparrow}$	$77.10_{1.95\uparrow}$	$75.05_{1.96\uparrow}$
NeMo-12B-Instruct															
`English`	56.50	70.30	52.90	83.90	78.70	51.50	73.60	56.50	65.70	63.50	65.90	51.30	60.93	65.82	64.19
`Chinese`	60.40	71.80	61.30	84.30	78.40	69.90	73.70	60.80	70.60	65.60	75.50	76.2	$68.80_{7.87\uparrow}$	$71.63_{5.81\uparrow}$	$70.71_{6.52\uparrow}$
`German`	60.00	70.70	61.30	84.50	79.10	69.40	74.40	60.70	69.40	65.90	74.80	72.40	$68.62_{7.69\uparrow}$	$71.01_{5.19\uparrow}$	$70.22_{6.03\uparrow}$
`Multilingual`	61.30	72.30	62.60	84.80	79.60	70.00	75.40	62.40	72.30	68.00	75.20	75.20	$70.08_{9.15\uparrow}$	$72.35_{6.53\uparrow}$	$71.59_{7.40\uparrow}$
`Native`	61.50	70.70	63.80	83.90	76.40	70.50	76.00	63.00	72.10	68.20	74.10	76.20	$70.60_{9.67\uparrow}$	$71.75_{5.93\uparrow}$	$71.37_{7.18\uparrow}$
Aya-Expanse-8B															
`English`	71.60	76.70	74.00	87.60	83.70	70.30	83.00	64.70	54.60	68.40	80.20	80.60	70.48	76.69	74.62
`Chinese`	72.00	72.20	72.00	66.10	82.50	69.20	81.60	63.70	54.40	63.50	79.50	82.10	$69.30_{1.18\downarrow}$	$72.70_{3.99\downarrow}$	$71.57_{3.05\downarrow}$
`German`	71.60	74.70	68.40	63.40	78.20	69.50	76.70	50.60	55.10	59.50	78.20	68.70	$67.42_{3.06\downarrow}$	$68.11_{8.58\downarrow}$	$67.88_{6.74\downarrow}$
`Multilingual`	72.50	74.20	72.30	78.20	83.10	71.20	82.40	64.10	55.10	67.90	79.80	78.20	$70.25_{0.23\downarrow}$	$74.75_{1.94\downarrow}$	$73.25_{1.37\downarrow}$
`Native`	72.00	74.70	72.00	87.60	82.80	67.70	81.30	64.60	51.50	67.60	79.60	82.10	$68.12_{2.36\downarrow}$	$76.38_{0.31\downarrow}$	$73.62_{1.00\downarrow}$

Table 14: Accuracies (%) of `English`, `Multilingual`, `Native` all `Monolingual` ICL modes (`German` and `Chinese`) across 12 languages of the XQuAD dataset. AVG represents the average accuracy of the language set (LRLs, HRLs or All languages). The <u>underlined languages</u> in the table header are <u>LRLs</u>, otherwise HRLs. The subscript indicates the performance increase↑ (or decrease↓) of all other modes compared to the `English` ICL mode.

McNemar's Test MGSM ICL Mode M1 vs M2 Comparison	Low-Resource Languages χ^2	p-value	Sig. Level	#Both ✗	#M1 ✗ M2 ✓	#M1 ✓ M2 ✗	#Both ✓	High-Resource Languages χ^2	p-value	Sig. Level	#Both ✗	#M1 ✗ M2 ✓	#M1 ✓ M2 ✗	#Both ✓
Llama3-8B-Instruct														
`English vs French`	0.85	3.56×10^{-1}		280	78	91	551	0.08	7.84×10^{-1}		300	104	109	1237
`English vs Chinese`	0.65	4.20×10^{-1}		271	87	99	543	1.09	2.97×10^{-1}		295	109	126	1220
`English vs Japanese`	2.01	1.57×10^{-1}		278	80	100	542	2.58	1.08×10^{-1}		296	108	134	1212
`English vs Multilingual`	7.53	6.07×10^{-3}	**	277	81	121	521	1.02	3.12×10^{-1}		302	102	118	1228
`English vs Native`	2.16	1.41×10^{-1}		276	82	103	539	0.00	1.00×10^{0}		294	110	109	1237
Llama3.1-8B-Instruct														
`English vs French`	11.75	6.08×10^{-4}	***	311	118	70	501	25.57	4.26×10^{-7}	***	261	192	104	1193
`English vs Chinese`	17.39	3.04×10^{-5}	***	287	142	79	492	19.63	9.39×10^{-6}	***	263	190	112	1185
`English vs Japanese`	49.92	1.60×10^{-12}	***	255	174	64	507	14.77	1.22×10^{-4}	***	267	186	118	1179
`English vs Multilingual`	33.24	8.16×10^{-9}	***	268	161	72	499	24.12	9.03×10^{-7}	***	248	205	116	1181
`English vs Native`	50.67	1.09×10^{-12}	***	246	183	69	502	27.57	1.52×10^{-7}	***	253	200	107	1190
Qwen2-7B-Instruct														
`English vs French`	0.40	5.30×10^{-1}		505	58	66	371	0.00	1.00×10^{0}		277	89	88	1296
`English vs Chinese`	0.53	4.68×10^{-1}		481	82	72	365	3.01	8.30×10^{-2}		266	100	76	1308
`English vs Japanese`	0.33	5.68×10^{-1}		492	71	79	358	0.08	7.82×10^{-1}		264	102	107	1277
`English vs Multilingual`	7.36	6.67×10^{-3}	**	451	112	74	363	2.41	1.21×10^{-1}		254	112	89	1295
`English vs Native`	56.78	4.88×10^{-14}	***	386	177	60	377	20.92	4.80×10^{-6}	***	232	134	68	1316
Qwen2.5-7B-Instruct														
`English vs French`	0.56	4.56×10^{-1}		344	62	53	541	0.15	7.02×10^{-1}		186	52	57	1455
`English vs Chinese`	0.21	6.48×10^{-1}		343	63	57	537	1.07	3.02×10^{-1}		164	74	61	1451
`English vs Japanese`	0.69	4.07×10^{-1}		342	64	54	540	3.20	7.38×10^{-2}		158	80	58	1454
`English vs Multilingual`	0.00	1.00×10^{0}		342	64	63	531	0.71	3.99×10^{-1}		176	62	52	1460
`English vs Native`	0.38	5.36×10^{-1}		318	88	79	515	1.76	1.85×10^{-1}		166	72	56	1456
NeMo-12B-Instruct														
`English vs French`	22.44	2.17×10^{-6}	***	278	152	79	491	8.98	2.73×10^{-3}	**	222	147	99	1282
`English vs Chinese`	23.24	1.43×10^{-6}	***	271	159	83	487	10.82	1.01×10^{-3}	**	238	131	82	1299
`English vs Japanese`	17.17	3.41×10^{-5}	***	274	156	90	480	16.32	5.35×10^{-5}	***	211	158	93	1288
`English vs Multilingual`	16.93	3.87×10^{-5}	***	285	145	82	488	5.66	1.74×10^{-2}	*	229	140	102	1279
`English vs Native`	29.80	4.79×10^{-8}	***	259	171	83	487	7.05	7.93×10^{-3}	**	235	134	93	1288
Aya-Expanse-8B														
`English vs French`	0.81	3.67×10^{-1}		611	95	82	212	0.17	6.78×10^{-1}		323	108	101	1218
`English vs Chinese`	2.70	1.00×10^{-1}		605	101	78	216	0.11	7.44×10^{-1}		317	114	120	1199
`English vs Japanese`	0.47	4.95×10^{-1}		614	92	82	212	0.31	5.75×10^{-1}		307	124	134	1185
`English vs Multilingual`	0.20	6.52×10^{-1}		614	92	85	209	0.00	9.51×10^{-1}		296	135	135	1184
`English vs Native`	4.62	3.16×10^{-2}	*	586	120	88	206	2.47	1.16×10^{-1}		312	119	95	1224
GPT3.5-turbo														
`English vs French`	1.93	1.64×10^{-1}		455	103	125	317	25.29	4.92×10^{-7}	***	281	235	137	1097
`English vs Chinese`	0.02	8.98×10^{-1}		438	120	123	319	32.82	1.01×10^{-8}	***	280	236	126	1108
`English vs Japanese`	0.44	5.09×10^{-1}		449	109	120	322	40.56	1.90×10^{-10}	***	278	238	117	1117
`English vs Multilingual`	17.81	2.44×10^{-5}	***	398	160	92	350	27.69	1.43×10^{-7}	***	289	227	127	1107
`English vs Native`	43.05	5.34×10^{-11}	***	374	184	77	365	65.82	4.93×10^{-16}	***	264	252	99	1135
GPT4o-mini														
`English vs French`	0.24	6.25×10^{-1}		106	36	31	827	0.81	3.68×10^{-1}		136	35	44	1535
`English vs Chinese`	0.77	3.82×10^{-1}		106	36	28	830	0.05	8.17×10^{-1}		132	39	36	1543
`English vs Japanese`	1.69	1.93×10^{-1}		106	36	49	809	0.19	6.61×10^{-1}		132	39	44	1535
`English vs Multilingual`	1.27	2.61×10^{-1}		108	34	45	813	0.01	9.11×10^{-1}		131	40	40	1539
`English vs Native`	0.30	5.85×10^{-1}		103	39	45	813	1.09	2.95×10^{-1}		139	32	42	1537

Table 15: McNemar's test results of ICL modes on LRL and HRL splits of MGSM dataset. Baseline is the `English` mode, compared with other `Monolingual`, `Multilingual`, and `Native` modes.

