Gradual Code-Switching In-Context Learning for Cross-Lingual Transfer of Large Language Models

Anonymous ACL submission

Abstract

While large language models (LLMs) have achieved notable progress in multilingual settings, their performance remains uneven across languages as LLMs often rely on English-centric latent representations. In this work, we introduce code-switching in-context learning (**CSICL**), a simple yet effective prompting strategy that gradually transitions from a target language to English within demonstrations and instruction to facilitate their latent reasoning in English. By explicitly scaffolding the reasoning process through controlled code-switching, **CSICL** acts as an implicit linguistic bridge that enhances cross-lingual alignment. We conduct extensive experiments across 4 LLMs, 6 datasets, and 10 languages, spanning both knowledge-intensive and reasoning-oriented domains. Our results demonstrate that **CSICL** consistently outperforms X-ICL baselines, achieving 6.0%p and 4.8%p higher performance in both target and unseen languages, respectively. The improvement is generalized across diverse language families and even more pronounced in low-resource settings, with gains of 14.7%p in target and 5.3%p in unseen languages. These findings establish code-switching as a robust and effective approach for overcoming the cross-lingual misalignment during inference, moving LLMs toward more equitable and effective multilingual systems. [1]

1 Introduction

Large language models (LLMs) have demonstrated remarkable multilingual capabilities, powering diverse tasks such as question answering (Tjuatja et al., 2024), translation (Xu et al., 2024), and reasoning (Shi et al., 2023) across languages. However, recent studies reveal that this competence is far from language-agnostic, as LLMs typically rely on English-centric latent representations (Wendler et al., 2024; Zhong et al., 2025). Consequently, cross-lingual misalignment, such as failures in internally translating entities into English, can lead to a sharp drop in non-English performance (Bafna et al., 2025). This highlights that multilingual competence in LLMs remains uneven, thereby limiting the inclusive deployment of these models across diverse linguistic communities around the world.

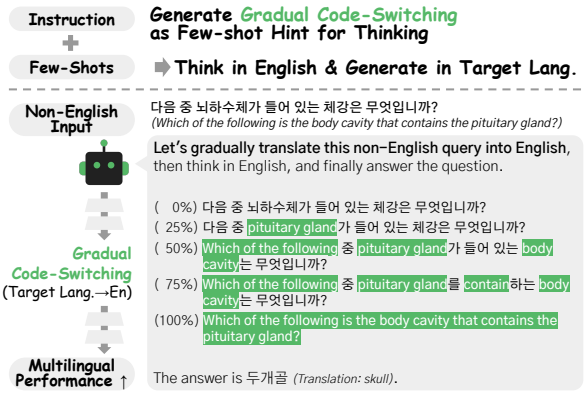


Figure 1: Overview of **CSICL**. We employ (1) gradual code-switching few-shot demonstrations and (2) gradual translation instruction to help the latent process of LLMs for non-English inputs and better align cross-lingual representations during LLM inference.

For robust multilingual generalization, LLMs should better align English-centric latent representations across languages, moving beyond improving a single target language performance. A common approach is cross-lingual in-context learning (X-ICL), which uses demonstrations in high-resource languages (*e.g.*, English) (Lin et al., 2022) to elicit behaviors in low-resource (Cahyawijaya et al., 2024) or unseen (Winata et al., 2022) languages. However, existing X-ICL often transfers superficial task patterns (*e.g.*, output format, label mapping, etc.), employing monolingual demonstrations without necessarily reducing the underlying cross-lingual representational gap.

Meanwhile, code-switching, a natural alterna-

[1] Code available at https://anonymous.4open.science/r/csicl/.

tion between languages, has been shown to benefit cross-lingual transfer in training phases (Wang et al., 2025b; Yoo et al., 2025a; Chai et al., 2025), pointing its potential to improve cross-lingual alignment at inference time. In this paper, we propose code-switching in-context learning (**CSICL**), a prompting strategy that explicitly guides the reasoning process of LLMs through a gradual transition from a target language to English during inference (Figure 1). Specifically, **CSICL** begins in the target language, progressively introduces English tokens via code-switching, and converges to a full English equivalent, using chain-of-thought prompting with few-shot demonstrations. This gradual shift acts as a *linguistic bridge*, nudging LLMs to align cross-lingual representations directly, rather than relying solely on latent translation.

We conduct large-scale, rigorous evaluations across 4 multilingual LLMs in 10 languages on 6 datasets spanning diverse tasks and knowledge domains. We observe that **CSICL** consistently outperforms X-ICL baselines, yielding improvements of 3.1%p and 1.9%p in target and unseen languages, respectively. We generalize this to diverse tasks and highlight its effectiveness, particularly in translation and reasoning tasks. These findings position **CSICL** as a robust, effective direction for cross-lingual transfer, moving LLMs closer to more inclusive, end-to-end multilingual systems.

Our contributions are threefold:

- We propose **CSICL**, a novel code-switching-based prompting strategy that gradually bridges target languages with English representations during inference.
- We conduct a systematic evaluation of **CSICL** across 4 LLMs, 6 tasks, and 10 languages, demonstrating consistent improvements over standard X-ICL baselines.
- We show that **CSICL** is effective across language families, particularly in unseen, low-resource languages.

2 Background

2.1 Code-switching

Code-switching, also known as code-mixing or language alternation, is a common linguistic phenomenon in which a speaker interleaves two or more languages within a single conversational context (Auer, 1998). It occurs in various switching levels: subwords such as at morpheme boundary (*i.e.*, intra-word switching), tag phrases (*i.e.*, tag-switching), words (*i.e.*, intra-sentential switching), and sentences or clauses (*i.e.*, inter-sentential switching). Matrix Language Frame (MLF) model, which distinguishes a grammatically dominant *matrix language* (ML) and an inserted *embedded language* (EL) (Myers-Scotton, 1997), has been widely adopted as a syntactic rule for code-switching. The following examples illustrate code-switching, with variation in the choice of matrix language and embedded language. The sentence originates from Global MMLU (Singh et al., 2025): "*Paper will burn at approximately what temperature in Fahrenheit?*"

(1) *Paper* *will* ***brûler*** *at* *approximately*
ML ML EL ML ML
quelle ***température*** *in* *Fahrenheit* *?*
EL EL ML ML

(ML: English, EL: French)

(2) *En* *degrés* *Fahrenheit,* *à*
ML ML ML ML
approximately ***what*** ***temperature*** *le*
EL EL EL ML
papier ***burn*** *-t-il* *?*
ML EL ML

(ML: French, EL: English)

2.2 In-context learning

In-context learning (ICL) is a paradigm that allows language models to learn tasks given only a few examples in the form of demonstrations (Dong et al., 2024). Brown et al. (2020) investigated a few-shot transfer that concatenates one or more examples as a demonstration, and follow-up studies have explored the selection (Rubin et al., 2022), formatting (Kim et al., 2022), and ordering (Lu et al., 2022) of the demonstration examples. As several tasks require more complex reasoning, Wei et al. (2022) proposed chain-of-thought (CoT) prompting that generates intermediate reasoning steps between inputs and outputs. Kojima et al. (2022); Wang et al. (2022); Zhang et al. (2023b); Zhou et al. (2023) further introduced zero-shot CoT, highlighting the step-by-step reasoning in LLMs.

3 Code-Switching In-Context Learning

In this paper, we propose code-switching in-context learning (**CSICL**) for cross-lingual transfer of LLMs. Figure 1 illustrates an abstract example of **CSICL**. Given (1) gradual translation instruction and (2) gradual code-switching few-shot demonstrations, **CSICL** facilitates the latent process of LLMs to understand non-English input. The exact system prompts for the demonstrations and the instructions are in Appendix A.2.

Instruction. We instruct LLMs to process non-English inputs by progressively translating them into English using code-switching. Following the few-shot demonstrations, we ask them to

1. Repeat a short, single-sentence instruction that guides a model to gradual translation (*i.e.*, "*Let's gradually translate this non-English query into English, then think in English, and finally answer the question.*");
2. Explicitly show their step-by-step, progressive translation process into English;
3. Provide the final answer in the target language, following the specified format (*e.g.*, "*The answer is ...*").

Demonstrations. We employ gradual code-switching as both a few-shot demonstration, which

1. Begins with a query in a target language (*i.e.*, En 0%);
2. Progressively transition to English by leveraging code-switching whose matrix language is the target language and the embedded language is English (*i.e.*, En 25→50→75%);
3. Finally concludes with the full English equivalent (*i.e.*, En 100%).

We randomly sample 5 instances from the test sets as demonstrations and transform them into gradual inter-sentential code-switching from a target language into English. Figure 2 illustrates the two-step pipeline to construct these demonstrations in **CSICL**. Following Kim et al. (2025), we first instruct GPT-5 [2] to generate the code-switching sentences given parallel inputs, by providing both a detailed description of the MLF model and five illustrative examples. We then prompt GPT-5 again with the generated output code-switching sentences and parallel sentences to produce gradual code-switching demonstrations. To ground the system prompt, we carefully curate the real-world code-switching examples from Finer (2014).

