Lost in the Mix: Evaluating LLM Understanding of Code-Switched Text

Amr Mohamed[1,2†], Yang Zhang[2], Michalis Vazirgiannis[1,2], Guokan Shang[1†]

[1]MBZUAI, [2]Ecole Polytechnique

[†]Correspondence: {amr.mohamed, guokan.shang}@mbzuai.ac.ae

Abstract

Code-switching (CSW) is the act of alternating between two or more languages within a single discourse. This phenomenon is widespread in multilingual communities and increasingly prevalent online, exposing large language models (LLMs) to mixed-language inputs. We present a systematic evaluation of LLM *comprehension* under code-switching by generating linguistically grounded CSW variants of established benchmarks (Belebele, MMLU, XNLI) across five typologically diverse languages. Our contributions are: (i) a controlled pipeline for producing CSW test sets that respect linguistic constraints on code-switching; (ii) a multi-model, multi-language analysis showing that inserting non-English tokens into English consistently reduces accuracy on comprehension and reasoning benchmarks, whereas embedding English into non-English contexts often improves it; and (iii) a mitigation study contrasting in-context learning (ICL) with fine-tuning. Across model families, ICL cues yield inconsistent, and sometimes negative, effects, while fine-tuning on CSW data provides modest but reliable gains, partially recovering accuracy under CSW.

1 Introduction

Code-switching (CSW)—the act of alternating between two or more languages within a single discourse (Das et al., 2023; Zhang et al., 2023; Ochieng et al., 2024)—is a common phenomenon in multilingual communities (Bullock and Toribio, 2009; Parekh et al., 2020; Doğruöz et al., 2021), and increasingly prevalent in online content (Kodali et al., 2024), where users naturally mix languages in everyday informal communications.

Large Language Models (LLMs) have demonstrated remarkable capabilities across a wide range of natural language processing tasks (Zhao et al., 2023). As they are increasingly used to process and generate content, the widespread availability of code-switched inputs makes it crucial to understand how LLMs reason about such mixed-language data, and whether their multilingual fluency reflects genuine understanding or superficial pattern matching (Zhang et al., 2023). To systematically assess LLMs' handling of such data, we turn to insights from linguistic theories that define the structural constraints governing natural CSW.

Linguistic theories of code-switching offer valuable guidance on where and how language mixing naturally occurs. Frameworks such as the Equivalence Constraint Theory (Poplack, 1978) and the Matrix Language Frame model (Myers-Scotton, 1993) describe grammatical boundaries that make mixed-language sentences well-formed. We draw on their core insights to guide controlled, linguistically plausible code-switching in our benchmarks, ensuring that our synthetic data adheres to natural switching patterns while remaining interpretable for systematic evaluation.

Despite extensive linguistic work on code-switching, existing evaluation benchmarks fail to leverage these insights to assess deeper comprehension in mixed-language contexts. Most current CSW benchmarks focus on surface-level tasks, such as language identification, sentiment analysis, or part-of-speech tagging, providing limited visibility into models' reasoning and semantic understanding (Khanuja et al., 2020; Aguilar et al., 2020; Patwa et al., 2020). Recent efforts have begun probing this gap (Yadav et al., 2024; Gupta et al., 2024; Ng and Chan, 2024), yet a systematic evaluation of LLM comprehension under controlled code-switching remains missing.

To fill this gap, we propose a systematic evaluation framework that uses a constrained, multi-step LLM pipeline to generate linguistically grounded, code-switched variants of established reasoning and comprehension benchmarks—*Belebele*, *MMLU*, and *XNLI*. Our approach draws on linguistic theories to ensure natural switch

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 44922–44938

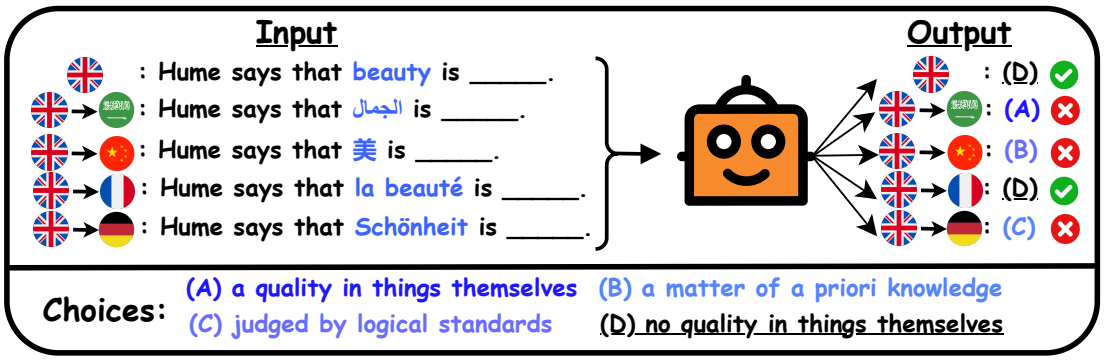

Figure 1: Illustration of the noun-token code-switching procedure used in Experiment 1. Each example replaces the English noun beauty with its aligned translation from a different embedded language (Arabic, Chinese, French, German) while keeping English as the matrix language. The model's varied answers across versions reveal how inserting foreign tokens into English sentences can alter its reasoning and output consistency.

points while maintaining grammatical coherence, enabling controlled studies of how switching direction and language pair affect understanding. Code and data are publicly available[1]

Our experiments show that the effect of code-switching on LLM comprehension is systematic but asymmetric (Figure 1):

• Code-switching into English text consistently reduces task accuracy, even when switches comply with linguistic constraints, revealing a structural sensitivity to foreign insertions beyond token-level unfamiliarity.

• Embedding English tokens into non-English text often enhances comprehension, particularly for models less proficient in the matrix language, highlighting the facilitative role of English in mixed-language contexts.

• In-context learning (ICL) cues yield inconsistent gains across model families, whereas fine-tuning on synthetic code-switched data provides reliable improvements, partially recovering accuracy under code-switching.