McNemar's Test XCOPA ICL Mode M1 vs M2 Comparison	Low-Resource Languages χ^2	p-value	Sig. Level	#Both ✗	#M1 ✗ M2 ✓	#M1 ✓ M2 ✗	#Both ✓	High-Resource Languages χ^2	p-value	Sig. Level	#Both ✗	#M1 ✗ M2 ✓	#M1 ✓ M2 ✗	#Both ✓
Llama3-8B-Instruct														
`English` vs `Italian`	274.27	1.33×10^{-61}	***	1246	636	166	1452	8.84	2.95×10^{-3}	**	287	111	70	2032
`English` vs `Chinese`	347.11	1.80×10^{-77}	***	1177	705	157	1461	5.85	1.55×10^{-2}	*	288	110	76	2026
`English` vs `Multilingual`	366.08	1.33×10^{-81}	***	1163	719	153	1465	15.76	7.21×10^{-5}	***	274	124	68	2034
`English` vs `Native`	468.19	7.94×10^{-104}	***	935	947	210	1408	45.38	1.63×10^{-11}	***	240	158	58	2044
Llama3.1-8B-Instruct														
`English` vs `Italian`	27.32	1.73×10^{-7}	***	1194	349	223	1734	9.27	2.32×10^{-3}	**	224	102	62	2112
`English` vs `Chinese`	74.68	5.53×10^{-18}	***	1105	438	216	1741	0.63	4.28×10^{-1}		241	85	74	2100
`English` vs `Multilingual`	157.05	5.00×10^{-36}	***	961	582	225	1732	21.04	4.49×10^{-6}	***	189	137	70	2104
`English` vs `Native`	276.51	4.32×10^{-62}	***	723	820	270	1687	46.05	1.16×10^{-11}	***	180	146	50	2124
Qwen2-7B-Instruct														
`English` vs `Italian`	8.45	3.65×10^{-3}	**	1084	236	176	2004	0.00	1.00×10^{0}		216	76	77	2131
`English` vs `Chinese`	2.59	1.07×10^{-1}		1106	214	181	1999	26.05	3.33×10^{-7}	***	229	63	136	2072
`English` vs `Multilingual`	6.11	1.35×10^{-2}	*	1063	257	203	1977	14.73	1.24×10^{-4}	***	193	99	51	2157
`English` vs `Native`	40.18	2.31×10^{-10}	***	796	524	337	1843	47.82	4.67×10^{-12}	***	157	135	42	2166
Qwen2.5-7B-Instruct														
`English` vs `Italian`	0.03	8.71×10^{-1}		937	299	304	1960	1.82	1.78×10^{-1}		152	79	62	2207
`English` vs `Chinese`	1.03	3.10×10^{-1}		990	246	223	2041	0.01	9.30×10^{-1}		167	64	64	2205
`English` vs `Multilingual`	0.00	9.68×10^{-1}		931	305	307	1957	4.60	3.20×10^{-2}	*	144	87	60	2209
`English` vs `Native`	11.49	6.98×10^{-4}	***	713	523	418	1846	13.65	2.20×10^{-4}	***	130	101	54	2215
NeMo-12B-Instruct														
`English` vs `Italian`	0.05	8.30×10^{-1}		1054	267	273	1906	0.06	8.13×10^{-1}		234	78	82	2106
`English` vs `Chinese`	0.93	3.35×10^{-1}		1048	273	297	1882	1.06	3.03×10^{-1}		227	85	100	2088
`English` vs `Multilingual`	0.72	3.95×10^{-1}		943	378	354	1825	1.22	2.69×10^{-1}		212	100	84	2104
`English` vs `Native`	88.46	5.18×10^{-21}	***	654	667	364	1815	23.24	1.43×10^{-6}	***	178	134	65	2123
Aya-Expanse-8B														
`English` vs `Italian`	18.11	2.09×10^{-5}	***	2268	303	206	723	28.96	7.39×10^{-8}	***	242	151	70	2037
`English` vs `Chinese`	100.08	1.46×10^{-23}	***	2194	377	147	782	1.51	2.19×10^{-1}		250	143	122	1985
`English` vs `Multilingual`	637.99	9.13×10^{-141}	***	1714	857	82	847	63.44	1.66×10^{-15}	***	219	174	53	2054
`English` vs `Native`	1003.90	2.55×10^{-220}	***	1176	1395	149	780	104.47	1.60×10^{-24}	***	192	201	41	2066
GPT3.5-turbo														
`English` vs `Italian`	0.00	1.00×10^{0}		890	390	391	1829	3.89	4.85×10^{-2}	*	213	148	115	2024
`English` vs `Chinese`	0.14	7.06×10^{-1}		860	420	432	1788	0.11	7.36×10^{-1}		199	162	155	1984
`English` vs `Multilingual`	0.67	4.12×10^{-1}		864	416	441	1779	22.46	2.15×10^{-6}	***	190	171	93	2046
`English` vs `Native`	15.10	1.02×10^{-4}	***	655	625	494	1726	39.95	2.61×10^{-10}	***	170	191	85	2054
GPT4o-mini														
`English` vs `Italian`	5.19	2.28×10^{-2}	*	529	113	151	2707	2.82	9.33×10^{-2}		57	37	23	2383
`English` vs `Chinese`	1.00	3.18×10^{-1}		521	121	105	2753	0.61	4.35×10^{-1}		68	26	33	2373
`English` vs `Multilingual`	9.63	1.91×10^{-3}	**	481	161	109	2749	3.32	6.84×10^{-2}		57	37	22	2384
`English` vs `Native`	51.49	7.21×10^{-13}	***	348	294	143	2715	0.40	5.25×10^{-1}		60	34	28	2378

Table 16: McNemar's test results of ICL modes on LRL and HRL splits of XCOPA dataset. Baseline is the `English` mode, compared with other `Monolingual`, `Multilingual`, and `Native` modes.