[2] Version: `gpt-5-2025-08-07`

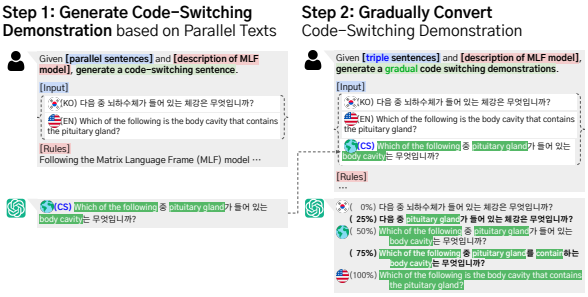


Figure 2: Two-step pipeline to generate gradual code-switching few-shot demonstrations in **CSICL**.

4 Experiments

4.1 Experimental Setup

4.1.1 Evaluation Setup

Following Lin et al. (2022), we use English prompts with cross-lingual 5-shot demonstrations. The samples used as few-shot demonstrations are excluded from the corresponding test set to avoid data leakage. For each experiment, we set 3 target languages in total—*i.e.*, French (*high*), Korean (*mid*), Yoruba (*low*)—and report performances in: (1) the target language and (2) random, unseen languages that do not appear in the demonstrations. Evaluating unseen languages is particularly important, as real-world multilingual systems are often deployed in settings where the test language rarely appears in the training or demonstration phase.

We compare **CSICL** against five conventional X-ICL baselines following Asai et al. (2024):

- **Monolingual** few-shot demonstrations in either English or a target language;
- **Parallel** few-shot demonstrations concatenating the target language and English;
- Instructions to **translate** the target language into either English or a random language, given the parallel few-shots (Chai et al., 2025).

For ablations, we additionally employ five baselines to validate the effectiveness of both (gradual) code-switching demonstrations and instructions:

- Inter-sentential **code-switching (CS)** few-shot demonstrations between the English and the

target language, controlling a matrix and an embedded language;

- **Gradual code-switching** demonstrations in both directions (*i.e.*, English to the target language and the target language to English);
- Zero-shot **gradual translation** instruction that converts the target language query into English.

4.1.2 Evaluation Models

We use four state-of-the-art multilingual LLMs (*i.e.*, two open and two proprietary models). Model versions and details are stated in Appendix A.1.

- `Qwen3-32B` (Yang et al., 2025);
- `deepseek-chat-v3.1` (DeepSeek-AI et al., 2025);
- `grok-4-fast` (xAI, 2025);
- `Gemini 2.5 Flash` (Comanici et al., 2025).

4.1.3 Evaluation Datasets

For the experiments in §4.2.1–4.2.3, we construct a balanced set of 36,000 instances by randomly sampling 600 questions per subject category (six in total) and per language (ten in total) from Global MMLU (Singh et al., 2025) for general knowledge evaluation. The languages include English, three target languages from three distinct language families, and six unseen languages across six different language families spanning resource levels—Chinese, Spanish (*high*); Indonesian, Turkish (*mid*); Swahili, Telugu (*low*).

In §4.2.4, we further conduct an ablation across tasks and knowledge domains using 5 datasets, fixing Spanish as the target and selecting one mid- to high-resource unseen language per dataset (shown in parentheses):

- FLORES+ (Costa-jussà et al., 2024) for machine translation to English (Japanese);
- MedExpQA (Alonso et al., 2024) for domain-specific (*i.e.*, medical) reasoning (French);
- PolyMath (Wang et al., 2025a) for mathematical reasoning (Chinese);
- BLEnD (Myung et al., 2024) for cultural knowledge (Korean);
- MBBQ (Neplenbroek et al., 2024) for social bias (Dutch).

For MBBQ, we randomly select ten samples (five ambiguous contexts and five unambiguous contexts) per template in the MBBQ dataset, totaling 980 samples, due to limited computational resources for processing the full 10k+ instances. For the other task ablation datasets, we employ the entire set for the other datasets. We use accuracy, exact match (EM), and COMET (Rei et al., 2022) as evaluation metrics for multiple-choice questions (Global MMLU, MBBQ), short-answer questions (MedExpQA, PolyMath, BLEnD), and translation (FLORES+), respectively. System prompts for inference and evaluation in each setting and dataset are described in Appendix A.2.

4.1.4 Statistical Significance Testing

Due to limited computational resources, we conduct the following experiments as a single run with a fixed seed. Instead, we apply bootstrap resampling with 2,000 iterations over the evaluation set to rigorously evaluate whether **`CSICL`** significantly outperforms the baseline systems. In each iteration, we compute the performance difference between **`CSICL`** and a given baseline and derive a 95% percentile bootstrap confidence interval (CI) from the resulting distribution. A difference is deemed statistically significant if the lower bound of the CI is greater than zero, indicating that **`CSICL`** consistently outperforms the corresponding baseline. We repeat this analysis against all baselines and mark the results with an asterisk only when **`CSICL`** achieves statistical significance over every baseline.

4.2 Results and Analyses

4.2.1 Main Results

Table 1 reports the results of **`CSICL`** on Global MMLU. Table 6 in Appendix C.5 shows additional experimental results controlling the combination of demonstration and instruction settings. We primarily report experimental results using Qwen 3 in the main text, while the full results for the other three models are stated in Appendix C.5. **`CSICL`** outperforms existing X-ICL baselines in both target and unseen languages. Specifically, **`CSICL`** produces 6.0%p higher performance than monolingual demonstrations in the target language, while the performance gap in English is within 0.2%p. In addition, **`CSICL`** shows substantial cross-lingual transfer in unseen, low-resource languages (+4.8%p), highlighting its practical need in multilingual scenarios. Interestingly, we observe that monolingual demonstrations in the target language

Method	X-ICL setting				En	Tgt.*	Unseen Lang.		
	Demonstration		Instruction				High*	Mid*	Low*
Zero-shot learning	✗		✗		88.6	68.6	86.2	62.1	39.4
Few-shot learning	✓	Monolingual (En)	✗		**88.8**	70.8	86.5	62.8	41.2
	✓	Monolingual (Tgt.)	✗		**88.8**	72.0	86.9	62.1	38.7
	✓	Parallel	✗		88.7	72.7	87.1	63.0	41.4
Zero-shot CoT	✗		✓	Translation (Tgt.→En)	**88.8**	74.5	87.4	63.7	42.0
	✗		✓	Translation (Tgt.→Rnd.)	88.6	73.8	87.5	63.8	42.3
CSICL	✓	Gradual CS (Tgt.→En)	✓	Gradual Translation (Tgt.→En)	88.6	**76.8**	**87.8**	**64.9**	**46.0**

Table 1: Experimental results comparing **CSICL** to existing X-ICL baselines using Qwen 3. X-ICL setting describes each method in terms of a combination of its few-shot demonstrations and instruction. **CSICL** outperforms X-ICL baselines in both target and unseen languages, yielding stable results in English. Bold and underline denote the best and the second-best results, respectively. Tgt. and Rnd. denote a target language and a random language, respectively. Asterisk indicates statistical significance against each and every baseline within the corresponding column.

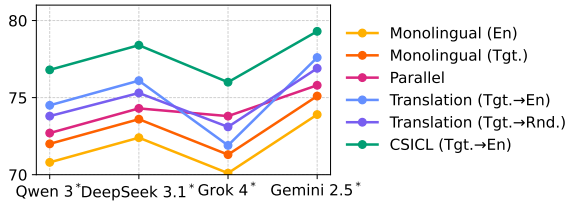


Figure 3: Experimental results of X-ICL approaches in target languages using four models. Tgt. and Rnd. denote a target language and a random language, respectively. Asterisk indicates statistical significance against each and every baseline within the corresponding model.

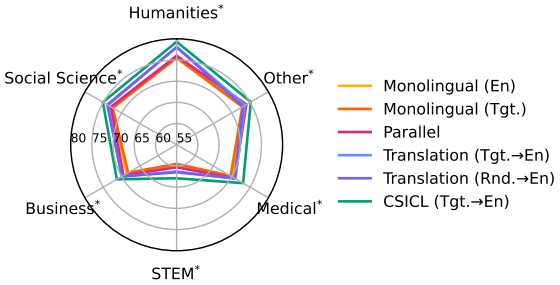


Figure 4: Experimental results of X-ICL approaches for each subject category on Global MMLU using Qwen 3. Tgt. and Rnd. denote a target language and a random language, respectively. Asterisk indicates statistical significance against each and every baseline within the corresponding subject.

yield higher performance, whereas the benefit diminishes in unseen languages. This contrast highlights a key limitation of monolingual prompting: it overfits the demonstrated language rather than fostering cross-lingual alignment. In contrast, **CSICL** explicitly facilitates this alignment through gradual code-switching, resulting in stronger transfer to unseen settings. We do not observe significant differences in the *out-of-format* ratio, which remains below 0.05% for all X-ICL settings. Figure 6 in Appendix C.1 visualizes how **CSICL** aligns cross-lingual representations. We further demonstrate the effectiveness of **CSICL** on across model sizes in Appendix C.2 (Figure 7).

Models. Figure 3 displays experimental results of each model with X-ICL approaches in target languages on Global MMLU. All four models achieve similar trends across X-ICL settings, where **CSICL** significantly outperforms existing baselines.

Subject categories. Figure 4 shows experimental results of X-ICL approaches using Qwen 3 for six subject categories on Global MMLU. **CSICL** significantly surpasses baselines across all categories.

Languages. Figure 5 presents performance differences (%p) of each X-ICL setting compared to zero-shot learning, where their target languages are French, Korean, and Yoruba, respectively. **CSICL** outperforms X-ICL baselines, particularly in target languages and mid- to low-resource unseen languages, while English performance remains stable. Interestingly, the benefits of **CSICL** extend beyond script similarity or language family, implying that it fosters language-agnostic cross-lingual alignment rather than relying on superficial linguistic overlap.