Our findings expose a persistent asymmetry in how LLMs process mixed-language text. When non-English tokens interrupt English sentences, model performance consistently declines; when English intrudes into other languages, comprehension often improves. This asymmetry reveals a structural bias toward English, an imprint of data imbalance in multilingual pretraining, and points to a broader challenge of linguistic equity as LLMs increasingly shape the global information ecosystem.

[1]https://github.com/amr-mohamedd/Lost-in-the-Mix

2 Related Work

Code-Switching in Language Models. Early multilingual encoders such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) achieved strong monolingual performance but struggled with code-switched input (Winata et al., 2021a). This limitation led to specialized architectures and training regimes tailored for mixed-language text (Winata et al., 2019; Liu et al., 2020; Winata et al., 2021b). Benchmarks like GLUE-CoS (Khanuja et al., 2020) enabled progress, yet most research remained confined to encoder-centric models (Tan and Joty, 2021; Zhu et al., 2023). Decoder-only models—now central to state-of-the-art LLMs—have received far less attention in this context. While a few studies explored adversarial code-mixing in autoregressive architectures (Das et al., 2022), systematic evaluation of their comprehension under controlled, linguistically valid switching has been lacking.

Code-Switched Text Generation. Synthetic code-switched text generation is key for augmenting multilingual data, and probing model robustness (Pratapa et al., 2018; Zhang et al., 2023). Approaches span from linguistically motivated frameworks—such as the Equivalence Constraint Theory (ECT) (Poplack, 1978), and Matrix Language Frame (MLF) model (Myers-Scotton, 1993), to heuristic or random token substitutions (Myslín, 2014; and, 2018; Chan et al., 2024). Alignment-based methods (Kuwanto et al., 2024) generate fluent outputs, but often overlook broader syntactic or contextual coherence. Recent work has begun ex-

ploiting LLMs' capacity for context-aware, linguistically grounded switching, via constrained prompting (Potter and Yuan, 2024), fine-tuning (Heredia et al., 2025), or hybrid pipelines (Kuwanto et al., 2024). Yet, reliable control over switch placement, scalability across language pairs, and rigorous evaluation of resulting text quality remain open challenges. Our work advances this line by employing modern LLMs in a controlled, theory-informed pipeline that produces high-quality, linguistically constrained code-switched benchmarks for systematic comprehension testing.

Evaluation of LLM CSW Capabilities. Existing evaluations of code-switching competence largely target surface-level tasks—language identification, sentiment, or part-of-speech tagging—through benchmarks such as GLUECoS (Khanuja et al., 2020), LINCE (Aguilar et al., 2020), and SemEval (Patwa et al., 2020). While informative, these tasks assess recognition rather than reasoning. A few recent studies extend to question answering and sentiment classification (Winata et al., 2021a; Huzaifah et al., 2024), but remain narrow in scope and metrics. In contrast, we evaluate deeper comprehension and reasoning under linguistically controlled code-switching, using established multi-domain benchmarks (*Belebele*, *MMLU*, *XNLI*). This shift enables a systematic analysis of how switching direction, language pair, and linguistic grounding jointly shape LLM comprehension under code-switching.

3 Methodology

3.1 Notations

$$\mathcal{B} = \{B_p\}_{p=1}^{P}$$

be a set of P standard benchmarks. Let

$$\mathcal{L} = \{l_j\}_{j=1}^{L}$$

be a set of L languages from which the matrix and embedded languages are selected for code-switched benchmarks generation. Let

$$\mathcal{M} = \{m_k\}_{k=1}^{K}$$

be a set of K LLMs.

To evaluate the performance of an LLM $m_k \in \mathcal{M}$ on code-switched text comprehension, we generate a code-switched version of benchmark $B_p \in \mathcal{B}$ using a single matrix language $l_{\text{matrix}} \in \mathcal{L}$ and a set of embedded languages $\mathcal{L}_{\text{embedded}}$, where $\mathcal{L}_{\text{embedded}} \subseteq \mathcal{L} \setminus l_{\text{matrix}}$ and $|\mathcal{L}_{\text{embedded}}| \geq 1$, which we denote by $B_p^{l_{\text{matrix}} \to \mathcal{L}_{\text{embedded}}}$.

3.2 Constructing Code-Switched Inputs

We operationalize code-switching through two input construction strategies that differ in their linguistic grounding. **(i) Noun-token CSW:** nouns in the matrix-language text are replaced with their aligned counterparts from a parallel sentence in the embedded language. Substitutions are applied only when they adhere to standard CSW constraints—the Equivalence Constraint Theory and the Matrix Language Frame model—keeping the matrix language as the grammatical frame (Poplack, 1978; Myers-Scotton, 1993). **(ii) Ratio-token CSW:** a fixed proportion of tokens ($\approx 20\%$) are replaced irrespective of syntactic structure, following prior heuristic setups (Chan et al., 2024).

An example of the noun-token method across embedded languages is shown in Figure 2.

Generation pipeline used in main experiments. Given each matrix/embedded parallel pair from the selected benchmarks, we use a two-step LLM-centric procedure: (1) identify and mask candidate switch points (nouns for the noun-token setting; random tokens for the ratio-token setting) under the stated constraints; (2) fill each placeholder with the aligned counterpart from the embedded-language sentence, adjusting inflection as needed so the sentence remains well-formed and semantically faithful.

Our choice of the LLM-centric pipeline follows a comparative evaluation against an alignment-first alternative (AWESOME (Dou and Neubig, 2021) + LaBSE (Feng et al., 2022) with POS-guided substitution via Stanza (Qi et al., 2020)). Across language pairs, bilingual annotators preferred the LLM-centric outputs in manual checks, and LLM-as-a-Judge results corroborated this trend (Zheng et al., 2023). We therefore adopt it for all main experiments. Full prompts, alignment settings, quality-control procedures, comparative results, and dataset-level statistics on token replacement ratios and switch densities are provided in Appendices B and C.