McNemar's Test XL-WiC ICL Mode M1 vs M2 Comparison	Low-Resource Languages χ^2	p-value	Sig. Level	#Both ✗	#M1 ✗ M2 ✓	#M1 ✓ M2 ✗	#Both ✓	High-Resource Languages χ^2	p-value	Sig. Level	#Both ✗	#M1 ✗ M2 ✓	#M1 ✓ M2 ✗	#Both ✓
Llama3-8B-Instruct														
`English` vs `French`	1.85	1.74×10^{-1}		483	182	156	739	3.70	5.45×10^{-2}		991	422	367	1730
`English` vs `Chinese`	0.00	9.60×10^{-1}		469	196	194	701	1.38	2.41×10^{-1}		1000	413	379	1718
`English` vs `Japanese`	3.67	5.53×10^{-2}		438	227	187	708	4.59	3.22×10^{-2}	*	905	508	441	1656
`English` vs `Multilingual`	1.89	1.70×10^{-1}		427	238	208	687	0.45	5.00×10^{-1}		917	496	474	1623
`English` vs `Native`	13.85	1.98×10^{-4}	***	334	331	241	654	19.84	8.43×10^{-6}	***	856	557	417	1680
Llama3.1-8B-Instruct														
`English` vs `French`	1.47	2.26×10^{-1}		641	226	200	493	113.05	2.10×10^{-26}	***	1130	654	321	1405
`English` vs `Chinese`	66.13	4.22×10^{-16}	***	563	304	133	560	94.72	2.19×10^{-22}	***	1125	659	349	1377
`English` vs `Japanese`	50.57	1.15×10^{-12}	***	555	312	157	536	96.34	9.66×10^{-23}	***	1195	589	296	1430
`English` vs `Multilingual`	82.26	1.19×10^{-19}	***	535	332	135	558	94.44	2.52×10^{-22}	***	985	799	454	1272
`English` vs `Native`	134.15	5.06×10^{-31}	***	416	451	163	530	197.07	9.10×10^{-45}	***	956	828	346	1380
Qwen2-7B-Instruct														
`English` vs `French`	26.64	2.45×10^{-7}	***	526	278	168	588	99.87	1.62×10^{-23}	***	1559	553	266	1132
`English` vs `Chinese`	25.96	3.48×10^{-7}	***	578	226	129	627	274.42	1.23×10^{-61}	***	1392	720	213	1185
`English` vs `Japanese`	34.50	4.26×10^{-9}	***	463	341	203	553	311.88	8.52×10^{-70}	***	1347	765	212	1186
`English` vs `Multilingual`	20.45	6.13×10^{-6}	***	385	419	297	459	468.48	6.86×10^{-104}	***	1070	1042	260	1138
`English` vs `Native`	35.09	3.15×10^{-9}	***	436	368	223	533	645.95	1.70×10^{-142}	***	1059	1053	165	1233
Qwen2.5-7B-Instruct														
`English` vs `French`	5.71	1.69×10^{-2}	*	563	95	132	770	2.29	1.30×10^{-1}		1028	235	270	1977
`English` vs `Chinese`	13.46	2.44×10^{-4}	***	570	88	145	757	1.80	1.80×10^{-1}		959	304	339	1908
`English` vs `Japanese`	17.89	2.34×10^{-5}	***	576	82	147	755	5.40	2.02×10^{-2}	*	991	272	330	1917
`English` vs `Multilingual`	3.10	7.83×10^{-2}		546	112	141	761	0.30	5.85×10^{-1}		973	290	276	1971
`English` vs `Native`	0.45	5.04×10^{-1}		462	196	182	720	25.30	4.91×10^{-7}	***	902	361	237	2010
NeMo-12B-Instruct														
`English` vs `French`	0.95	3.31×10^{-1}		686	70	58	746	25.69	4.01×10^{-7}	***	1167	400	268	1675
`English` vs `Chinese`	7.17	7.41×10^{-3}	**	705	51	83	721	28.59	8.95×10^{-8}	***	1098	469	318	1625
`English` vs `Japanese`	5.08	2.42×10^{-2}	*	703	53	80	724	16.60	4.62×10^{-5}	***	1147	420	309	1634
`English` vs `Multilingual`	0.17	6.83×10^{-1}		684	72	78	726	28.68	8.56×10^{-8}	***	1176	391	254	1689
`English` vs `Native`	44.12	3.09×10^{-11}	***	485	271	136	668	87.70	7.63×10^{-21}	***	1023	544	275	1668
Aya-Expanse-8B														
`English` vs `French`	0.50	4.79×10^{-1}		346	297	279	638	118.02	1.71×10^{-27}	***	829	751	384	1546
`English` vs `Chinese`	0.31	5.79×10^{-1}		361	282	268	649	89.50	3.06×10^{-21}	***	919	661	358	1572
`English` vs `Japanese`	2.26	1.33×10^{-1}		354	289	253	664	91.35	1.20×10^{-21}	***	855	725	403	1527
`English` vs `Multilingual`	0.43	5.12×10^{-1}		337	306	289	628	109.78	1.10×10^{-25}	***	813	767	407	1523
`English` vs `Native`	3.20	7.35×10^{-2}		304	339	293	624	133.37	7.51×10^{-31}	***	904	676	312	1618

Table 17: McNemar's test results of ICL modes on LRL and HRL splits of XL-WiC dataset. Baseline is the `English` mode, compared with other `Monolingual`, `Multilingual`, and `Native` modes.

McNemar's Test XQuAD ICL Mode M1 vs M2 Comparison	Low-Resource Languages χ^2	*p*-value	Sig. Level	#Both ✖	#M1 ✖ M2 ✔	#M1 ✔ M2 ✖	#Both ✔	High-Resource Languages χ^2	*p*-value	Sig. Level	#Both ✖	#M1 ✖ M2 ✔	#M1 ✔ M2 ✖	#Both ✔
Llama3-8B-Instruct														
`English` vs `German`	5.02	2.50×10^{-2}	*	979	140	104	2777	10.55	1.16×10^{-3}	**	1842	275	203	5680
`English` vs `Chinese`	10.61	1.12×10^{-3}	**	925	194	134	2747	42.92	5.72×10^{-11}	***	1734	383	221	5662
`English` vs `Multilingual`	27.27	1.77×10^{-7}	***	925	194	103	2778	68.30	1.40×10^{-16}	***	1744	373	178	5705
`English` vs `Native`	42.90	5.77×10^{-11}	***	867	252	124	2757	59.08	1.51×10^{-14}	***	1716	401	210	5673
Llama3.1-8B-Instruct														
`English` vs `German`	4.72	2.98×10^{-2}	*	1106	164	126	2604	32.10	1.47×10^{-8}	***	2094	275	426	5205
`English` vs `Chinese`	0.28	5.97×10^{-1}		1097	173	184	2546	25.24	5.07×10^{-7}	***	2095	274	406	5225
`English` vs `Multilingual`	40.49	1.98×10^{-10}	***	1052	218	103	2627	8.95	2.77×10^{-3}	**	2043	326	253	5378
`English` vs `Native`	35.73	2.26×10^{-9}	***	1008	262	141	2589	23.68	1.14×10^{-6}	***	2010	359	239	5392
Qwen2-7B-Instruct														
`English` vs `German`	135.57	2.47×10^{-31}	***	1508	356	105	2031	23.40	1.31×10^{-6}	***	1947	470	332	5251
`English` vs `Chinese`	3.90	4.83×10^{-2}	*	1689	175	215	1921	0.82	3.66×10^{-1}		2077	340	365	5218
`English` vs `Multilingual`	152.19	5.75×10^{-35}	***	1489	375	104	2032	36.32	1.67×10^{-9}	***	1918	499	325	5258
`English` vs `Native`	275.76	6.30×10^{-62}	***	1252	612	152	1984	52.53	4.23×10^{-13}	***	1798	619	388	5195
Qwen2.5-7B-Instruct														
`English` vs `German`	0.00	1.00×10^{0}		1111	130	131	2628	0.39	5.35×10^{-1}		1726	262	247	5765
`English` vs `Chinese`	13.56	2.31×10^{-4}	***	1105	136	205	2554	288.05	1.32×10^{-64}	***	1737	251	803	5209
`English` vs `Multilingual`	3.40	6.53×10^{-2}		1093	148	117	2642	0.04	8.36×10^{-1}		1695	293	287	5725
`English` vs `Native`	17.74	2.54×10^{-5}	***	1030	211	132	2627	42.00	9.12×10^{-11}	***	1624	364	208	5804
NeMo-12B-Instruct														
`English` vs `German`	177.16	2.02×10^{-40}	***	1143	420	112	2325	182.14	1.65×10^{-41}	***	2056	678	263	5003
`English` vs `Chinese`	180.26	4.26×10^{-41}	***	1123	440	121	2316	211.57	6.24×10^{-48}	***	1990	744	278	4988
`English` vs `Multilingual`	217.81	2.71×10^{-49}	***	1072	491	124	2313	286.72	2.58×10^{-64}	***	1995	739	215	5051
`English` vs `Native`	211.34	7.00×10^{-48}	***	1017	546	159	2278	200.47	1.65×10^{-45}	***	1939	795	321	4945
Aya-Expanse-8B														
`English` vs `German`	47.23	6.32×10^{-12}	***	1087	94	216	2603	459.12	7.46×10^{-102}	***	1697	168	854	5281
`English` vs `Chinese`	8.30	3.97×10^{-3}	**	1077	104	151	2668	136.10	1.90×10^{-31}	***	1653	212	531	5604
`English` vs `Multilingual`	0.31	5.80×10^{-1}		1081	100	109	2710	42.27	7.93×10^{-11}	***	1662	203	358	5777
`English` vs `Native`	24.43	7.70×10^{-7}	***	1051	130	224	2595	1.26	2.62×10^{-1}		1649	216	241	5894

Table 18: McNemar's test results of ICL modes on LRL and HRL splits of XQuAD dataset. Baseline is the `English` mode, compared with other `Monolingual`, `Multilingual`, and `Native` modes.