4.2.2 Ablation Study

Table 2 shows the experimental results of the ablation study. We observe that both gradual code-switching demonstrations and translation instruction contribute to cross-lingual transfer, by yield-

Method	X-ICL setting				En	Tgt.	Unseen Lang.		
	Demonstration		Instruction				High	Mid	Low
Few-shot learning	✓	CS (En+Tgt.)	✗		88.6	73.6	87.1	62.9	43.3
	✓	CS (Tgt.+En)	✗		88.7	73.7	87.0	62.8	43.1
	✓	Gradual CS (En→Tgt.)	✗		88.7	74.0	87.6	63.8	44.5
	✓	Gradual CS (Tgt.→En)	✗		88.6	74.2	87.7	64.3	45.7
Zero-shot CoT	✗		✓	Gradual Translation (Tgt.→En)	88.7	74.6	87.2	63.8	42.0
CSICL	✓	Gradual CS (En→Tgt.)	✓	Gradual Translation (En→Tgt.)	88.7	75.0	87.6	64.0	45.2
	✓	Gradual CS (Tgt.→En)	✓	Gradual Translation (Tgt.→En)	88.6	**76.8**	**87.8**	**64.9**	**46.0**

Table 2: Ablation results to verify the effectiveness of both (gradual) code-switching demonstrations and translation instruction in **CSICL**. Both components contribute to improved cross-lingual transfer. Bold and underline denote the best and the second-best results, respectively. Tgt. denotes a target language.

X-ICL setting	En	Tgt.*	Unseen Lang.		
			High*	Mid*	Low*
Paraphrasing (En)	**88.9**	71.0	86.6	63.2	41.6
Paraphrasing (Tgt.)	88.8	72.3	87.0	62.6	39.3
CSICL (Tgt.→En)	88.6	**76.8**	**87.8**	**64.9**	**46.0**

Table 3: Experimental results comparing **CSICL** to paraphrased, monolingual demonstrations, keeping the number of sentences per shot, using Qwen 3. Bold and underline denote the best and the second-best results, respectively. Tgt. denotes a target language. Asterisk indicates statistical significance against each and every baseline within the corresponding column.

ing 3.4%p and 3.8% higher accuracy compared to the monolingual baseline. In particular, gradual code-switching consistently achieves higher performances than code-switching (+0.5%p), parallel (+1.5%p), and monolingual demonstrations (+3.4%p). Furthermore, transitioning from a target language to English outperforms the reverse direction, supporting our hypothesis that **CSICL** alleviates the translation barrier by scaffolding the latent translation process of LLMs. This asymmetry indicates that LLMs benefit more when demonstrations converge to English representations, aligning with their latent space, rather than diverging away from it. Zero-shot and gradual translation instructions yield minimal differences within 0.1%p in the target language, as the model fails to understand the concept of gradual translation and to follow it without explicit demonstrations. Specifically, zero-shot gradual translation instruction often collapses into abrupt translation, where the model outputs two sentences entirely in Korean, followed by three equivalent sentences entirely in English.

4.2.3 CSICL vs. Paraphrasing

CSICL may benefit from using a larger number of sentences, since **CSICL** includes four additional sentences per demonstration. To control for this, we fix the number of sentences per demonstration to five, the same as in **CSICL**, and compare it against monolingual demonstrations. Specifically, for each shot, we add four paraphrased monolingual demonstrations, generated by GPT-5 and then ask the LLMs to solve the task.

Table 3 describes experimental results of **CSICL** and paraphrased demonstrations on Global MMLU, averaged over Qwen 3 and Gemini 2.5. Although paraphrasing provides a marginal improvement in cross-lingual transfer for both target (+0.2%p) and unseen languages (+0.3%p) on average, its effect is limited; **CSICL** consistently outperforms both baselines (+6.0%p and +4.8%p in the target and the unseen languages, respectively, on average). This suggests that the gains of **CSICL** cannot be attributed to a larger demonstration budget alone. Instead, the gradual transition across languages supplies a distinct cross-lingual signal that paraphrasing fails to capture, reinforcing the role of code-switching in aligning multilingual representations. Table 5 in Appendix C.4 reports an ablation study controlling the language of paraphrasing demonstrations.

4.2.4 Task and Domain Knowledge

Table 4 presents experimental results of X-ICL approaches using Qwen 3 across five additional tasks spanning translation, reasoning, and knowledge-intensive evaluation. In line with earlier results, **CSICL** consistently outperforms existing baselines across settings. Following the descriptions provided in the original papers, we categorize tasks into three types:

Task type	Translation		Reasoning-oriented						Knowledge-intensive					
			Domain-specific			Math			Cultural			Social Bias		
X-ICL setting	Tgt.*	Uns.*	En	Tgt.*	Uns.*	En	Tgt.*	Uns.*	En	Tgt.*	Uns.*	En	Tgt.*	Uns.
Monolingual (En)	76.6	72.3	**46.4**	38.9	32.1	51.2	49.3	47.4	83.7	77.8	73.2	92.0	88.9	86.3
Monolingual (Tgt.)	75.3	71.6	45.8	38.5	30.3	50.5	49.7	46.1	83.1	78.9	71.5	91.6	89.3	85.8
Parallel	78.1	73.0	46.3	40.2	31.9	**51.4**	49.7	47.7	**83.8**	79.2	72.0	91.5	89.5	86.0
Translation (Tgt.→En)	74.8	72.1	46.1	39.6	32.3	50.8	50.8	48.2	83.3	78.5	71.7	**92.1**	90.1	86.4
Translation (Tgt.→Rnd.)	73.2	73.7	46.3	40.1	32.7	50.3	51.0	47.9	83.6	78.1	72.4	91.9	89.8	**87.3**
CSICL (Tgt.→En)	**83.4**	**75.3**	46.2	**44.4**	**35.2**	50.9	**54.5**	**51.8**	83.5	**81.3**	**74.5**	91.8	**90.5**	87.2

Table 4: Experimental results on 5 diverse tasks and domain knowledge using Qwen 3. **CSICL** achieves larger improvements in translation and reasoning tasks. Bold and underline denote the best and the second-best results, respectively. Tgt. and Rnd. denote a target language and a random language, respectively. Asterisk indicates statistical significance against each and every baseline within the corresponding column.

- Machine translation (FLORES+);
- Reasoning-oriented tasks (MedExpQA, PolyMath);
- Knowledge-intensive tasks (BLEnD, MBBQ).

Translation. **CSICL** achieves the largest gains in the translation task—gains of +6.8%p and +3.0%p COMET in the target and the unseen languages, respectively, over monolingual demonstrations. This strong effect is expected, as progressive transition directly scaffolds the latent translation process, aligning the input with English-centric representations. It is noteworthy that **CSICL** is effective even in a purely generative setting.

Reasoning. Beyond translation, **CSICL** also improves reasoning-oriented tasks, with average gains of +5.4%p and +3.8%p of EM over monolingual baselines. We suppose that gradual transitions mitigate representational interference: instead of reasoning entirely in a non-English space, the demonstrations nudge the model to "*think in English,*" where reasoning capabilities are strongest. This highlights the potential of **CSICL** as a lightweight test-time alignment mechanism, complementary to scaling methods such as self-consistency.

Knowledge. For cultural knowledge and social bias, the improvements are modest but still consistent (+2.6%p and +1.1%p in the target and the unseen languages, respectively, on average). This indicates that **CSICL** not only facilitates translation-like tasks but also improves general factual consistency across languages.

5 Related Work

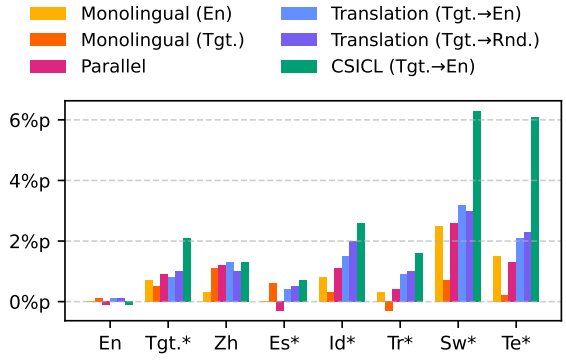
5.1 Cross-lingual Transfer in LLMs

Recent studies suggest that LLMs tend to rely on English as a pivot for cross-lingual processing. Zhao et al. (2024) showed that LLMs exhibit language-specific neurons to process multilingual inputs, and Schut et al. (2025) demonstrated that LLMs internally translate multilingual inputs into English representations and think in English. This reliance on internal translation introduces a critical vulnerability: *translation barrier*, where a failure in the initial translation stage is propagated to the quality of the final outputs (Bafna et al., 2025). While external machine translation can slightly mitigate this issue (Etxaniz et al., 2024; Intrator et al., 2024), it undermines the goal of using LLMs as seamless, end-to-end multilingual systems. This motivates the exploration of mechanisms that can directly activate cross-lingual capabilities of LLMs without relying on latent English translation.