3.3 Evaluation Metrics

We report three standard metrics.

Accuracy. For each benchmark B, accuracy is the proportion of correctly answered questions.

Weighted average accuracy. To summarize performance across benchmarks, we compute a size-

Questions:

EN: Find the degree for the given field extension $\mathbb{Q}(\sqrt{2}, \sqrt{3}, \sqrt{18})$ over $\mathbb{Q}$.

EN→AR: Find the درجة for the given امتداد الحقل $\mathbb{Q}(\sqrt{2}, \sqrt{3}, \sqrt{18})$ over $\mathbb{Q}$.

EN→FR: Find the degré for the given extension de champ $\mathbb{Q}(\sqrt{2}, \sqrt{3}, \sqrt{18})$ over $\mathbb{Q}$.

EN→DE: Find the Grad for the given Felderweiterung $\mathbb{Q}(\sqrt{2}, \sqrt{3}, \sqrt{18})$ over $\mathbb{Q}$.

EN→ZH: Find the 次 for the given 域 $\mathbb{Q}(\sqrt{2}, \sqrt{3}, \sqrt{18})$ over $\mathbb{Q}$.

Choices: 0, 4, 2, 6

Answer: 2

Figure 2: Example of the noun-token code-switching method applied to an MMLU question (*Abstract Algebra*). Embedded tokens from Arabic, French, German, and Chinese are shown in blue, illustrating aligned noun replacements within an English matrix sentence while preserving grammatical structure and meaning.

weighted mean:

$$\mathrm{Acc}_{\text{weighted}}(m_k, l_{\text{matrix}}, \mathcal{L}_{\text{embedded}}) = \frac{\sum_{B_p \in \mathcal{B}} |B_p| \cdot \mathrm{Acc}(m_k, B_p^{l_{\text{matrix}} \to \mathcal{L}_{\text{embedded}}})}{\sum_{B_p \in \mathcal{B}} |B_p|}, \quad (1)$$

Accuracy delta. The CSW impact for a benchmark B is

$$\Delta\mathrm{Acc}(m_k, B_p^{l_{\text{matrix}} \to \mathcal{L}_{\text{embedded}}}) = \mathrm{Acc}(m_k, B_p^{l_{\text{matrix}} \to \mathcal{L}_{\text{embedded}}}) - \mathrm{Acc}(m_k, B_p). \quad (2)$$

with negative values indicating degradation under code-switching. Unless noted, tables show weighted averages; per-benchmark results appear in the figures and appendix.

4 Experimental Setting

Language selection. We consider the set

$$\mathcal{L} = \{\text{English}, \text{Arabic}, \text{German}, \text{French}, \text{Chinese}\},$$

chosen to represent languages with varying degrees of semantic, lexical, and syntactic similarity to English. Prior work has shown that such typological diversity can influence cross-lingual transfer and degradation patterns in multilingual modeling and translation (Guerin et al., 2024; Mohamed et al., 2025). This selection therefore enables controlled analysis of how linguistic distance and representation coverage affect model robustness under CSW.

Model selection. We evaluate LLaMA 3.2 Instruct (3B) and LLaMA 3.1 Instruct (8B, 70B) (Grattafiori et al., 2024), Qwen 2.5 Instruct (3B, 7B, 72B) (Yang et al., 2025), Mistral 7B Instruct (v0.3) (Albert et al., 2023), and ALLaM 7B (Bari et al., 2024), covering a broad range of scales, architectures, and pretraining curricula. *Allam* represents the state of the art among Arabic-focused LLMs, while *Qwen* demonstrates strong coverage of Chinese and other languages. The *LLaMA* and *Mistral* families provide robust multilingual baselines with balanced proficiency across high-resource languages. This diversity allows us to analyze how architecture type and model scale influence robustness to code-switching.

Benchmark selection. We evaluate LLM comprehension using three well-established tasks adapted into linguistically controlled code-switched variants. Specifically, we use *Belebele* (Bandarkar et al., 2023) for passage-level reading comprehension (with both passages and questions code-switched), *MMLU*[2] (Hendrycks et al., 2020) for broad-domain multiple-choice reasoning (code-switching applied to questions), and *XNLI* (Conneau et al., 2018) for natural language inference (both premise and hypothesis code-switched). All non-English content is drawn from the official parallel translations accompanying these benchmarks, ensuring one-to-one alignment with the English source data rather than region-specific or independently translated variants. This guarantees semantic fidelity across languages and supports consistent evaluation of switching direction and language pair effects. To standardize evaluation across models, we adapt EleutherAI's Language Model Evaluation Harness (Gao et al., 2024) for our code-switched benchmarks.

Full implementation details (hardware, GPU-hours, and evaluation harness details) are provided

[2]https://huggingface.co/datasets/openai/MMMLU

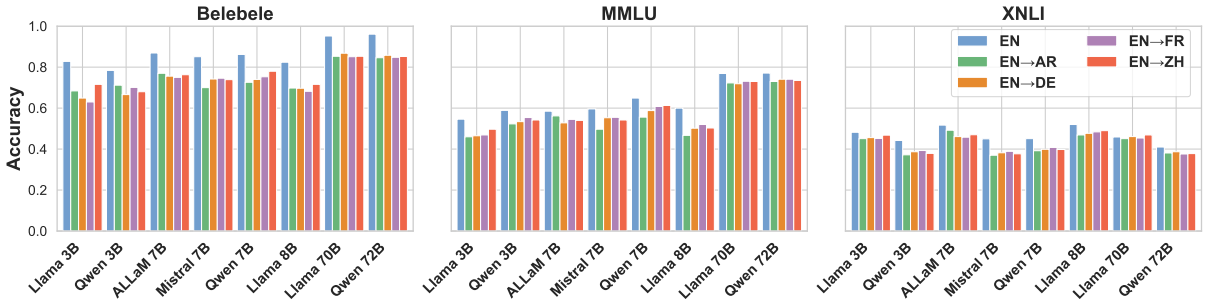

Figure 3: Comparison of LLM accuracy on monolingual English benchmarks versus their noun-token code-switched variants. English serves as the matrix language, with Arabic (EN→AR), Chinese (EN→ZH), French (EN→FR), and German (EN→DE) as embedded languages.