MGSM CIS	bn	de	en	es	fr	ja	ru	sw	te	th	zh	LRL AVG	HRL AVG	ALL AVG
Llama3-8B-Instruct														
English + CIS-En	64.40	75.20	84.80	78.40	74.00	69.20	74.00	56.80	49.20	73.60	72.40	61.00	75.43	70.18
English + CIS-Fr	66.00	77.20	85.20	80.40	76.40	67.20	75.60	54.00	58.40	68.00	70.80	61.60	76.11	70.84
English + CIS-Ja	66.40	75.60	83.20	78.80	74.80	66.80	77.20	53.20	53.60	72.00	70.80	61.30	75.31	70.22
English + CIS-Zh	63.60	76.80	83.20	77.20	76.40	69.20	77.60	52.40	52.40	68.80	71.60	59.30	76.00	69.93
English + CIS-Multi	65.60	75.20	83.20	78.40	77.20	66.40	76.40	51.60	55.60	69.20	70.40	60.50	75.31	69.93
Llama3.1-8B-Instruct														
English + CIS-En	52.40	70.80	88.40	70.80	75.20	70.40	74.00	67.60	38.40	65.20	73.20	55.90	74.69	67.85
English + CIS-Fr	52.80	74.40	89.20	79.60	76.80	69.20	78.80	60.00	37.60	58.00	76.40	52.10	77.77	68.44
English + CIS-Ja	61.20	76.00	87.60	78.40	75.20	70.80	78.00	66.40	41.20	66.40	75.60	58.80	77.37	70.62
English + CIS-Zh	58.00	72.40	89.60	80.80	76.00	66.00	77.20	69.60	38.80	53.60	76.00	55.00	76.86	68.91
English + CIS-Multi	62.00	74.80	88.80	81.60	78.00	69.20	79.60	68.80	46.40	72.80	78.80	62.50	78.69	72.80
Qwen2-7B-Instruct														
English + CIS-En	54.40	74.00	89.60	82.00	78.80	69.60	79.20	24.80	19.20	73.60	81.20	43.00	79.20	66.04
English + CIS-Fr	55.20	75.20	90.00	84.00	72.80	68.00	80.80	26.40	18.00	74.40	82.00	43.50	78.97	66.07
English + CIS-Ja	56.00	75.20	90.40	78.00	76.80	69.20	79.20	24.80	20.40	74.00	80.00	43.80	78.40	65.82
English + CIS-Zh	55.60	74.80	90.80	79.20	76.80	67.60	80.80	21.60	18.40	75.20	80.40	42.70	78.63	65.56
English + CIS-Multi	54.80	76.40	89.60	82.00	76.80	69.20	80.00	24.40	18.00	73.20	81.20	42.60	79.31	65.96
Qwen2.5-7B-Instruct														
English + CIS-En	75.20	86.40	92.40	88.00	87.60	80.80	88.00	31.60	48.00	82.80	85.20	59.40	86.91	76.91
English + CIS-Fr	75.60	85.20	94.00	89.60	86.00	78.80	87.20	31.60	46.40	84.00	82.80	59.40	86.23	76.47
English + CIS-Ja	76.80	86.40	92.40	90.40	85.60	81.20	86.00	29.60	50.40	83.60	82.00	60.10	86.29	76.76
English + CIS-Zh	78.00	88.40	93.60	89.20	88.00	80.40	90.00	29.20	52.40	84.00	82.80	60.90	87.49	77.82
English + CIS-Multi	76.00	86.40	94.00	90.40	84.80	80.80	86.80	30.00	51.60	80.40	82.80	59.50	86.57	76.73
NeMo-12B-Instruct														
English + CIS-En	61.20	80.00	92.00	82.00	82.40	71.60	82.00	48.40	62.40	70.80	78.00	60.70	81.14	73.71
English + CIS-Fr	64.80	83.20	90.40	84.00	82.80	72.00	84.40	46.40	63.20	71.20	78.00	61.40	82.11	74.58
English + CIS-Ja	67.20	82.80	92.40	81.20	83.60	73.60	84.80	43.20	63.20	67.20	76.80	60.20	82.17	74.18
English + CIS-Zh	62.80	82.00	91.60	80.40	83.20	70.40	86.00	44.40	65.60	69.20	80.40	60.50	82.00	74.18
English + CIS-Multi	72.00	81.60	92.00	84.00	83.60	73.60	86.40	47.20	67.20	73.20	79.20	64.90	82.91	76.36
Aya-Expanse-8B														
English + CIS-En	36.00	76.40	82.40	83.20	74.80	69.20	77.60	24.00	15.60	35.60	70.80	27.80	76.34	58.69
English + CIS-Fr	42.00	72.80	83.60	79.20	73.20	72.00	76.80	20.80	17.20	34.80	72.40	28.70	75.71	58.62
English + CIS-Ja	40.80	76.40	82.00	79.60	74.80	68.80	76.00	23.20	18.00	35.20	73.20	29.30	75.83	58.91
English + CIS-Zh	38.40	74.40	82.00	79.60	75.20	72.00	80.00	20.40	18.40	37.20	72.40	28.60	76.51	59.09
English + CIS-Multi	39.60	76.40	81.60	81.60	73.60	71.60	77.60	22.80	18.80	30.40	73.20	27.90	76.51	58.84

Table 19: Accuracies (%) of CIS modes across 11 languages of the MGSM dataset. AVG represents the average accuracy of the language set (LRLs, HRLs or All languages). The underlined languages in the table header are LRLs, otherwise HRLs.

XCOPA CIS	en	et	ht	id	it	qu	sw	ta	th	tr	vi	zh	LRL AVG	HRL AVG	ALL AVG
Llama3-8B-Instruct															
English + CIS-En	95.40	56.60	16.00	80.20	84.80	18.60	39.00	58.40	69.80	70.80	80.20	87.40	48.37	83.72	63.10
English + CIS-Fr	95.40	57.00	47.00	80.40	85.80	42.20	52.40	57.80	72.40	74.20	80.60	86.60	58.49	84.48	69.32
English + CIS-Ja	95.00	57.20	39.60	82.20	86.60	46.40	56.00	57.00	73.20	74.00	81.40	85.60	58.69	84.68	69.52
English + CIS-Zh	95.00	58.40	35.80	82.00	86.00	43.00	56.80	57.00	73.80	73.60	81.40	87.60	58.03	84.84	69.20
English + CIS-Multi	95.60	58.40	51.00	81.60	86.80	49.00	59.20	56.40	73.60	74.60	80.40	88.00	61.14	85.32	71.22
Llama3.1-8B-Instruct															
English + CIS-En	96.00	64.00	25.00	84.80	88.60	27.60	51.60	62.20	72.80	76.80	85.00	90.00	55.46	87.24	68.70
English + CIS-Fr	95.20	65.80	23.00	85.00	88.20	42.80	53.20	69.80	76.00	78.00	85.20	89.20	59.40	87.12	70.95
English + CIS-Ja	96.40	65.80	53.60	86.60	89.00	46.60	61.40	68.60	77.00	79.20	86.60	88.40	65.66	87.92	74.93
English + CIS-Zh	95.80	65.40	52.00	86.80	88.80	47.40	58.40	68.40	76.00	78.40	85.00	89.00	64.66	87.76	74.28
English + CIS-Multi	96.40	63.20	51.20	86.80	89.80	48.80	59.80	69.20	75.60	80.80	85.40	89.40	64.74	88.64	74.70
Qwen2-7B-Instruct															
English + CIS-En	97.00	62.00	51.40	87.80	90.20	51.00	52.40	56.40	79.00	76.20	84.00	90.20	62.31	88.28	73.13
English + CIS-Fr	97.00	62.60	49.80	87.20	90.20	50.20	54.00	57.20	79.20	74.40	83.20	90.60	62.31	87.88	72.97
English + CIS-Ja	97.00	62.60	51.40	88.40	89.60	50.80	54.40	58.40	80.20	75.00	83.80	91.80	63.09	88.36	73.62
English + CIS-Zh	97.20	60.80	51.00	86.80	88.80	50.60	53.60	57.80	80.20	76.20	83.40	92.80	62.49	88.36	73.27
English + CIS-Multi	97.20	62.20	52.20	87.80	90.80	51.80	52.20	56.80	78.80	74.60	83.60	90.40	62.51	88.16	73.20
Qwen2.5-7B-Instruct															
English + CIS-En	97.00	63.00	56.60	89.40	91.00	50.20	52.40	56.00	83.40	78.20	87.80	94.20	64.20	89.96	74.93
English + CIS-Fr	96.40	63.60	58.80	90.00	93.00	49.00	51.00	56.80	82.20	79.20	87.40	94.00	64.11	90.52	75.12
English + CIS-Ja	96.60	63.60	60.00	90.00	92.60	50.00	48.80	56.00	82.80	78.80	88.20	93.80	64.20	90.36	75.10
English + CIS-Zh	97.20	64.00	59.80	89.80	92.60	50.20	52.00	58.00	84.00	79.20	88.40	93.60	65.20	90.48	75.73
English + CIS-Multi	96.60	64.60	61.20	90.20	93.00	52.00	48.40	59.20	82.20	80.00	87.80	93.60	65.06	90.68	75.73
NeMo-12B-Instruct															
English + CIS-En	96.40	54.40	55.20	79.80	90.80	49.60	54.20	69.80	63.80	70.40	80.00	90.00	61.00	85.48	71.20
English + CIS-Fr	96.00	55.80	57.40	80.20	90.00	49.40	53.80	71.20	61.20	69.60	79.80	88.60	61.23	84.88	71.08
English + CIS-Ja	95.60	58.40	54.60	82.40	91.20	51.20	54.40	72.00	61.40	69.20	80.60	88.20	61.80	85.32	71.60
English + CIS-Zh	95.60	56.40	53.20	82.40	91.40	50.40	54.80	72.20	61.00	71.00	80.00	90.00	61.14	86.08	71.53
English + CIS-Multi	96.20	57.20	56.20	83.40	92.20	50.60	55.40	73.40	61.80	72.80	81.00	90.20	62.23	86.96	72.53
Aya-Expanse-8B															
English + CIS-En	93.80	16.40	8.60	82.80	85.20	1.80	3.20	40.00	17.60	81.40	74.80	74.80	23.20	83.60	48.37
English + CIS-Fr	94.20	27.80	13.00	84.00	87.80	0.60	5.00	47.00	28.20	81.60	79.40	79.40	28.71	85.40	52.33
English + CIS-Ja	94.00	31.60	12.80	85.00	87.40	0.80	5.40	62.00	40.60	81.40	78.20	83.60	33.06	86.28	55.23
English + CIS-Zh	94.80	28.60	12.00	84.80	87.20	0.40	5.40	43.60	21.80	81.60	76.80	83.60	26.94	86.40	51.72
English + CIS-Multi	94.60	36.40	18.80	86.40	89.00	0.80	12.80	49.80	37.00	82.80	79.60	83.60	33.60	87.28	55.97

Table 20: Accuracies (%) of CIS modes across 12 languages of the XCOPA dataset. AVG represents the average accuracy of the language set (LRLs, HRLs or All languages). The underlined languages in the table header are LRLs, otherwise HRLs.