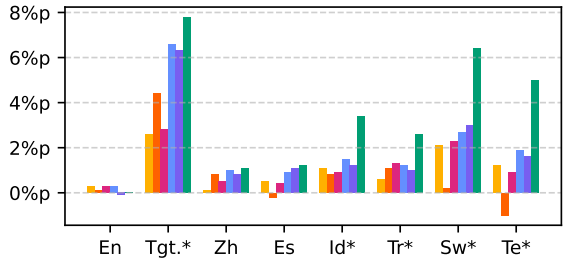
To bridge this gap, cross-lingual transfer has been widely studied (Pallucchini et al., 2025). For example, Tanwar et al. (2023); Li et al. (2024); Lin et al. (2025) investigated X-ICL approaches to better align representations across languages. However, these approaches are primarily based on *monolingual* demonstrations, requiring a more effective signal that is multilingual itself and can mitigate cross-lingual misalignment during inference.

5.2 Code-switching for Cross-lingual Transfer

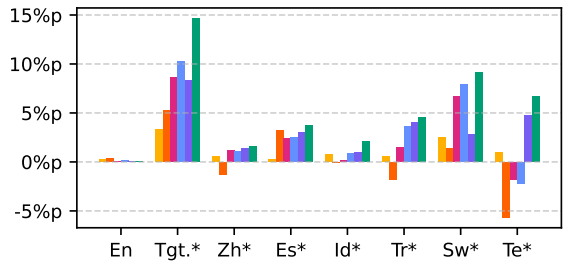
Code-switching, which has been studied over decades by the natural language processing community (Winata et al., 2023), is a promising candidate for such a direct signal. Code-switching has shown



(a) Target: French (*high*)



(b) Target: Korean (*mid*)



(c) Target: Yoruba (*low*)

Figure 5: Performance differences (%p) of **CSICL** and X-ICL baselines compared to zero-shot learning setting per a target language using Qwen 3. Tgt. and Rnd. denote a target and a random language. Asterisk indicates statistical significance against each and every baseline within the corresponding language.

its efficacy for cross-lingual transfer in various training stages, including pre-training (Wang et al., 2025b), continual pre-training (Yoo et al., 2025a), supervised fine-tuning (SFT) (Lee et al., 2024; Chai et al., 2025), and selective constrained decoding (Li et al., 2025). In particular, Chai et al. (2025) demonstrated that (1) SFT using inter-sentential code-switching data and (2) chain-of-thought prompting, translating a non-English query into a random language and answering in the target language, enhances cross-lingual transfer. Li et al. (2025) introduced probe-guided decoding with an additional classifier to enhance bilingual LLM reasoning with a case study of English-Chinese code-switching. Kang et al. (2026) suggested to selectively incorporate English translations into the latent reasoning of LLMs using additional classifiers.

However, these approaches require substantial additional training resources, while exploration remains limited at the inference level, focusing only on a specific task and domain knowledge. For machine reading comprehension, Kim et al. (2024) provided a passage in English and a QA pair in a target language, which is a form of intra-sentential code-switching. Kim et al. (2025) showcased a case study where Korean-English inter-sentential code-switching can unlock Korean cultural knowledge that remains inaccessible in English queries. Our work, in contrast, introduces a training-free, task- and language-agnostic approach that systemically integrates inter-sentential code-switching into X-ICL. By progressively transitioning from the target language to English within demonstrations and instruction, **CSICL** improves cross-lingual transfer of LLMs during inference.

6 Conclusion

In this work, we address cross-lingual misalignment in multilingual LLMs—their over-reliance on English-centric latent representations that often leads to degraded performance in non-English. To mitigate this issue, we introduce **CSICL**, a code-switching based prompting strategy that better aligns cross-lingual representations of multilingual LLMs. By gradually transitioning from a target language to English, **CSICL** serves as a linguistic bridge that improves cross-lingual alignment without requiring additional training or resources. We rigorously examine **CSICL** across 4 multilingual LLMs, 6 datasets, and 10 languages. **CSICL** consistently outperforms conventional X-ICL baselines, achieving gains of 6.0%p and 4.8%p in both target and unseen languages, respectively. We demonstrate its pronounced improvements in low-resource settings, with gains of 14.7%p in target and 5.3%p in unseen languages. Beyond empirical improvements, this paper positions code-switching as a new lens for cross-lingual modeling during LLM inference. Rather than treating language alternation as noise, we frame it as a resource for bridging linguistic gaps, opening a promising view in robust and inclusive multilingual LLM research.

Limitations

In this paper, we focus on inter-sentential code-switching, which can occur at various switching levels. We consider this as a natural first step, and extending the approach to other switching levels is an important direction for future work.

For evaluation, we largely employ automated metrics (*i.e.*, accuracy, exact match, and COMET), which could be considered a limitation. We adopt automated metrics for consistency and scalability across multilingual settings, which is central to our study. Further, a small-scale human evaluation in Appendix C.3 shows strong agreement with the COMET-based results, indicating that our main findings are robust.

We conduct experiments on 10 languages spanning different resource levels. While this set cannot cover all language families, it offers a broad and representative sample that supports the generality of **CSICL**.

In our task- and domain-specific evaluation (§4.2.4), we fix Spanish as the target language and vary the unseen language across datasets due to the limited availability of parallel data in specialized tasks and domains. While this design enables broad coverage of domains, part of the performance differences observed across tasks may stem from the choice of unseen languages rather from the tasks themselves. It reflects the practical challenges of multilingual evaluation and still provides a diverse and informative basis for analysis.

Ethical Considerations

All datasets used in this study are publicly available and employed solely for research purposes, consistent with their intended licenses. We prevent data leakage by excluding the samples used as few-shot demonstrations from the corresponding test set. **CSICL** aims to improve the multilingual performance of LLMs, with the intended use for both research and practical applications, which is not used for malicious purposes. We use ChatGPT, Gemini, and Copilot for writing and coding assistance.

References

Samir Abdaljalil, Erchin Serpedin, Khalid Qaraqe, and Hasan Kurban. 2025. Evaluating multilingual and code-switched alignment in LLMs via synthetic natural language inference. *arXiv preprint arXiv:2508.14735*.

Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. MedExpQA: Multilingual benchmarking of large language models for medical question answering. *Artificial Intelligence in Medicine*, 155:102938.

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.

Peter Auer. 1998. *Code-switching in conversation*. Routledge, London, England.

Niyati Bafna, Tianjian Li, Kenton Murray, David R. Mortensen, David Yarowsky, Hale Sirin, and Daniel Khashabi. 2025. The translation barrier hypothesis: Multilingual generation with large language models suffers from implicit translation failure. *arXiv preprint arXiv:2506.22724*.

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, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. LLMs are few-shot in-context low-resource language learners. 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 405–433, Mexico City, Mexico. Association for Computational Linguistics.

Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xinnian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, and Zhoujun Li. 2025. XCOT: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22):23550–23558.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

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

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de 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.

Daniel L. Finer. 2014. Movement triggers and reflexivization in Korean-English codeswitching. In *Grammatical Theory and Bilingual Codeswitching*. The MIT Press.

Muhammad Huzaifah, Weihua Zheng, Nattapol Chanpaisit, and Kui Wu. 2024. Evaluating code-switching translation with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6381–6394, Torino, Italia. ELRA and ICCL.

Yotam Intrator, Matan Halfon, Roman Goldenberg, Reut Tsarfaty, Matan Eyal, Ehud Rivlin, Yossi Matias, and Natalia Aizenberg. 2024. Breaking the language barrier: Can direct inference outperform pre-translation in multilingual LLM applications? 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 829–844, Mexico City, Mexico. Association for Computational Linguistics.

Sai Muralidhar Jayanthi, Kavya Nerella, Khyathi Raghavi Chandu, and Alan W Black. 2021. CodemixedNLP: An extensible and open NLP toolkit for code-mixing. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 113–118, Online. Association for Computational Linguistics.

Deokhyung Kang, Seonjeong Hwang, Daehui Kim, Hyounghun Kim, and Gary Geunbae Lee. 2026. Why do multilingual reasoning gaps emerge in reasoning language models? *arXiv preprint arXiv:2510.27269*.

Hyuhng Joon Kim, Hyunsoo Cho, Junyeob Kim, Taeuk Kim, Kang Min Yoo, and Sang goo Lee. 2022. Self-generated in-context learning: Leveraging auto-regressive language models as a demonstration generator. *arXiv preprint arXiv:2206.08082*.

Seoyeon Kim, Huiseo Kim, Chanjun Park, Jinyoung Yeo, and Dongha Lee. 2025. Can code-switched texts activate a knowledge switch in LLMs? a case study on English-Korean code-switching. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22303–22327, Suzhou, China. Association for Computational Linguistics.

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

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*, volume 35, pages 22199–22213. Curran Associates, Inc.

Jaeseong Lee, YeonJoon Jung, and Seung-won Hwang. 2024. COMMIT: Code-mixing English-centric large language model for multilingual instruction tuning. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3130–3137, Mexico City, Mexico. Association for Computational Linguistics.

Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2024. Improving in-context learning of multilingual generative language models with cross-lingual alignment. 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 8058–8076, Mexico City, Mexico. Association for Computational Linguistics.

Yihao Li, Jiayi Xin, Miranda Muqing Miao, Qi Long, and Lyle Ungar. 2025. The impact of language mixing on bilingual LLM reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32519–32536, Suzhou, China. Association for Computational Linguistics.