Model	EN→AR	EN→DE	EN→FR	EN→ZH	EN	Placeholder
Llama 3B	0.47	0.47	0.47	0.50	**0.54**	0.41
Qwen 3B	0.49	0.50	0.52	0.51	**0.56**	0.44
Allam 7B	0.55	0.52	0.53	0.53	**0.58**	0.48
Mistral 7B	0.47	0.52	0.52	0.51	**0.57**	0.46
Qwen 7B	0.52	0.55	0.56	0.57	**0.61**	0.47
Llama 8B	0.48	0.51	0.52	0.51	**0.59**	0.43
Llama 70B	0.66	0.66	0.67	0.67	**0.70**	0.57
Qwen 72B	0.65	0.66	0.65	0.65	**0.69**	0.59

Table 1: Weighted average accuracy of LLMs on noun-token code-switched benchmarks compared to the monolingual English and placeholder baselines. Cell colors indicate relative performance (green = higher, red = lower). The placeholder baseline isolates structural disruption without foreign semantics.

in Appendix A.

5 Experiments

We conduct two complementary experiments to analyze how linguistic grounding and switch structure affect LLM comprehension under code-switching. The first evaluates linguistically constrained noun substitutions; the second uses random replacements independent of syntax. To isolate structural disruption from semantic effects, we include a *placeholder baseline* where all switch points are replaced with a neutral token ("####") instead of foreign words.

5.1 Experiment 1: Linguistically Grounded Code-Switching

Setup. English serves as the matrix language (l_{matrix}). For each non-English language $l \in \mathcal{L} \setminus l_{\text{matrix}}$, we construct code-switched variants of all benchmarks using the noun-token method (§3.2), in which aligned nouns from the embedded language replace their English counterparts while preserving grammatical structure and sentence meaning. Models are evaluated on the original English versions, the noun-token CSW variants, and the placeholder baseline to disentangle the impact of mixed-language semantics from mere surface disruption.

Results. Figure 3 and Table 1 summarize results across models and languages. All models show consistent degradation when foreign nouns are embedded into English text, confirming that even linguistically valid switches impair comprehension. The extent of degradation correlates with model scale and language coverage: larger models (*LLaMA 70B*, *Qwen 72B*) are most robust, while smaller ones exhibit larger relative losses. *Allam 7B* remains comparatively strong for Arabic due to its language-focused pretraining. Importantly, all models outperform the placeholder baseline, indicating that meaningful foreign tokens are less harmful than nonsensical replacements and carry recoverable semantic information.

5.2 Experiment 2: Random Code-Switching Without Linguistic Constraints

Setup. This experiment follows the same setup as Experiment 1 but replaces linguistically valid noun substitutions with random token swaps ($\approx 20\%$ of tokens per sentence), disregarding syntactic

Model	EN→AR	EN→DE	EN→FR	EN→ZH	EN
Llama 3B	0.47	0.43	0.46	0.51	**0.54**
Qwen 3B	0.50	0.51	0.52	0.51	**0.56**
Allam 7B	0.56	0.51	0.53	0.54	**0.58**
Mistral 7B	0.49	0.52	0.53	0.52	**0.57**
Qwen 7B	0.53	0.55	0.56	0.57	**0.61**
Llama 8B	0.50	0.52	0.53	0.54	**0.59**
Llama 70B	0.68	0.67	0.68	0.68	**0.70**
Qwen 72B	0.66	0.66	0.66	0.66	**0.69**

Table 2: Weighted average accuracy of LLMs on ratio-token (random) code-switched benchmarks compared to the monolingual English baseline. Cell colors indicate relative performance from highest (green) to lowest (red).

Model	AR→EN		DE→EN		FR→EN		ZH→EN	
	Orig	CSW	Orig	CSW	Orig	CSW	Orig	CSW
LLaMA 3B	0.37	**0.45**	0.35	**0.38**	0.43	**0.45**	0.42	**0.47**
Qwen 3B	0.40	**0.48**	0.49	**0.52**	0.50	**0.53**	*0.48*	*0.48*
Allam 7B	0.51	**0.52**	0.39	**0.43**	0.49	**0.52**	0.44	**0.51**
Mistral 7B	0.35	**0.48**	0.50	**0.54**	0.52	**0.55**	0.46	**0.53**
Qwen 7B	0.47	**0.52**	0.51	**0.53**	0.56	**0.57**	**0.56**	0.55
LLaMA 8B	0.38	**0.44**	*0.50*	*0.50*	0.50	**0.52**	0.49	**0.53**
LLaMA 70B	0.61	**0.66**	0.67	**0.67**	*0.68*	*0.68*	0.64	**0.66**
Qwen 72B	0.63	**0.66**	*0.68*	*0.68*	*0.68*	*0.68*	*0.66*	*0.66*

Table 3: Weighted average accuracy of LLMs on monolingual (Orig) versus English-embedded code-switched (CSW) benchmarks across Arabic, German, French, and Chinese. **Bold** indicates improvement ($\Delta > 0$); *Italic* indicates no change. Deltas are computed relative to Orig.

structure (§3.2). This configuration tests whether degradation primarily stems from grammatical disruption or from the general presence of mixed-language tokens.

Results. Table 2 presents the outcomes. All models again show performance degradation relative to the monolingual baseline, consistent with Experiment 1. The overall magnitude of loss is similar to the noun-token setting, indicating that linguistic well-formedness only partly mitigates CSW difficulty. Smaller models are most affected (*LLaMA 3B*, -0.11 on EN→DE), while larger ones (*LLaMA 70B*, *Qwen 72B*) remain more stable ($\Delta \approx -0.02$). The persistence of comparable declines across both structured and random mixing suggests a broader limitation in LLMs' ability to integrate multilingual input, independent of syntactic plausibility.