XL-WiC CIS	bg	da	de	en	et	fa	fr	hr	it	ja	ko	nl	zh	LRL AVG	HRL AVG	ALL AVG
Llama3-8B-Instruct																
English + CIS-En	55.13	65.90	61.54	65.64	53.85	64.36	59.23	53.59	52.05	52.82	54.62	56.92	63.59	56.73	59.15	58.40
English + CIS-Fr	56.15	67.18	63.08	64.62	54.62	65.13	60.00	54.10	53.85	52.82	55.64	60.00	60.26	57.50	59.72	59.03
English + CIS-Ja	56.67	67.95	62.05	64.87	54.10	66.67	60.77	55.13	56.67	53.85	53.85	59.23	62.05	58.14	60.14	59.53
English + CIS-Zh	57.44	68.72	63.33	66.15	54.36	65.64	58.97	55.64	54.10	53.33	54.87	60.26	58.97	58.27	59.86	59.37
English + CIS-Multi	56.67	65.90	62.31	64.36	54.62	65.90	60.77	54.10	54.62	51.79	52.31	59.49	61.03	57.82	59.17	58.76
Llama3.1-8B-Instruct																
English + CIS-En	53.08	52.05	62.05	64.62	41.79	44.36	54.36	52.31	31.03	48.46	52.56	54.36	57.95	47.88	53.05	51.46
English + CIS-Fr	55.13	58.97	63.08	64.87	53.08	47.95	60.77	55.13	53.59	50.77	55.38	60.51	58.97	52.82	58.55	56.79
English + CIS-Ja	55.13	60.77	64.87	65.13	53.33	61.79	59.74	53.59	54.62	50.77	53.85	62.05	54.87	55.96	58.52	57.73
English + CIS-Zh	54.62	58.97	64.36	65.64	53.59	58.46	61.28	52.05	55.13	51.28	54.36	61.28	57.95	54.68	58.92	57.61
English + CIS-Multi	55.38	62.56	62.56	64.62	52.31	61.03	61.28	55.38	52.82	51.79	51.03	64.10	58.46	56.03	58.80	57.95
Qwen2-7B-Instruct																
English + CIS-En	39.74	40.26	61.54	66.15	55.38	54.36	61.28	54.10	37.95	8.21	13.85	55.90	0.26	50.90	38.38	42.23
English + CIS-Fr	53.33	43.59	63.85	65.90	52.05	55.64	62.05	53.59	40.26	18.21	15.90	56.41	3.33	53.65	41.05	44.93
English + CIS-Ja	56.41	50.77	62.05	65.64	53.33	61.28	60.26	53.85	43.85	48.46	34.10	60.26	15.13	56.22	48.95	51.18
English + CIS-Zh	56.41	50.00	62.56	65.38	54.87	61.79	61.03	54.10	43.59	37.69	30.77	59.49	14.10	56.79	47.18	50.14
English + CIS-Multi	55.13	46.67	62.82	64.10	52.05	59.49	61.03	53.08	40.26	26.67	20.00	55.64	6.92	54.94	42.68	46.45
Qwen2.5-7B-Instruct																
English + CIS-En	59.49	61.79	72.05	73.33	57.44	60.51	66.67	58.72	60.77	52.56	64.87	74.62	63.59	59.04	65.58	63.57
English + CIS-Fr	55.90	62.56	71.03	72.05	57.18	62.31	64.36	60.77	59.49	57.95	66.67	71.28	64.10	59.04	65.50	63.51
English + CIS-Ja	57.95	62.31	72.05	71.54	58.46	61.28	64.10	59.74	58.72	65.13	64.62	73.59	65.64	59.36	66.41	64.24
English + CIS-Zh	56.92	61.54	72.56	71.28	57.44	61.54	64.36	61.03	58.72	65.13	65.13	71.79	65.64	59.23	66.24	64.08
English + CIS-Multi	57.18	62.31	71.03	72.82	57.95	63.59	65.13	58.72	57.95	64.10	64.87	73.08	65.90	59.36	66.35	64.20
NeMo-12B-Instruct																
English + CIS-En	48.97	62.56	69.74	66.67	51.79	47.95	58.21	49.74	34.10	38.46	52.56	67.44	53.85	49.62	55.95	54.00
English + CIS-Fr	48.72	62.31	69.74	65.90	51.79	48.97	59.23	52.82	42.56	52.82	51.79	62.82	55.13	50.58	58.03	55.74
English + CIS-Ja	48.21	62.05	71.28	66.15	51.03	47.95	58.72	52.31	37.18	57.69	53.85	65.13	54.62	49.87	58.52	55.86
English + CIS-Zh	48.46	60.00	69.49	65.64	51.54	49.74	58.46	50.51	37.18	54.36	53.85	65.13	57.18	50.06	57.92	55.50
English + CIS-Multi	48.46	58.97	69.74	66.15	52.31	47.95	61.28	52.05	39.23	54.10	54.36	65.90	56.15	50.19	58.43	55.90
Aya-Expanse-8B																
English + CIS-En	52.31	55.64	57.44	65.13	56.92	64.62	54.10	57.44	19.49	49.74	54.62	55.38	57.95	57.82	52.17	53.91
English + CIS-Fr	56.92	60.00	64.10	65.13	59.49	73.59	59.74	56.15	38.97	63.33	64.62	63.59	59.74	61.54	59.91	60.41
English + CIS-Ja	55.90	58.21	62.82	65.38	61.28	71.54	54.62	55.38	28.21	64.36	64.10	61.03	58.46	61.03	57.46	58.56
English + CIS-Zh	56.15	57.95	63.08	65.13	60.51	70.77	54.36	54.87	29.23	63.08	64.62	61.03	60.51	60.58	57.66	58.56
English + CIS-Multi	57.95	60.77	64.36	65.64	59.49	74.36	56.41	57.18	39.49	62.31	65.64	64.87	58.21	62.24	59.74	60.51

Table 21: Accuracies (%) of CIS modes across 13 languages of the XL-WiC dataset. AVG represents the average accuracy of the language set (LRLs, HRLs or All languages). The underlined languages in the table header are LRLs, otherwise HRLs.