Peiqin Lin, Andre Martins, and Hinrich Schuetze. 2025. XAMPLER: Learning to retrieve cross-lingual in-context examples. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3968–3977, Albuquerque, New 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 Diab, and 2 others. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Amr Mohamed, Yang Zhang, Michalis Vazirgiannis, and Guokan Shang. 2025. Lost in the mix: Evaluating LLM understanding of code-switched text. *arXiv preprint arXiv:2506.14012*.

Carol Myers-Scotton. 1997. *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2024. BLEnD: A benchmark for LLMs on everyday knowledge in diverse cultures and languages. In *Advances in Neural Information Processing Systems*, volume 37, pages 78104–78146. Curran Associates, Inc.

Vera Neplenbroek, Arianna Bisazza, and Raquel Fernández. 2024. MBBQ: A dataset for cross-lingual comparison of stereotypes in generative LLMs. In *First Conference on Language Modeling*.

Filippo Pallucchini, Lorenzo Malandri, Fabio Mercorio, and Mario Mezzanzanica. 2025. Lost in alignment: A survey on cross-lingual alignment methods for contextualized representation. *ACM Comput. Surv.*

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Mohd Sanad Zaki Rizvi, Anirudh Srinivasan, Tanuja Ganu, Monojit Choudhury, and Sunayana Sitaram. 2021. GCM: A toolkit for generating synthetic code-mixed text. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 205–211, Online. Association for Computational Linguistics.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. Do multilingual LLMs think in English? *arXiv preprint arXiv:2502.15603*.

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*.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Jiayang Song, Yuheng Huang, Zhehua Zhou, and Lei Ma. 2025. Multilingual blending: Large language model safety alignment evaluation with language mixture. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3433–3449, Albuquerque, New Mexico. Association for Computational Linguistics.

Sathya Krishnan Suresh, Tanmay Surana, Lim Zhi Hao, and Eng Siong Chng. 2025. CS-sum: A benchmark for code-switching dialogue summarization and the limits of large language models. In *Proceedings of The 5th New Frontiers in Summarization Workshop*, pages 31–47, Hybrid. Association for Computational Linguistics.

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. 2023. Multilingual LLMs are better cross-lingual in-context learners with alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6292–6307, Toronto, Canada. Association for Computational Linguistics.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. 2024. Do LLMs exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

Boshi Wang, Xiang Deng, and Huan Sun. 2022. Iteratively prompt pre-trained language models for chain of thought. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2714–2730, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yiming Wang, Pei Zhang, Jialong Tang, Haoran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, Qiqian Cang, Yichang Zhang, Fei Huang, Junyang Lin, Fei Huang, and Jingren Zhou. 2025a. PolyMath: Evaluating mathematical reasoning in multilingual contexts. *arXiv preprint arXiv:2504.18428*.

Zhijun Wang, Jiahuan Li, Hao Zhou, Rongxiang Weng, Jingang Wang, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2025b. Investigating and scaling up code-switching for multilingual language model pre-training. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11032–11046, Vienna, Austria. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

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 Winata, Alham Fikri Aji, Zheng Xin Yong, and Thamar Solorio. 2023. The decades progress on code-switching research in NLP: A systematic survey on trends and challenges. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978, Toronto, Canada. Association for Computational Linguistics.

Genta Winata, Shijie Wu, Mayank Kulkarni, Thamar Solorio, and Daniel Preotiuc-Pietro. 2022. Cross-lingual few-shot learning on unseen languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 777–791, Online only. Association for Computational Linguistics.

xAI. 2025. Grok 4 fast model card.

Peng Xie, Xingyuan Liu, Tsz Wai Chan, Yequan Bie, Yangqiu Song, Yang Wang, Hao Chen, and Kani Chen. 2025. SwitchLingua: The first large-scale multilingual and multi-ethnic code-switching dataset. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yilun Yang and Yekun Chai. 2025. CodeMixBench: Evaluating code-mixing capabilities of LLMs across 18 languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2139–2169, Suzhou, China. Association for Computational Linguistics.

Haneul Yoo, Cheonbok Park, Sangdoo Yun, Alice Oh, and Hwaran Lee. 2025a. Code-switching curriculum learning for multilingual transfer in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7816–7836, Vienna, Austria. Association for Computational Linguistics.

Haneul Yoo, Yongjin Yang, and Hwaran Lee. 2025b. Code-switching red-teaming: LLM evaluation for safety and multilingual understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13392–13413, Vienna, Austria. Association for Computational Linguistics.

Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, and Alham Fikri Aji. 2023a. Multilingual large language models are not (yet) code-switchers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12567–12582, Singapore. Association for Computational Linguistics.

Wenbo Zhang, Aditya Majumdar, and Amulya Yadav. 2025. CHAI for LLMs: Improving code-mixed translation in large language models through reinforcement learning with AI feedback. *arXiv preprint arXiv:2411.09073*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023b. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

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.

Chengzhi Zhong, Qianying Liu, Fei Cheng, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. 2025. What language do non-English-centric large language models think in? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26333–26346, Vienna, Austria. Association for Computational Linguistics.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

Appendix

A Reproducibility Statement

A.1 Evaluation Details

For open-source LLM inference, we use 4 H200 GPUs with 564 GB memory and 4 RTX 8000 GPUs with 188 GB memory. For closed-source LLMs, we access them via OpenRouter[3]. All experiments and sample selections are conducted with a random seed of 42. For reliable reproduction, we use greedy decoding with a temperature of 0.0, if possible. All COMET scores in this paper are COMET-22 (`unbabel/wmt22-comet-da` (Rei et al., 2022).

A.2 System Prompts

A.2.1 Few-shot Demonstrations Generation

Prompt for generating English-Korean code-switching few-shot demonstrations

You are a bilingual rewriting assistant.

[TASK]
• Input : an English sentence (E) and its Korean translation (K)
• Output : the code-switching version of the parallel sentences following Matrix Language Frame (MLF) model
- Replace about 50% percent of words/phrases in E with their Korean equivalents taken from K
- Keep the original English word order and follow English syntax (S-V-O)
- DO NOT add explanations, examples, tags, prefix or extra sentences
- If there is no suitable Korean equivalent, keep the English word

[EXAMPLE]
<English> I ate dinner quickly.
<Korean> 나는 저녁을 빨리 먹었다.
<Code-Switching> I ate 저녁 빨리.

<English> Hana, put the toys in the basket quickly and go home.
<Korean> 하나야, 바구니에 장난감을 빨리 넣고 집에 가자.
<Code-Switching> Hana, put 장난감 in the basket quickly and 집에 가자.

<English> Dad was about to throw away my tooth.
<Korean> 아빠가 내 이빨을 빼려고 했어.
<Code-Switching> 아빠 was about to 빨래 my 이빨.

<English> I have to wash my hand.
<Korean> 나는 손을 씻어야 해.
<Code-Switching> I have to 닦아 my hand.

<English> Tom thinks Bill likes himself.
<Korean> 톰은 빌이 자기 자신을 좋아한다고 생각한다.
<Code-Switching> Tom thinks that Bill이 자기를 좋아한다.

<English> Tom thinks Bill likes himself.
<Korean> 톰은 빌이 자기 자신을 좋아한다고 생각한다.
<Code-Switching> Tom이 생각하기를 Bill likes himself.

[BEGIN TASK]

Prompt for generating Korean-English code-switching few-shot demonstrations

You are a bilingual rewriting assistant.

[TASK]
• Input : an English sentence (E) and its Korean translation (K)
• Output : the code-switching version of the parallel sentences following Matrix Language Frame (MLF) model
- Replace about 50% percent of words/phrases in K with their English equivalents taken from E
- Keep the original English word order and follow Korean syntax (S-O-V)
- DO NOT add explanations, examples, tags, prefix or extra sentences
- If there is no suitable English equivalent, keep the Korean word

[EXAMPLE]
<English> Meena, put all the toys in the basket quickly, and go home.
<Korean> 미나야, 바구니에 장난감을 다 넣고 빨리 집에 가자.
<Code-Switching> Meena, basket 안에다 all the toys를 빨리 put하고 집에 가자.

<English> Last time, by mistake, they did not renew my driver's license at the DMV.
<Korean> 지난 번에 도로교통공단이 내 운전면허증을 실수로 갱신하지 않았어요.
<Code-Switching> 지난번에 motor vehicle department 에서 내 driver's license를 mistake로 갱신하지 않았어요.

<English> They often do a thing like that.
<Korean> 그들이 일을 종종 그렇게 해요.
<Code-Switching> 그들이 일을 often 그렇게 해요.

<English> Tom thinks Bill likes himself.
<Korean> 톰은 빌이 자기 자신을 좋아한다고 생각한다.
<Code-Switching> Tom은 Bill이 himself를 좋아한다고 생각한다.

<English> John wonders what Mary bought yesterday.
<Korean> 존은 메리가 어제 무엇을 샀는지 궁금해한다.
<Code-Switching> John은 Mary가 yesterday 무엇을 샀는지 궁금해한다.

[BEGIN TASK]

Prompt for generating gradual code-switching (En→Ko) few-shot demonstrations

You are a bilingual rewriting assistant.
Your task is to generate five versions of a sentence that gradually transition from English to Korean.