6 Ablations

Building on Section 5, which found comparable degradation from noun-token and ratio-token CSW, we conduct two additional ablations to test (i) how switching direction affects comprehension, and (ii) whether model robustness persists under extreme multilingual mixing. All ablations use the linguistically grounded noun-token method described in §3.2, with switch density fixed at roughly 30% of content nouns per sentence for consistency across languages.

6.1 English as an Embedded Language

To assess whether embedding English improves comprehension in other matrix languages, we reverse the language roles from the main experiments. Each non-English language in $\mathcal{L} \setminus l_{\text{matrix}}$ serves as the matrix language, while English becomes the embedded language. Code-switched variants ($B_p^{l_{\text{matrix}} \to \{\text{English}\}}$) of *Belebele*, *MMLU*, and *XNLI* are generated using their official multilingual parallel versions rather than region-specific or independently translated variants, ensuring one-to-one alignment with the English originals. This guarantees that inserted English tokens correspond exactly to their aligned nouns in the parallel English item rather than being back-translated. All ablation datasets are produced with the same LLM-centric pipeline vetted for linguistic quality (§3.2).

Embedding English into lower-resource matrix languages consistently improves comprehension, whereas effects are minimal for languages already well represented in training corpora. For example, *Mistral 7B* gains +0.13 in Arabic (0.35→0.48) and +0.07 in Chinese, but only +0.03 in French, while high-capacity models (*LLaMA 70B*, *Qwen 72B*) show near-parity between monolingual and CSW conditions. These asymmetric patterns support the view that improvements arise from greater data familiarity with English rather than typological distance. Qualitative examples of reverse direction code-switching are included in Appendix B.

6.2 Extreme Multilingual Mixing

To test robustness under multi-language mixing, we create "extreme" CSW variants of the *MMLU* benchmark with English as the matrix language and multiple embedded languages per sentence. We define three settings: **(S1)** non-Latin scripts ({Arabic, Chinese}), **(S2)** Latin scripts ({French, German}),

Model	S1	S2	S3	EN
Llama 3B	0.48	0.46	0.47	**0.55**
Qwen 3B	0.54	0.55	0.53	**0.59**
Allam 7B	0.56	0.54	0.54	**0.58**
Mistral 7B	0.53	0.56	0.55	**0.59**
Qwen 7B	0.58	0.60	0.59	**0.65**
Llama 8B	0.49	0.51	0.49	**0.60**
Llama 70B	0.72	0.70	0.70	**0.77**
Qwen 72B	0.74	0.74	0.73	**0.77**

Table 4: *MMLU* accuracy for extreme CSW with English as the matrix language and the embedded languages being Arabic and Chinese (Setting 1), French and German (Setting 2), and Arabic, Chinese, French, and German (Setting 3), alongside the monolingual English baseline. The highest scores are indicated in bold.

Model	EN→AR	EN→DE	EN→FR	EN→ZH	EN
Llama 3B	0.31	0.34	0.32	0.32	**0.54**
Qwen 3B	0.51	0.53	0.54	0.53	**0.56**
Allam 7B	0.56	0.53	0.54	0.53	**0.58**
Mistral 7B	0.46	0.50	0.50	0.50	**0.57**
Qwen 7B	0.54	0.56	0.58	0.59	**0.61**
Llama 8B	0.41	0.47	0.48	0.47	**0.59**
Llama 70B	0.53	0.53	0.64	0.50	**0.70**
Qwen 72B	0.70	0.71	0.71	**0.72**	0.69

Table 5: Effect of ICL-based mitigation on noun-token CSW benchmarks. English serves as the matrix language; results are weighted-average accuracies. The highest scores are in **bold**.

and **(S3)** all four. Each instance borrows embedded words evenly from the specified set using the same linguistically constrained pipeline as in §3.2.

All models degrade moderately under extreme CSW, with drops of 0.03–0.07 relative to English. Script type (non-Latin vs Latin) has no consistent effect: *Allam 7B* performs slightly better with non-Latin (0.56 vs 0.54), whereas *Mistral 7B* favors Latin (0.56 vs 0.53). Even when all four languages are mixed, high-capacity models (*LLaMA 70B*, *Qwen 72B*) lose less than 0.07, indicating stable multilingual alignment. These findings suggest that degradation arises from representational interference rather than script complexity or language count.

7 Mitigation Strategies

To mitigate the performance declines induced by code-switching, we explore two complementary strategies: an **in-context learning (ICL)** approach, which prepends explicit cues identifying the input as mixed-language, and a **model-based** approach, which fine-tunes LLMs on synthetic CSW data.

7.1 ICL-Based Mitigation

Each noun-token CSW instance was preceded by a short meta-instruction indicating that the input text mixes English with another language and should be interpreted without translation. The goal is to test whether minimal contextual framing can help the model activate multilingual priors relevant for processing mixed-language inputs. The exact instructions used per benchmark are listed in Appendix D.

ICL cues yield mixed outcomes across model families (Table 5). Models with stronger multilingual pretraining (*Qwen*, *Allam*) show modest, consistent gains, occasionally matching or exceeding their English-only accuracy (e.g., *Qwen 72B* on EN→ZH),whereas *Llama* and *Mistral* are neutral to negative. Taken together, this pattern indicates that ICL mitigates CSW degradation primarily when robust multilingual representations are already present, rather than as a function of scale; in models with limited cross-lingual priors, the cue can instead disrupt processing.

7.2 Model-Based Mitigation

Fine-tuning on code-switched data offers a complementary approach by directly exposing the model to mixed-language patterns. We fine-tuned *Llama 8B*, a model that showed limited responsiveness to IC, on a small, linguistically constrained CSW corpus derived from TED Talk translations (Qi et al., 2018) spanning English, Arabic, Chinese, French, and German. We filtered to English sentences >70 words to emphasize multi-clause structure and increase potential switch sites, yielding roughly 3.6k examples per language pair. Applying the noun-token CSW process produced a total of 14.6k training samples. Each example instructed the model to generate the code-switched variant given the English and embedded-language sentences, using five paraphrased templates to maintain prompt diversity (Appendix E).