XQuAD CIS	ar	de	<u>el</u>	en	es	<u>hi</u>	<u>ro</u>	ru	<u>th</u>	tr	vi	zh	<u>LRL AVG</u>	HRL AVG	ALL AVG
Llama3-8B-Instruct															
English + CIS-En	60.70	75.00	68.30	86.70	79.30	68.70	79.10	62.20	69.60	72.40	74.40	77.80	71.43	73.56	72.85
English + CIS-Fr	58.80	75.30	68.30	86.70	79.50	68.00	79.00	62.00	68.60	71.30	73.60	76.70	70.98	72.99	72.32
English + CIS-Ja	58.60	75.00	67.70	86.80	79.70	66.90	79.20	61.70	69.60	71.50	73.00	77.80	70.85	73.01	72.29
English + CIS-Zh	58.30	75.30	68.40	87.00	79.40	68.20	78.80	61.90	69.70	71.40	73.20	77.90	71.28	73.05	72.46
English + CIS-Multi	58.50	75.30	68.70	87.30	79.80	66.40	79.20	62.30	69.30	71.60	74.10	76.70	70.90	73.20	72.43
Llama3.1-8B-Instruct															
English + CIS-En	60.30	73.20	65.50	84.40	78.20	67.10	74.60	54.70	66.60	67.10	70.90	74.20	68.45	70.38	69.73
English + CIS-Fr	61.00	73.10	65.30	85.20	77.80	67.40	74.50	55.70	68.60	65.80	71.90	73.40	68.95	70.49	69.98
English + CIS-Ja	60.80	72.60	65.00	84.30	77.30	67.80	74.80	55.30	68.00	66.70	70.70	74.70	68.90	70.30	69.83
English + CIS-Zh	61.10	73.00	65.90	84.70	77.30	67.70	74.60	55.80	68.20	66.20	71.30	74.40	69.10	70.48	70.02
English + CIS-Multi	61.00	72.80	66.30	84.80	77.90	67.90	75.00	55.60	69.20	66.10	71.40	73.20	69.60	70.35	70.10
Qwen2-7B-Instruct															
English + CIS-En	61.00	70.50	47.70	80.90	77.20	54.50	72.30	57.30	59.70	63.00	73.70	81.80	58.55	70.68	66.63
English + CIS-Fr	60.30	71.10	50.80	81.40	77.00	55.70	73.00	59.50	64.10	64.10	73.10	81.80	60.90	71.04	67.66
English + CIS-Ja	60.80	69.70	50.10	82.40	77.50	55.10	72.10	58.90	65.00	63.80	74.10	81.50	60.58	71.09	67.58
English + CIS-Zh	60.40	69.40	47.80	81.20	75.90	54.30	71.50	57.90	62.20	62.40	72.90	81.70	58.95	70.22	66.47
English + CIS-Multi	59.80	70.10	50.00	82.00	77.80	54.40	72.30	59.80	64.60	64.50	73.50	81.40	60.32	71.11	67.52
Qwen2.5-7B-Instruct															
English + CIS-En	64.20	75.60	56.90	85.60	80.50	62.60	80.30	66.20	76.40	70.60	75.40	85.20	69.05	75.41	73.29
English + CIS-Fr	62.80	74.90	57.30	86.10	80.60	62.60	80.20	66.40	76.40	70.10	75.60	85.10	69.12	75.20	73.17
English + CIS-Ja	58.40	73.40	57.70	85.10	77.60	62.20	79.90	64.10	74.80	69.80	71.50	85.70	68.65	73.20	71.68
English + CIS-Zh	60.40	73.90	56.50	85.40	77.80	61.90	78.80	64.30	75.80	69.70	70.50	85.70	68.25	73.46	71.72
English + CIS-Multi	62.80	74.60	56.90	85.90	80.10	61.80	79.80	66.50	76.00	69.40	74.90	85.50	68.62	74.96	72.85
NeMo-12B-Instruct															
English + CIS-En	54.40	69.60	46.70	82.90	76.70	44.30	71.40	55.20	63.50	61.70	64.10	49.50	56.48	64.26	61.67
English + CIS-Fr	47.80	63.00	41.90	83.50	75.50	37.20	66.30	43.40	59.20	53.90	53.40	37.20	51.15	57.21	55.19
English + CIS-Ja	46.20	63.30	40.70	83.00	75.60	33.40	66.10	45.40	59.40	55.20	54.60	34.90	49.90	57.28	54.82
English + CIS-Zh	46.30	62.70	42.50	83.10	75.10	33.00	66.50	41.70	58.20	55.10	54.40	32.70	50.05	56.39	54.28
English + CIS-Multi	47.00	63.50	42.00	83.30	75.00	37.60	67.90	43.70	59.90	56.00	54.10	39.90	51.85	57.81	55.82
Aya-Expanse-8B															
English + CIS-En	72.00	77.20	73.80	88.30	84.10	70.40	82.10	64.30	54.90	68.50	79.70	81.40	70.30	76.94	74.73
English + CIS-Fr	72.10	75.60	72.40	88.50	84.30	70.30	81.30	63.70	54.70	66.80	78.60	78.60	69.68	76.02	73.91
English + CIS-Ja	71.40	75.90	72.70	88.20	83.80	69.90	82.20	64.40	55.20	68.80	78.50	77.30	70.00	76.04	74.02
English + CIS-Zh	71.10	76.40	73.10	88.70	83.90	70.40	82.50	64.80	55.00	68.70	79.00	81.70	70.25	76.79	74.61
English + CIS-Multi	71.70	76.80	73.00	88.50	83.80	70.60	82.30	64.00	54.50	67.70	79.60	80.40	70.10	76.56	74.41

Table 22: Accuracies (%) of CIS modes across 12 languages of the XQuAD dataset. AVG represents the average accuracy of the language set (LRLs, HRLs or All languages). The <u>underlined languages</u> in the table header are <u>LRLs</u>, otherwise HRLs.

McNemar's Test MGSM CIS Mode M1 vs M2 Comparison	Low-Resource Languages							High-Resource Languages						
	χ^2	p-value	Sig. Level	#Both ✖	#M1 ✖ M2 ✔	#M1 ✔ M2 ✖	#Both ✔	χ^2	p-value	Sig. Level	#Both ✖	#M1 ✖ M2 ✔	#M1 ✔ M2 ✖	#Both ✔
Llama3-8B-Instruct														
`English + CIS-En vs English + CIS-Zh`	1.57	2.10E-01		317	73	90	520	0.47	4.93E-01		339	91	81	1239
`English + CIS-En vs English + CIS-Fr`	0.16	6.85E-01		311	79	73	537	0.60	4.39E-01		323	107	95	1225
`English + CIS-En vs English + CIS-Ja`	0.03	8.72E-01		311	79	76	534	0.01	9.40E-01		343	87	89	1231
`English + CIS-En vs English + CIS-Multi`	0.08	7.72E-01		297	93	98	512	0.01	9.41E-01		341	89	91	1229
`Multilingual + CIS-Multi vs English + CIS-Multi`	0.06	8.00E-01		327	72	68	533	0.16	6.93E-01		355	83	77	1235
Llama3.1-8B-Instruct														
`English + CIS-En vs English + CIS-Zh`	0.27	6.02E-01		328	113	122	437	5.35	2.08E-02	*	296	147	109	1198
`English + CIS-En vs English + CIS-Fr`	5.90	1.51E-02	*	344	97	135	424	10.48	1.21E-03	**	282	161	107	1200
`English + CIS-En vs English + CIS-Ja`	3.79	5.16E-02		323	118	89	470	7.69	5.54E-03	**	282	161	114	1193
`English + CIS-En vs English + CIS-Multi`	19.20	1.17E-05	***	298	143	77	482	18.31	1.88E-05	***	278	165	95	1212
`Multilingual + CIS-Multi vs English + CIS-Multi`	18.27	1.91E-05	***	246	68	129	557	0.07	7.97E-01		250	118	123	1259
Qwen2-7B-Instruct														
`English + CIS-En vs English + CIS-Zh`	0.03	8.69E-01		498	72	75	355	0.50	4.80E-01		288	76	86	1300
`English + CIS-En vs English + CIS-Fr`	0.12	7.27E-01		502	68	63	367	0.06	8.06E-01		291	73	77	1309
`English + CIS-En vs English + CIS-Ja`	0.37	5.42E-01		500	70	62	368	0.96	3.27E-01		283	81	95	1291
`English + CIS-En vs English + CIS-Multi`	0.07	7.92E-01		507	63	67	363	0.01	9.37E-01		284	80	78	1308
`Multilingual + CIS-Multi vs English + CIS-Multi`	12.23	4.70E-04	***	464	63	110	363	0.35	5.54E-01		266	87	96	1301
Qwen2.5-7B-Instruct														
`English + CIS-En vs English + CIS-Zh`	2.06	1.51E-01		351	55	40	554	0.86	3.53E-01		177	52	42	1479
`English + CIS-En vs English + CIS-Fr`	0.01	9.23E-01		352	54	54	540	1.41	2.36E-01		192	37	49	1472
`English + CIS-En vs English + CIS-Ja`	0.34	5.62E-01		349	57	50	544	0.97	3.24E-01		183	46	57	1464
`English + CIS-En vs English + CIS-Multi`	0.00	1.00E+00		347	59	58	536	0.27	6.06E-01		185	44	50	1471
`Multilingual + CIS-Multi vs English + CIS-Multi`	0.03	8.55E-01		347	61	58	534	0.01	9.28E-01		174	63	61	1452
NeMo-12B-Instruct														
`English + CIS-En vs English + CIS-Zh`	0.01	9.42E-01		300	93	95	512	1.30	2.55E-01		247	83	68	1352
`English + CIS-En vs English + CIS-Fr`	0.20	6.57E-01		298	95	88	519	1.46	2.26E-01		234	96	79	1341
`English + CIS-En vs English + CIS-Ja`	0.08	7.80E-01		293	100	105	502	1.66	1.97E-01		234	96	78	1342
`English + CIS-En vs English + CIS-Multi`	9.66	1.88E-03	**	285	108	66	541	5.96	1.46E-02	*	239	91	60	1360
`Multilingual + CIS-Multi vs English + CIS-Multi`	2.31	1.28E-01		283	88	68	561	12.66	3.74E-04	***	242	103	57	1348
Aya-Expanse-8B														
`English + CIS-En vs English + CIS-Zh`	0.34	5.60E-01		646	76	68	210	0.02	8.78E-01		327	87	84	1252
`English + CIS-En vs English + CIS-Fr`	0.47	4.91E-01		650	72	63	215	0.56	4.55E-01		330	84	95	1241
`English + CIS-En vs English + CIS-Ja`	1.22	2.70E-01		634	88	73	205	0.35	5.52E-01		328	86	95	1241
`English + CIS-En vs English + CIS-Multi`	0.00	1.00E+00		662	60	59	219	0.02	8.77E-01		329	85	82	1254
`Multilingual + CIS-Multi vs English + CIS-Multi`	6.88	8.71E-03	**	614	71	107	208	0.68	4.09E-01		294	131	117	1208

Table 23: McNemar's test results of CIS modes comparison on LRL and HRL splits of MGSM dataset across 6 MLLMs. Sig. – Significance. ✖– the model response is incorrect; ✔– correct.