[INPUT]
• One English sentence (E)
• Its Korean translation (K)
• A code-switching version of the sentence (C), where about 50% of English words are replaced by Korean

[3] https://openrouter.ai/

equivalents

[OUTPUT]
Generate a sequence of five sentences showing a smooth progression from English to Korean:
1. English only (100% English, source syntax S-V-O)
2. 75% English + 25% Korean (matrix language: English, embedded language: Korean)
3. 50% English + 50% Korean (matrix language: English, embedded language: Korean)
4. 25% English + 75% Korean (matrix language: English, embedded language: Korean)
5. Korean only (100% Korean, target syntax S-O-V)

[RULES]
Following the Matrix Language Frame (MLF) model,
• Preserve English word order (S-V-O) and syntax until version 5 (full Korean).
• Use Korean equivalents from K when inserting Korean into English sentences.
• Keep the code-switching natural and consistent, not random.
• Do not add explanations, notes, or extra text — output only the five sentences in order.

[EXAMPLES]

EXAMPLE 1
Input:
<English> I ate dinner quickly.
<Korean> 나는 저녁을 빨리 먹었다.
<Code-Switching> I ate 저녁 빨리.

Output:
1. I ate dinner quickly.
2. I ate dinner 빨리.
3. I ate 저녁 빨리.
4. 나는 저녁 빨리 ate.
5. 나는 저녁을 빨리 먹었다.

EXAMPLE 2
Input:
<English> Dad was about to throw away my tooth.
<Korean> 아빠가 내 이빨을 빼려고 했어.
<Code-Switching> 아빠 was about to 뺄래 my 이빨.

Output:
1. Dad was about to throw away my tooth.
2. Dad was about to throw away 내 이빨.
3. 아빠 was about to 뺄래 my 이빨.
4. 아빠 was about to 내 이빨 빼려고 했어.
5. 아빠가 내 이빨을 빼려고 했어.

Example 3
Input:
<English> Tom thinks Bill likes himself.
<Korean> 톰은 빌이 자기 자신을 좋아한다고 생각한다.
<Code-Switching> Tom thinks that Bill이 자기를 좋아한다.

Output:
1. Tom thinks Bill likes himself.
2. Tom thinks Bill likes 자기.
3. Tom thinks that Bill이 자기를 좋아한다.
4. Tom은 Bill이 자기를 좋아한다고 think한다.
5. 톰은 빌이 자기 자신을 좋아한다고 생각한다.

[BEGIN TASK]

Prompt for generating gradual code-switching (Ko→En) few-shot demonstrations

You are a bilingual rewriting assistant.
Your task is to generate five versions of a sentence that gradually transition from Korean to English.

[INPUT]
• One English sentence (E)
• Its Korean translation (K)
• A code-switching version of the sentence (C), where about 50% of Korean words are replaced by English equivalents

[OUTPUT]
Generate a sequence of five sentences showing a smooth progression from Korean to English:
1. Korean only (100% Korean, source syntax S-O-V)
2. 75% Korean + 25% English (matrix language: Korean, embedded language: English)
3. 50% Korean + 50% English (matrix language: Korean, embedded language: English)
4. 25% Korean + 75% English (matrix language: Korean, embedded language: English)
5. Korean only (100% English, target syntax S-V-O)

[RULES]
Following the Matrix Language Frame (MLF) model,
• Preserve English word order (S-O-V) and syntax until version 5 (full English).
• Use Korean equivalents from E when inserting English into Korean sentences.
• Keep the code-switching natural and consistent, not random.
• Do not add explanations, notes, or extra text — output only the five sentences in order.

[EXAMPLES]

EXAMPLE 1
Input:
<Korean> 미나야, 바구니에 장난감을 다 넣고 빨리 집에 가자.
<English> Meena, put all the toys in the basket quickly, and go home.
<Code-Switching> Meena, basket 안에다 all the toys를 빨리 put하고 집에 가자.

Output:
1. 미나야, 바구니에 장난감을 다 넣고 빨리 집에 가자.
2. Meena, 바구니에 장난감을 다 put하고 빨리 집에 가자.
3. Meena, basket 안에다 all the toys를 빨리 put하고 집에 가자.
4. Meena, put all the toys in the basket quickly, 집에 가자.
5. Meena, put all the toys in the basket quickly, and go home.

EXAMPLE 2
Input:
<Korean> 지난 번에 도로교통공단이 내 운전면허증을 실수로 갱신하지 않았어요.
<English> Last time, by mistake, they did not renew my driver's license at the DMV.
<Code-Switching> 지난번에 motor vehicle department에서 내 driver's license를 mistake로 갱신하지 않았어요.

Output:
1. 지난 번에 도로교통공단이 내 운전면허증을 실수로 갱신하지 않았어요.

2. 지난 번에 motor vehicle department가 내 운전면허증을 실수로 갱신하지 않았어요.
3. 지난번에 motor vehicle department에서 내 driver's license를 mistake로 갱신하지 않았어요.
4. Last time, motor vehicle department에서 my driver's license를 mistake로 renew하지 않았어요.
5. Last time, by mistake, they did not renew my driver's license at the DMV.

Example 3
Input:
<Korean> 존은 메리가 어제 무엇을 샀는지 궁금해한다.
<English> John wonders what Mary bought yesterday.
<Code-Switching> John은 Mary가 yesterday 무엇을 샀는지 궁금해한다.

Output:
1. 존은 메리가 어제 무엇을 샀는지 궁금해한다.
2. John은 메리가 yesterday 무엇을 샀는지 궁금해한다.
3. John은 Mary가 yesterday what을 샀는지 궁금해한다.
4. John wonders what Mary가 yesterday 샀는지 궁금해한다.
5. John wonders what Mary bought yesterday.

[BEGIN TASK]

A.2.2 Code-switching In-context Learning

Prompt for `CSICL`

Answer the following questions written in non-English by explicitly showing a step-by-step translation process into English, then provide the final answer.

Core Behaviors

1. Mandatory Opening Self-Instruction
- Always begin with the exact sentence: "Let's gradually translate this non-English query into English, then think in English, and finally answer the question."
2. Gradual Code-Switching Output:
- Show the transformation from the original query to English in 5 steps.
- Each step progressively replaces more non-English words with English until the query is fully natural English.
- End with a clean, natural English rendering of the question.
3. Answer Format:
- End with the final answer in the exact format: "The answer is X", where 'X' is a single uppercase letter of the correct choice (e.g., A, B, C, D).
- No explanations or justifications after the answer.

Unlike hidden scratchpad reasoning, note that the translation and reasoning process must be explicitly output. Keep the structure identical across all responses.

`{Few-shot demonstrations}`

A.2.3 Evaluation Experiments

Prompt for multiple-choice QA experiments

You will be asked to answer a multiple-choice question. Read the following question and choices then select the single best answer.

- Output only the single uppercase letter of the correct choice (e.g., A, B, C, D).
- Do not output anything else: no reasoning, no explanation, no restating the question, no prefixes like "Answer:", no punctuation, spaces, or newlines.

The final output must be exactly one letter.

Prompt for short-answer QA experiments

You will be asked to answer a short-answer question. Read the following question and provide a single answer without any explanations.

- Output only the single answer.
- Do not output anything else: no reasoning, no explanation, no restating the question, no prefixes like "Answer:", no punctuation, spaces, or newlines.

Prompt for machine translation experiments

You will be asked to translate a `{target language}` text. Read the following sentence and translate it into English without any explanations.

- Output only the English equivalent.
- Do not output anything else: no reasoning, no explanation, no restating the question, no prefixes like "Answer:", no punctuation, spaces, or newlines.

A.3 Licenses

All datasets used in this paper are publicly available. Global MMLU (Singh et al., 2025) is released under the `Apache-2.0` license. BLEnD (Myung et al., 2024), FLORES+ (Costa-jussà et al., 2024), and MedExpQA (Alonso et al., 2024) are distributed under the `CC-BY-SA-4.0` and `CC-BY-4.0` licenses, respectively. The licenses of MBBQ (Neplenbroek et al., 2024) and PolyMath (Wang et al., 2025a) are not explicitly specified.

B Additional Related Work

In this section, we provide further discussions on prior studies related to code-switching.

Code-switching Text Generation. Only a limited number of code-switching corpora and labeled datasets exist for specific language pairs, and code-switching among non-English languages is hardly available (Winata et al., 2023). Hence, the synthetic generation of code-switching has become the primary strategy. Early efforts such as Jayanthi et al. (2021); Rizvi et al. (2021) proposed toolkits for Hindi-English, although they lack generalizability across languages. More recently, prompting LLMs (*i.e.*, GPT-4) has been widely adopted for synthetic data generation (Yoo et al., 2025a,b; Kim et al., 2025; Yang and Chai, 2025). In particular, Yoo et al. (2025a) further provided quantitative and qualitative analyses comparing LLM-generated code-switching text with human bilingual data, showing

that such text largely follows natural switching patterns with one distinctive feature—redundant synonyms appearing in both languages—which may actually help LLMs by providing explicit lexical alignment cues. Therefore, a slight "unnaturalness" of synthetic code-switching may not be a weakness but rather a source of beneficial noise that facilitates multilingual transfer. In addition, Xie et al. (2025) introduced LinguaMaster, a LLM multi-agent collaboration framework specifically designed for controlled code-switching generation.

Code-switching for Multilingual Evaluations. As the multilingual capabilities of LLMs advance, recent benchmarks range from broad NLP tasks (Yang and Chai, 2025; Abdaljalil et al., 2025; Mohamed et al., 2025) to domain-specific evaluations, including translation (Huzaifah et al., 2024; Zhang et al., 2025), dialogue summarization (Suresh et al., 2025), and red-teaming (Song et al., 2025; Yoo et al., 2025b). Despite these efforts, Zhang et al. (2023a) reported that LLMs still struggle with code-switching, sometimes underperforming compared to smaller, fine-tuned systems. This suggests that code-switching remains an unsolved challenge for multilingual evaluation, rather than a solved proxy task.