Instruction tuning on CSW data leads to consistent, though modest, performance recovery across all embedded languages (Figure 4). The largest improvement is observed for EN→AR, aligning with its lower representation in pretraining data. Notably, the model was fine-tuned *for generation of code-switched text*, yet evaluated on comprehension benchmarks. The observed gains there-

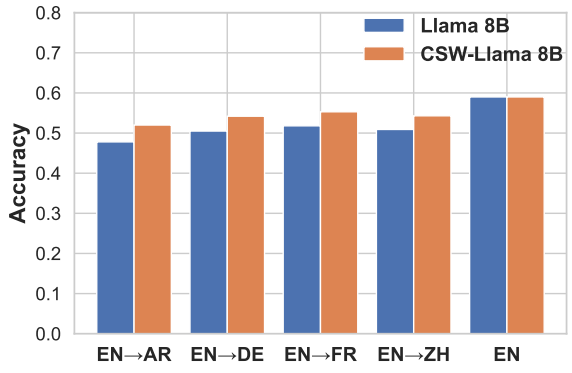

Figure 4: Comparison of *Llama 8B* and its CSW-tuned variant (*CSW-Llama 8B*) on English and code-switched benchmarks (*Belebele*, *MMLU*, *XNLI*). English is the matrix language; Arabic, Chinese, French, and German serve as embedded languages.

fore suggest that exposure to the structural patterns of CSW, rather than task-specific supervision, enhances the model's ability to interpret mixed-language input. Crucially, English-only accuracy remains stable, indicating no trade-off between monolingual and code-switched comprehension.

Compared to ICL, fine-tuning offers a more *stable and architecture-agnostic* mitigation: even small-scale exposure to code-switched text enables the model to internalize switch-boundary patterns and reduce sensitivity to linguistic mixing.

8 Discussion and Conclusion

This work provides the first systematic evaluation of LLM *comprehension* under linguistically controlled code-switching (CSW). By generating parallel, theory-grounded CSW variants of established reasoning and comprehension benchmarks, we revealed how language mixing interacts with multilingual pretraining, linguistic structure, and mitigation strategies. Our findings expose consistent asymmetries in how models process mixed-language text and point to concrete paths for improving robustness.

Code-switching disrupts comprehension in structurally predictable ways. When English serves as the matrix language, introducing tokens from other languages systematically degrades accuracy across models and tasks. This degradation persists even when switch points comply with linguistic theories such as the Equivalence Constraint and Matrix Language Frame models. The comparable losses between linguistically grounded and random substitutions indicate that the issue is not grammatical violation, but rather a representational limitation: LLMs trained on monolingual data struggle to integrate multilingual cues within a single syntactic frame.

The direction of switching reveals a structural bias toward English. Embedding English tokens into non-English text often improves performance, sometimes exceeding monolingual baselines, while inserting non-English tokens into English sentences consistently hurts comprehension. This asymmetry reflects the data imbalance in multilingual pretraining corpora: English serves as a dominant anchor representation that other languages map into, but not vice versa. Consequently, English functions as a *facilitative embedded language* rather than a robust matrix language, underscoring that LLM multilingualism remains uneven and data-driven rather than typologically general.

Mitigation requires representation-level adaptation, not only contextual cues. Our mitigation experiments compare lightweight in-context learning (ICL) cues with direct fine-tuning on CSW data. As shown in Table 5, ICL occasionally aids multilingual models with broad language coverage (*Qwen*, *Allam*), but fails, or even backfires, for models less aligned to instruction-following or cross-lingual processing (*Llama*, *Mistral*). In contrast, model-based instruction tuning on code-switched corpora (Figure 4) yields small yet consistent gains across all languages, improving robustness without harming monolingual performance. This suggests that structural adaptation to CSW distributions, rather than contextual re-instruction, is the more stable path to multilingual resilience.

Our results demonstrate that current LLMs exhibit systematic and direction-dependent vulnerabilities when processing mixed-language input. Linguistically controlled evaluation reveals that these failures arise not from superficial noise but from deeper representational asymmetries rooted in data imbalance. While ICL can sometimes compensate for multilingual exposure, stable comprehension of code-switched text ultimately demands explicit modeling of mixed-language structures during training. As LLMs continue to mediate global communication, ensuring equitable comprehension across languages requires moving beyond English-centric pretraining toward training paradigms that *natively include code-switching as a first-class linguistic setting*.

Limitations

While our study adopts a controlled evaluation setup for both linguistically and non-linguistically motivated code-switching, the noun-token approach we employ reflects one of the fundamental forms of linguistically grounded, naturalistic switching. However, more complex forms of code-switching may induce more severe performance degradation. Future work should investigate how higher-complexity switching patterns affect LLMs' understanding.

Additionally, in our non-linguistically motivated ratio-token experiments, the substitution rate was fixed at 20%. Exploring how variation in this ratio affects model behavior could yield a more nuanced understanding of the impact of non-linguistically grounded switching on LLM comprehension.

Finally, our benchmarks are synthetically constructed yet human-verified to ensure linguistic naturalness, and this controlled design enables reproducible analysis of how LLMs process mixed-language inputs, even if it does not capture all nuances of spontaneous human code-switching.

References

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. LinCE: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1803–1813, Marseille, France. European Language Resources Association.

Q Jiang Albert, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, and Devendra Singh Chaplot. 2023. Mistral 7b. *arXiv*.

Li Nguyen and. 2018. Borrowing or code-switching? traces of community norms in vietnamese-english speech. *Australian Journal of Linguistics*, 38(4):443–466.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint arXiv:2308.16884*.