McNemar's Test XCOPA CIS Mode M1 vs M2 Comparison	Low-Resource Languages							High-Resource Languages						
	χ^2	p-value	Sig. Level	#Both ✖	#M1 ✖ M2 ✔	#M1 ✔ M2 ✖	#Both ✔	χ^2	p-value	Sig. Level	#Both ✖	#M1 ✖ M2 ✔	#M1 ✔ M2 ✖	#Both ✔
Llama3-8B-Instruct														
English + CIS-En vs English + CIS-Zh	199.24	3.05E-45	***	1353	454	116	1577	8.28	4.00E-03	**	349	58	30	2063
English + CIS-En vs English + CIS-Fr	214.84	1.21E-48	***	1340	467	113	1580	3.34	6.76E-02		349	58	39	2054
English + CIS-En vs English + CIS-Ja	209.37	1.89E-47	***	1317	490	129	1564	4.90	2.69E-02	*	341	66	42	2051
English + CIS-En vs English + CIS-Multi	278.98	1.25E-62	***	1227	580	133	1560	13.34	2.60E-04	***	330	77	37	2056
Multilingual + CIS-Multi vs English + CIS-Multi	0.33	5.64E-01		1099	247	261	1893	0.65	4.19E-01		285	71	82	2062
Llama3.1-8B-Instruct														
English + CIS-En vs English + CIS-Zh	194.42	3.45E-44	***	1133	426	104	1837	1.55	2.13E-01		266	53	40	2141
English + CIS-En vs English + CIS-Fr	39.60	3.12E-10	***	1253	306	168	1773	0.04	8.45E-01		268	51	54	2127
English + CIS-En vs English + CIS-Ja	206.75	7.04E-47	***	1074	485	128	1813	2.44	1.18E-01		258	61	44	2137
English + CIS-En vs English + CIS-Multi	164.28	1.31E-37	***	1077	482	157	1784	10.41	1.25E-03	**	246	73	38	2143
Multilingual + CIS-Multi vs English + CIS-Multi	4.95	2.61E-02	*	956	227	278	2039	0.06	8.13E-01		206	82	78	2134
Qwen2-7B-Instruct														
English + CIS-En vs English + CIS-Zh	0.13	7.15E-01		1222	97	91	2090	0.01	9.07E-01		255	38	36	2171
English + CIS-En vs English + CIS-Fr	0.00	9.48E-01		1200	119	119	2062	1.35	2.45E-01		268	25	35	2172
English + CIS-En vs English + CIS-Ja	2.78	9.53E-02		1184	135	108	2073	0.01	9.09E-01		254	39	37	2170
English + CIS-En vs English + CIS-Multi	0.14	7.09E-01		1186	133	126	2055	0.05	8.15E-01		258	35	38	2169
Multilingual + CIS-Multi vs English + CIS-Multi	2.96	8.55E-02		1099	178	213	2010	9.60	1.95E-03	**	210	49	86	2155
Qwen2.5-7B-Instruct														
English + CIS-En vs English + CIS-Zh	4.03	4.48E-02	*	1092	161	126	2121	2.53	1.12E-01		216	35	22	2227
English + CIS-En vs English + CIS-Fr	0.01	9.10E-01		1098	155	158	2089	1.84	1.75E-01		198	53	39	2210
English + CIS-En vs English + CIS-Ja	0.00	9.57E-01		1078	175	175	2072	0.86	3.53E-01		199	52	42	2207
English + CIS-En vs English + CIS-Multi	2.06	1.51E-01		1034	219	189	2058	3.14	7.63E-02		196	55	37	2212
Multilingual + CIS-Multi vs English + CIS-Multi	2.08	1.49E-01		1008	247	215	2030	0.92	3.38E-01		161	60	72	2207
NeMo-12B-Instruct														
English + CIS-En vs English + CIS-Zh	0.05	8.24E-01		1201	164	159	1976	1.87	1.72E-01		303	60	45	2092
English + CIS-En vs English + CIS-Fr	0.13	7.15E-01		1177	188	180	1955	1.98	1.59E-01		321	42	57	2080
English + CIS-En vs English + CIS-Ja	1.78	1.82E-01		1146	219	191	1944	0.08	7.84E-01		305	58	62	2075
English + CIS-En vs English + CIS-Multi	4.27	3.88E-02	*	1137	228	185	1950	10.54	1.17E-03	**	283	80	43	2094
Multilingual + CIS-Multi vs English + CIS-Multi	0.12	7.24E-01		1069	262	253	1916	0.88	3.47E-01		238	75	88	2099
Aya-Expanse-8B														
English + CIS-En vs English + CIS-Zh	70.71	4.14E-17	***	2503	185	54	758	36.62	1.43E-09	***	310	100	30	2060
English + CIS-En vs English + CIS-Fr	124.96	5.19E-29	***	2444	244	51	761	15.24	9.45E-05	***	324	86	41	2049
English + CIS-En vs English + CIS-Ja	275.84	6.05E-62	***	2301	387	42	770	34.30	4.73E-09	***	313	97	30	2060
English + CIS-En vs English + CIS-Multi	298.12	8.46E-67	***	2285	403	39	773	57.51	3.37E-14	***	292	118	26	2064
Multilingual + CIS-Multi vs English + CIS-Multi	294.90	4.26E-66	***	1795	98	529	1078	9.14	2.50E-03	**	233	49	85	2133

Table 24: McNemar's test results of CIS modes comparison on LRL and HRL splits of XCOPA dataset across 6 MLLMs. Sig. – Significance. ✖– the model response is incorrect; ✔– correct.