C Additional Results & Discussions

C.1 Layer-wise Visualization of CSICL

We analyze whether **CSICL** helps an LLM resolve cross-lingual misalignment by explicitly facilitating an earlier transition to English latent representations. Using `Qwen3-32B` on the randomly sampled Global-MMLU, we evaluate three input conditions: the Korean original (Ko), a parallel English translation as an anchor (En), and **CSICL** applied to the same Korean set (**CSICL**). For each condition, we run forward passes and extract hidden representations from every layer, focusing on the test query span (excluding few-shot context) via span-based pooling to obtain a single vector per layer per example. We then visualize how these representations evolve across depth by projecting layer-wise vectors with PCA, plotting matched triples (*i.e.*, En, Ko, **CSICL**) in a shared space. Figure 6 visualizes the representations of `Qwen3-32B` with **CSICL** on Global MMLU, compared to English results and Korean results as anchors. We reveal that **CSICL** systematically shifts the non-English inputs toward the English anchor.

C.2 Scalability of CSICL

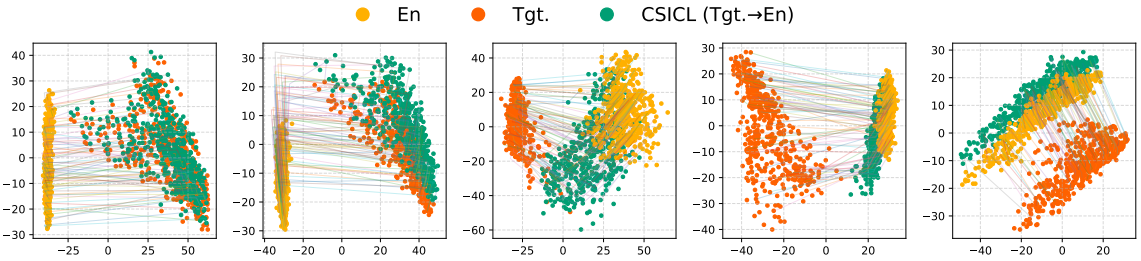
We examine the effectiveness of **CSICL** on small language models (SLMs), which fall behind larger models in instruction-following performance. We test different sizes of `Qwen 3` (*i.e.*, 0.6B, 1.7B, 4B, 8B, 14B, and 32B) on the sampled set of Global MMLU in Korean as a target language. Figure 7 reports the target language performance of X-ICL baselines and **CSICL** across model sizes, shown in a log scale. While the gap between **CSICL** and X-ICL baselines is particularly pronounced when the model size is larger than 4B, we observe that **CSICL** consistently outperforms all X-ICL baselines across model sizes.

C.3 Human Validation on COMET

We additionally conduct a pairwise human evaluation to validate the experimental results obtained with COMET. We randomly sample 200 FLORES+ examples and compare **CSICL** against the best-performing baseline for each source language (*i.e.*, Parallel for target language translation and Translation (Tgt.→Rnd.) for unseen language translation, respectively). Following the WMT practice, one author, who are native in Korean and fluent in English, are shown two anonymized translations (**CSICL** vs. baseline) and asked to choose "A wins / B wins / tie" based on overall adequacy and fluency. **CSICL** is preferred in 63% of cases (and 16% ties) on target-language performance, and in 68% of cases (and 9% ties) on unseen-language performance, respectively. These results confirm that **CSICL** produces more natural and rational translations, as measured by both automatic metrics and human preference.

C.4 Ablation Study on Paraphrasing

We conduct an ablation study on Section 4.2.3 by controlling the number of English and Korean few-shot demonstrations. Table 5 shows experimental results on Global MMLU using `Qwen 3` under different combinations of paraphrasing baselines. We observe mixed results across the paraphrasing baselines for target language and for unseen languages in high- and mid-resource languages. Interestingly, we find that increasing English demonstrations helps only in unseen low-resource languages. However, **CSICL** consistently outperforms all paraphrasing baselines across target and unseen languages.



(a) Layer 1 (b) Layer 8 (c) Layer 16 (d) Layer 32 (e) Layer 64

Figure 6: PCA visualizations of Qwen 3 with **CSICL** on Global MMLU, compared to English results and Korean results as anchors.

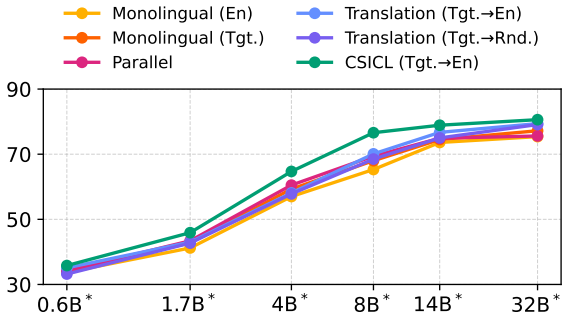


Figure 7: Experimental results of X-ICL approaches in the target language (Ko) on Global MMLU using different sizes of Qwen 3. X-axis is on a log scale. Tgt. and Rnd. denote a target language and a random language, respectively.

# En	# Tgt.	En	Tgt.*	Unseen Lang.		
				High*	Mid*	Low*
5	0	**88.9**	71.0	86.6	63.2	41.6
4	1	88.6	71.3	86.8	63.2	41.0
3	2	**88.9**	71.7	86.9	62.7	40.7
2	3	88.7	72.0	87.1	63.0	40.2
1	4	88.7	72.5	86.7	62.7	39.5
0	5	88.8	72.3	87.0	62.6	39.3
CSICL		88.6	**76.8**	**87.8**	**64.9**	**46.0**

Table 5: Experimental results comparing **CSICL** to paraphrasing baselines, controlling the number of English and target language few-shot demonstrations.# En and # Tgt. denote the number of English demonstrations and the number of target language demonstrations, respectively. Bold and underline denote the best and the second-best results, respectively. Asterisk indicates statistical significance against each and every baseline with the corresponding column.

C.5 Full Results

Table 6 reports experimental results on `Qwen3-32B` using other X-ICL baselines, controlling the combinations of few-shot demonstrations setting and instruction setting, where **CSICL** outperforms all X-ICL baselines.

Tables 7, 8, 9, and 10 present experimental results of X-ICL approaches on Global MMLU using Qwen 3, DeepSeek 3.1, Grok 4, and Gemini 2.5.

Method	X-ICL setting		En	Tgt.*	Unseen Lang.		
	Demonstration	Instruction			High*	Mid*	Low*
Few-shot learning	✓ Monolingual (Tgt.)	✓ Translation (Tgt.→En)	**88.8**	75.3	87.2	63.3	41.8
	✓ Monolingual (Tgt.)	✓ Translation (Tgt.→Rnd.)	88.6	74.8	87.5	63.5	42.0
	✓ Parallel	✓ Translation (Tgt.→En)	88.7	75.0	87.4	64.0	43.2
CSICL	✓ Gradual CS (Tgt.→En)	✓ Gradual Translation (Tgt.→En)	88.6	**76.8**	**87.8**	**64.9**	**46.0**

Table 6: Ablation results of other X-ICL baselines using Qwen 3, controlling the few-shot demonstrations setting and the instruction setting. Bold and underline denote the best and the second-best results, respectively. Tgt. and Rnd. denote a target language and a random language, respectively. Asterisk indicates statistical significance against each and every baseline within the corresponding column.

X-ICL setting	En	Tgt.*	Zh	Es*	Id*	Tr*	Sw*	Te*
Monolingual (En)	88.6	83.5	89.5	83.2	64.8	60.5	44.6	38.2
Monolingual (Tgt.)	**88.7**	83.3	90.3	83.8	64.3	59.9	42.8	36.9
Parallel	88.5	83.7	90.4	82.9	65.1	60.6	44.7	38.0
Translation (Tgt.→En)	**88.7**	83.6	**90.5**	83.6	65.5	61.1	45.3	38.8
Translation (Rnd.→En)	**88.7**	83.8	90.2	83.7	66.0	61.2	45.1	39.0
CSICL (Tgt.→En)	88.5	**84.9**	**90.5**	**83.9**	**66.6**	**61.8**	**48.4**	**42.8**

(a) Target: French (*high*)

X-ICL setting	En	Tgt.*	Zh	Es	Id*	Tr*	Sw*	Te*
Monolingual (En)	**88.9**	75.4	89.3	83.7	65.1	60.8	44.2	37.9
Monolingual (Tgt.)	88.7	77.2	90.0	83.0	64.8	61.3	42.3	35.7
Parallel	**88.9**	75.6	89.7	83.6	64.9	61.5	44.4	37.6
Translation (Tgt.→En)	**88.9**	79.4	90.2	84.1	65.5	61.4	44.8	38.6
Translation (Rnd.→En)	88.5	79.1	90.0	84.3	65.2	61.2	45.1	38.3
CSICL (Tgt.→En)	88.6	**80.6**	**90.3**	**84.4**	**67.4**	**62.8**	**48.5**	**41.7**

(b) Target: Korean (*mid*)

X-ICL setting	En	Tgt.*	Zh*	Es*	Id*	Tr*	Sw*	Te*
Monolingual (En)	88.9	53.5	89.8	83.5	64.8	60.8	44.6	37.7
Monolingual (Tgt.)	**89.0**	55.5	87.9	86.4	63.9	58.4	43.5	31.0
Parallel	88.7	58.8	90.4	85.6	64.2	61.7	48.8	34.9
Translation (Tgt.→En)	88.8	60.5	90.3	85.7	64.9	63.8	50.0	34.5
Translation (Rnd.→En)	88.6	58.5	90.6	86.2	65.0	64.2	44.9	41.4
CSICL (Tgt.→En)	88.7	**64.9**	**90.8**	**86.9**	**66.1**	**64.7**	**51.2**	**43.4**

(c) Target: Yoruba (*low*)

Table 7: Full experimental results comparing **CSICL** to X-ICL baselines using `Qwen3-32B`. Bold and underline denote the best and the second-best results, respectively. Tgt. and Rnd. denote a target language and a random language, respectively. Asterisk indicates statistical significance against each and every baseline within the corresponding column.