M Saiful Bari, Yazeed Alnumay, Norah A Alzahrani, Nouf M Alotaibi, Hisham A Alyahya, Sultan AlRashed, Faisal A Mirza, Shaykhah Z Alsubaie, Hassan A Alahmed, Ghadah Alabduljabbar, et al. 2024. Allam: Large language models for arabic and english. *arXiv preprint arXiv:2407.15390*.

Barbara E. Bullock and Almeida Jacqueline Toribio. 2009. *The Cambridge Handbook of Linguistic Code-switching*. Cambridge Handbooks in Language and Linguistics. Cambridge University Press.

Kelvin Wey Han Chan, Christopher Bryant, Li Nguyen, Andrew Caines, and Zheng Yuan. 2024. Grammatical error correction for code-switched sentences by learners of English. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7926–7938, Torino, Italia. ELRA and ICCL.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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

Richeek Das, Sahasra Ranjan, Shreya Pathak, and Preethi Jyothi. 2023. Improving pretraining techniques for code-switched NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1176–1191, Toronto, Canada. Association for Computational Linguistics.

Sourya Dipta Das, Ayan Basak, Soumil Mandal, and Dipankar Das. 2022. Advcodemix: Adversarial attack on code-mixed data. In *Proceedings of the 5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD)*, pages 125–129.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1654–1666, Online. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. The language model evaluation harness.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Nicolas Guerin, Shane Steinert-Threlkeld, and Emmanuel Chemla. 2024. The impact of syntactic and semantic proximity on machine translation with back-translation. *arXiv preprint arXiv:2403.18031*.

Ayushman Gupta, Akhil Bhogal, and Kripabandhu Ghosh. 2024. Code-mixer ya nahi: Novel approaches to measuring multilingual llms' code-mixing capabilities. *arXiv preprint arXiv:2410.11079*.

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

Maite Heredia, Gorka Labaka, Jeremy Barnes, and Aitor Soroa. 2025. Conditioning llms to generate code-switched text: A methodology grounded in naturally occurring data. *arXiv preprint arXiv:2502.12924*.

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.

Pranjal Khanuja et al. 2020. Improving code-switched nlp using data augmentation. In *Proceedings of ACL 2020*, pages 1860–1871.

Prashant Kodali, Anmol Goel, Likhith Asapu, Vamshi Krishna Bonagiri, Anirudh Govil, Monojit Choudhury, Manish Shrivastava, and Ponnurangam Kumaraguru. 2024. From human judgements to predictive models: Unravelling acceptability in code-mixed sentences. *arXiv preprint arXiv:2405.05572*.

Garry Kuwanto, Chaitanya Agarwal, Genta Indra Winata, and Derry Tanti Wijaya. 2024. Linguistics theory meets llm: Code-switched text generation via equivalence constrained large language models. *arXiv preprint arXiv:2410.22660*.

Zihan Liu, Genta Indra Winata, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2020. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8433–8440.

Amr Mohamed, Mingmeng Geng, Michalis Vazirgiannis, and Guokan Shang. 2025. Llm as a broken telephone: Iterative generation distorts information. *arXiv preprint arXiv:2502.20258*.

R. Myers-Scotton. 1993. *Social Motivations for Code-Switching: Evidence from Africa*. Oxford University Press.

Mark Myslín. 2014. Codeswitching and predictability of meaning in discourse. In *Codeswitching and predictability of meaning in discourse*.

Lynnette Hui Xian Ng and Luo Qi Chan. 2024. What talking you?: Translating code-mixed messaging texts to english. *arXiv preprint arXiv:2411.05253*.

Millicent Ochieng, Varun Gumma, Sunayana Sitaram, Jindong Wang, Vishrav Chaudhary, Keshet Ronen, Kalika Bali, and Jacki O'Neill. 2024. Beyond metrics: evaluating llms' effectiveness in culturally nuanced, low-resource real-world scenarios. *arXiv preprint arXiv:2406.00343*.

Tanmay Parekh, Emily Ahn, Yulia Tsvetkov, and Alan W Black. 2020. Understanding linguistic accommodation in code-switched human-machine dialogues. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 565–577, Online. Association for Computational Linguistics.

Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas PYKL, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio, and Amitava Das. 2020. SemEval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 774–790, Barcelona (online). International Committee for Computational Linguistics.

Susan Poplack. 1978. Sometimes i'll start a sentence in spanish y termino en español: Toward a typology of code-switching. *Linguistics*, 16(7-8):581–618.

Tom Potter and Zheng Yuan. 2024. LLM-based code-switched text generation for grammatical error correction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*,

pages 16957–16965, Miami, Florida, USA. Association for Computational Linguistics.

Adithya Pratapa, Gayatri Bhat, Monojit Choudhury, Sunayana Sitaram, Sandipan Dandapat, and Kalika Bali. 2018. Language modeling for code-mixing: The role of linguistic theory based synthetic data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1543–1553, Melbourne, Australia. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Samson Tan and Shafiq Joty. 2021. Code-mixing on sesame street: Dawn of the adversarial polyglots. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3596–3616, Online. Association for Computational Linguistics.

Genta Winata et al. 2021a. Multilingual pretrained models are effective for code-switching nlp. In *Proceedings of EMNLP 2021*, pages 2345–2356.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021b. 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.

Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2019. Code-switched language models using neural based synthetic data from parallel sentences. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 271–280, Hong Kong, China. Association for Computational Linguistics.

Anjali Yadav, Tanya Garg, Matej Klemen, Matej Ulcar, Basant Agarwal, and Marko Robnik Sikonja. 2024. Code-mixed sentiment and hate-speech prediction. *arXiv preprint arXiv:2405.12929*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, and Alham Fikri Aji. 2023. 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.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Zhihong Zhu, Xuxin Cheng, Zhiqi Huang, Dongsheng Chen, and Yuexian Zou. 2023. Enhancing code-switching for cross-lingual SLU: A unified view of semantic and grammatical coherence. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7849–7856, Singapore. Association for Computational Linguistics.