McNemar's Test XL-WiC CIS Mode M1 vs M2 Comparison	Low-Resource Languages χ^2	p-value	Sig. Level	#Both ✖	#M1 ✖ M2 ✔	#M1 ✔ M2 ✖	#Both ✔	High-Resource Languages χ^2	p-value	Sig. Level	#Both ✖	#M1 ✖ M2 ✔	#M1 ✔ M2 ✖	#Both ✔
Llama3-8B-Instruct														
`English` + `CIS-En` vs `English` + `CIS-Zh`	3.31	6.90E-02		583	92	68	817	1.69	1.94E-01		1251	183	158	1918
`English` + `CIS-En` vs `English` + `CIS-Fr`	0.81	3.69E-01		594	81	69	816	1.16	2.81E-01		1269	165	145	1931
`English` + `CIS-En` vs `English` + `CIS-Ja`	2.25	1.34E-01		566	109	87	798	2.96	8.55E-02		1221	213	178	1898
`English` + `CIS-En` vs `English` + `CIS-Multi`	1.29	2.57E-01		567	108	91	794	0.00	1.00E+00		1226	208	207	1869
`Multilingual` + `CIS-Multi` vs `English` + `CIS-Multi`	0.13	7.20E-01		464	186	194	716	4.03	4.46E-02	*	1044	334	389	1743
Llama3.1-8B-Instruct														
`English` + `CIS-En` vs `English` + `CIS-Zh`	34.03	5.43E-09	***	598	215	109	638	67.56	2.04E-16	***	1234	414	208	1654
`English` + `CIS-En` vs `English` + `CIS-Fr`	16.94	3.86E-05	***	604	209	132	615	72.71	1.50E-17	***	1298	350	157	1705
`English` + `CIS-En` vs `English` + `CIS-Ja`	47.64	5.13E-12	***	586	227	101	646	52.87	3.56E-13	***	1207	441	249	1613
`English` + `CIS-En` vs `English` + `CIS-Multi`	45.23	1.75E-11	***	574	239	112	635	53.72	2.31E-13	***	1171	477	275	1587
`Multilingual` + `CIS-Multi` vs `English` + `CIS-Multi`	0.97	3.25E-01		509	158	177	716	0.83	3.63E-01		992	426	454	1638
Qwen2-7B-Instruct														
`English` + `CIS-En` vs `English` + `CIS-Zh`	26.89	2.16E-07	***	566	200	108	686	200.56	1.58E-45	***	1772	391	82	1265
`English` + `CIS-En` vs `English` + `CIS-Fr`	8.28	4.00E-03	**	638	128	85	709	32.76	1.04E-08	***	1984	179	85	1262
`English` + `CIS-En` vs `English` + `CIS-Ja`	19.38	1.07E-05	***	551	215	132	662	242.30	1.24E-54	***	1695	468	97	1250
`English` + `CIS-En` vs `English` + `CIS-Multi`	14.08	1.75E-04	***	598	168	105	689	61.98	3.46E-15	***	1906	257	106	1241
`Multilingual` + `CIS-Multi` vs `English` + `CIS-Multi`	0.34	5.61E-01		446	243	257	614	392.40	2.48E-87	***	1089	245	923	1253
Qwen2.5-7B-Instruct														
`English` + `CIS-En` vs `English` + `CIS-Zh`	0.03	8.64E-01		569	70	67	854	1.77	1.83E-01		1060	148	125	2177
`English` + `CIS-En` vs `English` + `CIS-Fr`	0.01	9.36E-01		561	78	78	843	0.01	9.07E-01		1063	145	148	2154
`English` + `CIS-En` vs `English` + `CIS-Ja`	0.10	7.48E-01		559	80	75	846	2.30	1.29E-01		1023	185	156	2146
`English` + `CIS-En` vs `English` + `CIS-Multi`	0.09	7.70E-01		543	96	91	830	2.01	1.57E-01		1026	182	155	2147
`Multilingual` + `CIS-Multi` vs `English` + `CIS-Multi`	4.75	2.92E-02	*	494	180	140	746	0.24	6.27E-01		883	311	298	2018
NeMo-12B-Instruct														
`English` + `CIS-En` vs `English` + `CIS-Zh`	0.71	4.01E-01		757	29	22	752	16.69	4.39E-05	***	1373	173	104	1860
`English` + `CIS-En` vs `English` + `CIS-Fr`	4.17	4.11E-02	*	755	31	16	758	16.67	4.45E-05	***	1354	192	119	1845
`English` + `CIS-En` vs `English` + `CIS-Ja`	0.20	6.51E-01		762	24	20	754	26.40	2.77E-07	***	1351	195	105	1859
`English` + `CIS-En` vs `English` + `CIS-Multi`	0.93	3.36E-01		747	39	30	744	23.18	1.47E-06	***	1343	203	116	1848
`Multilingual` + `CIS-Multi` vs `English` + `CIS-Multi`	10.28	1.35E-03	**	683	54	94	729	9.39	2.18E-03	**	1130	254	329	1797
Aya-Expanse-8B														
`English` + `CIS-En` vs `English` + `CIS-Zh`	7.64	5.72E-03	**	521	137	94	808	102.12	5.24E-24	***	1402	277	84	1747
`English` + `CIS-En` vs `English` + `CIS-Fr`	10.90	9.60E-04	***	480	178	120	782	146.30	1.12E-33	***	1292	387	115	1716
`English` + `CIS-En` vs `English` + `CIS-Ja`	9.03	2.66E-03	**	500	158	108	794	94.02	3.12E-22	***	1404	275	89	1742
`English` + `CIS-En` vs `English` + `CIS-Multi`	14.59	1.34E-04	***	465	193	124	778	143.32	5.01E-33	***	1301	378	112	1719
`Multilingual` + `CIS-Multi` vs `English` + `CIS-Multi`	0.80	3.72E-01		417	190	172	781	43.63	3.96E-11	***	965	270	448	1827

Table 25: McNemar's test results of CIS modes comparison on LRL and HRL splits of XL-WiC dataset across 6 MLLMs. Sig. – Significance. ✖– the model response is incorrect; ✔– correct.

McNemar's Test XQuAD CIS Mode M1 vs M2 Comparison	Low-Resource Languages							High-Resource Languages						
	χ^2	p-value	Sig. Level	#Both ✖	#M1 ✖ M2 ✔	#M1 ✔ M2 ✖	#Both ✔	χ^2	p-value	Sig. Level	#Both ✖	#M1 ✖ M2 ✔	#M1 ✔ M2 ✖	#Both ✔
Llama3-8B-Instruct														
`English + CIS-En vs English + CIS-Zh`	0.19	6.66E-01		1079	64	70	2787	6.53	1.06E-02	*	2013	102	143	5742
`English + CIS-En vs English + CIS-Fr`	2.37	1.24E-01		1091	52	70	2787	8.80	3.01E-03	**	2023	92	138	5747
`English + CIS-En vs English + CIS-Ja`	3.21	7.34E-02		1079	64	87	2770	7.06	7.89E-03	**	2006	109	153	5732
`English + CIS-En vs English + CIS-Multi`	2.55	1.10E-01		1075	68	89	2768	3.23	7.25E-02		2008	107	136	5749
`Multilingual + CIS-Multi vs English + CIS-Multi`	51.05	9.01E-13	***	945	92	219	2744	65.38	6.18E-16	***	1769	183	375	5673
Llama3.1-8B-Instruct														
`English + CIS-En vs English + CIS-Zh`	3.47	6.24E-02		1159	103	77	2661	0.15	6.98E-01		2203	167	159	5471
`English + CIS-En vs English + CIS-Fr`	2.44	1.18E-01		1178	84	64	2674	0.25	6.20E-01		2235	135	126	5504
`English + CIS-En vs English + CIS-Ja`	1.54	2.15E-01		1159	103	85	2653	0.07	7.86E-01		2204	166	172	5458
`English + CIS-En vs English + CIS-Multi`	11.13	8.51E-04	***	1148	114	68	2670	0.00	9.55E-01		2216	154	156	5474
`Multilingual + CIS-Multi vs English + CIS-Multi`	15.06	1.04E-04	***	1066	89	150	2695	18.72	1.51E-05	***	2083	193	289	5435
Qwen2-7B-Instruct														
`English + CIS-En vs English + CIS-Zh`	1.14	2.86E-01		1551	107	91	2251	3.83	5.04E-02		2204	142	178	5476
`English + CIS-En vs English + CIS-Fr`	40.04	2.49E-10	***	1503	155	61	2281	2.37	1.24E-01		2166	180	151	5503
`English + CIS-En vs English + CIS-Ja`	25.50	4.43E-07	***	1492	166	85	2257	2.76	9.66E-02		2144	202	169	5485
`English + CIS-En vs English + CIS-Multi`	23.90	1.01E-06	***	1520	138	67	2275	3.31	6.88E-02		2154	192	157	5497
`Multilingual + CIS-Multi vs English + CIS-Multi`	13.06	3.02E-04	***	1403	120	184	2293	8.32	3.91E-03	**	1990	251	321	5438
Qwen2.5-7B-Instruct														
`English + CIS-En vs English + CIS-Zh`	6.08	1.37E-02	*	1175	63	95	2667	54.60	1.48E-13	***	1825	142	298	5735
`English + CIS-En vs English + CIS-Fr`	0.03	8.55E-01		1177	61	58	2704	1.00	3.16E-01		1848	119	136	5897
`English + CIS-En vs English + CIS-Ja`	1.29	2.55E-01		1159	79	95	2667	63.09	1.98E-15	***	1810	157	334	5699
`English + CIS-En vs English + CIS-Multi`	1.77	1.84E-01		1174	64	81	2681	4.06	4.40E-02	*	1834	133	169	5864
`Multilingual + CIS-Multi vs English + CIS-Multi`	6.34	1.18E-02	*	1115	100	140	2645	0.05	8.31E-01		1724	273	279	5724
NeMo-12B-Instruct														
`English + CIS-En vs English + CIS-Zh`	175.70	4.21E-40	***	1683	58	315	1944	462.20	1.60E-102	***	2746	113	743	4398
`English + CIS-En vs English + CIS-Fr`	125.89	3.24E-29	***	1669	72	285	1974	380.06	1.21E-84	***	2724	135	699	4442
`English + CIS-En vs English + CIS-Ja`	175.56	4.52E-40	***	1677	64	327	1932	380.18	1.14E-84	***	2729	130	689	4452
`English + CIS-En vs English + CIS-Multi`	98.13	3.91E-23	***	1661	80	265	1994	329.88	1.02E-73	***	2715	144	660	4481
`Multilingual + CIS-Multi vs English + CIS-Multi`	562.57	2.32E-124	***	1140	85	786	1989	871.12	1.86E-191	***	2073	169	1302	4456
Aya-Expanse-8B														
`English + CIS-En vs English + CIS-Zh`	0.01	9.17E-01		1143	45	47	2765	0.78	3.78E-01		1773	72	84	6071
`English + CIS-En vs English + CIS-Fr`	5.49	1.92E-02	*	1148	40	65	2747	20.99	4.62E-06	***	1758	87	160	5995
`English + CIS-En vs English + CIS-Ja`	1.12	2.90E-01		1140	48	60	2752	18.81	1.44E-05	***	1747	98	170	5985
`English + CIS-En vs English + CIS-Multi`	0.57	4.50E-01		1149	39	47	2765	6.01	1.42E-02	*	1790	55	85	6070
`Multilingual + CIS-Multi vs English + CIS-Multi`	1.28	2.58E-01		1116	96	80	2708	39.10	4.02E-10	***	1704	309	171	5816

Table 26: McNemar's test results of CIS modes comparison on LRL and HRL splits of XQuAD dataset across 6 MLLMs. Sig. – Significance. ✖– the model response is incorrect; ✔– correct.