X-ICL setting	En	Tgt.*	Zh	Es	Id*	Tr*	Sw*	Te*
Monolingual (En)	**89.6**	84.6	91.2	84.3	66.2	61.5	45.3	38.8
Monolingual (Tgt.)	89.5	84.4	91.8	84.9	65.7	61.0	43.6	37.5
Parallel	89.4	84.7	92.0	84.0	66.5	61.7	45.4	38.6
Translation (Tgt.→En)	**89.6**	84.8	92.1	84.8	66.8	62.1	46.0	39.2
Translation (Rnd.→En)	**89.6**	85.0	91.9	84.9	67.2	62.2	45.8	39.3
CSICL (Tgt.→En)	89.4	**86.0**	**92.3**	**85.1**	**68.0**	**62.9**	**49.1**	**43.1**

(a) Target: French (*high*)

X-ICL setting	En	Tgt.*	Zh	Es	Id*	Tr*	Sw*	Te*
Monolingual (En)	**89.8**	77.3	91.0	84.6	66.4	61.8	45.0	38.6
Monolingual (Tgt.)	89.6	79.1	91.6	83.9	66.0	62.3	43.2	36.5
Parallel	**89.8**	77.5	91.4	84.5	66.1	62.5	45.2	38.3
Translation (Tgt.→En)	**89.8**	81.1	91.8	85.0	66.7	62.4	45.6	39.0
Translation (Rnd.→En)	89.4	80.8	91.5	85.1	66.4	62.2	45.9	38.7
CSICL (Tgt.→En)	89.5	**82.4**	**92.0**	**85.3**	**68.6**	**63.6**	**49.2**	**41.7**

(b) Target: Korean (*mid*)

X-ICL setting	En	Tgt.*	Zh	Es*	Id*	Tr	Sw*	Te*
Monolingual (En)	89.8	55.3	91.5	84.7	66.3	61.9	45.2	38.2
Monolingual (Tgt.)	**89.9**	57.3	89.6	87.4	65.5	59.5	44.1	32.0
Parallel	89.7	60.6	92.0	86.6	65.8	62.7	49.3	35.9
Translation (Tgt.→En)	89.8	62.3	91.9	86.7	66.6	64.5	50.5	35.6
Translation (Rnd.→En)	89.6	60.2	92.3	87.2	66.7	65.0	45.4	42.2
CSICL (Tgt.→En)	89.7	**66.7**	**92.5**	**87.8**	**67.7**	**65.4**	**51.6**	**44.2**

(c) Target: Yoruba (*low*)

Table 8: Full experimental results comparing **CSICL** to X-ICL baselines using `DeepSeek-chat-v3.1`. Bold and underline denote the best and the second-best results, respectively. Tgt. and Rnd. denote a target language and a random language, respectively. Asterisk indicates statistical significance against each and every baseline within the corresponding column.

X-ICL setting	En	Tgt.*	Zh	Es	Id*	Tr*	Sw*	Te*
Monolingual (En)	**88.3**	83.0	89.8	83.6	65.0	60.3	44.0	37.5
Monolingual (Tgt.)	88.1	82.9	90.5	84.1	64.6	60.6	42.2	36.2
Parallel	88.2	83.3	90.6	83.4	65.2	60.9	44.2	37.2
Translation (Tgt.→En)	**88.3**	83.2	90.7	84.0	65.6	61.3	44.7	37.9
Translation (Rnd.→En)	88.1	83.5	90.4	84.1	65.7	61.3	44.6	38.1
CSICL (Tgt.→En)	88.0	**84.5**	**90.8**	**84.3**	**66.5**	**61.9**	**47.7**	**41.7**

(a) Target: French (*high*)

X-ICL setting	En	Tgt.*	Zh	Es	Id*	Tr*	Sw*	Te*
Monolingual (En)	**88.4**	74.6	89.6	83.9	65.3	60.6	43.7	37.2
Monolingual (Tgt.)	88.2	76.4	90.2	83.2	65.0	61.1	41.9	35.1
Parallel	**88.4**	78.6	89.9	83.8	65.1	61.3	44.0	36.9
Translation (Tgt.→En)	**88.4**	74.8	90.4	84.2	65.7	61.2	44.3	37.8
Translation (Rnd.→En)	88.1	78.3	90.1	84.4	65.4	61.1	44.6	37.5
CSICL (Tgt.→En)	88.2	**79.7**	**90.6**	**84.6**	**67.1**	**62.3**	**47.8**	**40.8**

(b) Target: Korean (*mid*)

X-ICL setting	En	Tgt.*	Zh	Es*	Id*	Tr*	Sw*	Te*
Monolingual (En)	88.4	52.7	90.1	83.8	65.1	60.7	43.9	36.9
Monolingual (Tgt.)	**88.5**	54.6	88.2	86.7	64.2	58.2	42.8	30.2
Parallel	88.2	59.5	90.7	85.9	64.7	61.4	47.9	34.3
Translation (Tgt.→En)	88.3	57.8	90.6	86.0	65.4	63.5	49.1	34.0
Translation (Rnd.→En)	88.1	57.6	90.9	86.5	65.5	63.8	44.0	40.6
CSICL (Tgt.→En)	88.2	**63.7**	**91.0**	**87.0**	**66.5**	**64.2**	**50.3**	**42.5**

(c) Target: Yoruba (*low*)

Table 9: Full experimental results comparing **CSICL** to X-ICL baselines using `grok-4-fast`. Bold and underline denote the best and the second-best results, respectively. Tgt. and Rnd. denote a target language and a random language, respectively. Asterisk indicates statistical significance against each and every baseline within the corresponding column.

X-ICL setting	En	Tgt.*	Zh	Es	Id*	Tr*	Sw*	Te*
Monolingual (En)	**90.2**	85.6	91.5	85.1	67.2	62.7	46.1	39.6
Monolingual (Tgt.)	90.0	85.4	92.1	85.7	66.7	62.3	44.3	38.3
Parallel	90.1	85.8	92.2	84.9	67.5	63.0	46.3	39.4
Translation (Tgt.→En)	**90.2**	85.9	92.3	85.6	67.9	63.4	46.9	40.1
Translation (Rnd.→En)	90.1	86.1	92.1	85.7	68.3	63.5	46.8	40.3
CSICL (Tgt.→En)	90.0	**87.3**	**92.5**	**85.9**	**69.2**	**64.1**	**49.9**	**44.3**

(a) Target: French (*high*)

X-ICL setting	En	Tgt.*	Zh	Es*	Id*	Tr*	Sw*	Te*
Monolingual (En)	**90.3**	79.3	91.3	85.5	67.6	63.0	45.8	39.4
Monolingual (Tgt.)	90.1	81.1	91.9	84.8	67.2	63.5	44.0	37.3
Parallel	**90.3**	79.5	91.6	85.4	67.3	63.7	46.0	39.1
Translation (Tgt.→En)	**90.3**	83.0	92.0	85.9	67.9	63.6	46.4	39.9
Translation (Rnd.→En)	90.0	82.7	91.7	86.1	67.6	63.5	46.7	39.6
CSICL (Tgt.→En)	90.1	**84.3**	**92.2**	**86.4**	**69.7**	**64.8**	**50.1**	**42.8**

(b) Target: Korean (*mid*)

X-ICL setting	En	Tgt.*	Zh	Es	Id*	Tr*	Sw*	Te*
Monolingual (En)	90.3	56.8	91.8	85.4	67.4	63.1	46.0	39.0
Monolingual (Tgt.)	**90.4**	58.9	89.8	88.3	66.6	60.8	44.9	31.7
Parallel	90.2	62.1	92.4	87.5	66.9	63.8	50.1	35.6
Translation (Tgt.→En)	90.3	63.9	92.3	87.6	67.7	65.7	51.3	35.3
Translation (Rnd.→En)	90.1	61.8	92.6	88.1	67.8	66.2	46.3	42.0
CSICL (Tgt.→En)	90.2	**66.2**	**92.8**	**88.6**	**68.9**	**66.6**	**52.5**	**44.8**

(c) Target: Yoruba (*low*)

Table 10: Full experimental results comparing **CSICL** to X-ICL baselines using `Gemini 2.5 Flash`. Bold and underline denote the best and the second-best results, respectively. Tgt. and Rnd. denote a target language and a random language, respectively. Asterisk indicates statistical significance against each and every baseline within the corresponding column.