A Additional Experimental Details

Hardware and budget. Experiments ran on mixed GPU clusters with NVIDIA A100 (40GB) and A10 (24GB) devices. We report device-time explicitly: **528 GPU-hours** in total: A100: 192 GPU-h; A10: 336 GPU-h.

Evaluation harness. We adapted EleutherAI's LM Evaluation Harness to our CSW variants with identical prompting across models (zero-shot for main experiments; mitigation prompts only in §D).

B Code-Switched Text Generation: Components and Selection

This appendix complements §3.2 with concrete generation components, prompts, quality controls, and selection criteria.

B.1 LLM Selection for Generation

We compared Claude 3.5 Sonnet and GPT-4o as *generators* in both Alignment-Based and LLM-Centric pipelines. For each language pair (EN→AR/FR/DE/ZH), we sampled 100 items across *Belebele*, *MMLU*, and *XNLI*, and generated *noun-token* CSW sentences under explicit ECT/MLF guidance.

Human preference. Bilingual author-annotators conducted blind, pairwise preferences on naturalness ("which sounds more like plausible bilingual speech?"). Results (Table 6) favored Claude across all languages by a modest but consistent margin; thus Claude was selected as the *generator* for the main benchmark construction. We also performed manual sanity checks during prompt refinement to verify ECT/MLF adherence (see §B.7).

Embedded Language	Claude (%)	GPT-4o (%)
Arabic	55	45
Chinese	57	43
French	52	48
German	62	38

Table 6: Human preferences for CSW text generated by Claude vs. GPT-4o (100 examples per language pair).

B.2 Embedding Backbone for Alignment (Alignment-Based Pipeline)

We evaluated AWESOME+{mBERT, LaBSE} as alignment backbones to support POS-guided substitutions. For each pair, 300 noun-token CSW sentences were built using each backbone, with substitutions performed by Claude under the same prompts. GPT-4o served as an *independent judge* to avoid self-evaluation by Claude. LaBSE-based alignments yielded more natural outputs (Table 7), and we adopt LaBSE for alignment-dependent experiments.

Embedded Language	LaBSE (%)	mBERT (%)
Arabic	89.0	11.0
Chinese	91.3	8.7
French	74.7	25.3
German	55.3	44.7

Table 7: GPT-4o preferences for LaBSE vs. mBERT alignments (300 judgments per language).

B.3 LLM-Centric Prompt Templates (Two-Step, Noun-Token)

Step 1: Placeholder identification (ECT/MLF-aware). We mark switchable English nouns with "#######", preserving English as matrix language and ECT/MLF constraints.

```
You are an expert linguist in code-switching. Using the Equivalence Constraint
Theory (ECT) and the Matrix Language Frame (MLF) model, identify nouns in the
English sentence that are natural switch points. Replace each such noun with
"#######" and return only the transformed sentence (no commentary). Ensure the
English sentence remains the grammatical frame (MLF), and avoid fixed
expressions or idioms that should not be switched.

[English sentence]
{text}
```

Figure 5: Step 1 prompt: placeholder identification (noun-token).

Step 2: Placeholder filling (aligned insertions). We fill each placeholder using only aligned tokens from the parallel target-language sentence, adjusting inflection to maintain well-formedness.

```
You are given parallel texts in English and {LANGUAGE}. Produce a code-switched
English sentence by replacing each "#######" with its aligned
{LANGUAGE} counterpart from the {LANGUAGE} sentence. Preserve meaning and
English as the matrix language (MLF). Respect ECT; adjust inflection as needed.
Return only the final code-switched sentence.

[English with placeholders]
{placeholder_text}

[{LANGUAGE} sentence]
{target_text}
```

Figure 6: Step 2 prompt: placeholder filling (noun-token).

B.4 Ratio-Token Variant (Random)

For the random CSW variant, we replace a 20% of tokens with aligned counterparts irrespective of syntax:

```
You are given an English sentence with "#######" placeholders and its  {LANGUAGE} parallel. Replace each
    "#######" with the aligned {LANGUAGE}  segment so that the result reads naturally as mixed English {
    LANGUAGE}.  Preserve order and return only the final sentence.

[English with placeholders]
{placeholder_text}

[{LANGUAGE} parallel]
{target_text}
```

Figure 7: Prompt used in the ratio-token variant (after random placeholder placement).

B.5 Final Generation Approach Selection

We compared Alignment-Based vs. LLM-Centric for noun-token generation on 100 items per language and benchmark, using GPT-4o as the *independent* judge (single binary preference per pair). LLM-Centric was preferred across the board (Table 8), hence adopted for main benchmarks; the Alignment-Based pipeline is retained when explicit control of replacement ratios is needed.

Embedded Language	LLM-Centric (%)	Alignment-Based (%)
Arabic	56.1	43.9
Chinese	66.0	34.0
French	63.8	36.2
German	53.4	46.6

Table 8: GPT-4o preferences between generation methods (noun-token).

B.6 LLM-as-Judge Protocol and Prompt

We used GPT-4o to avoid self-evaluation by Claude. Each comparison returns a single preference (A or B). Ties default to A. Criteria are presented jointly and elicit a single overall choice to avoid ambiguous

multi-criterion aggregation.

```
You will compare two code-switched sentences, A and B, mixing English (matrix
language) with {LANGUAGE}. Choose the better sentence by considering:

1) Fluency (plausible bilingual speech),
2) Depth of mixing (meaningful integration beyond isolated tokens),
3) Switch grammar under Equivalence Constraint Theory (ECT),
4) English as grammatical frame (Matrix Language Frame, MLF),
5) Overall coherence.

Return exactly "A" or "B". If you cannot decide, choose "A".
A: {sentence_one}
B: {sentence_two}
```

Figure 8: GPT-4o prompt for LLM-as-Judge preference between two CSW sentences.

B.7 Manual Sanity Checks and Human Protocol

Beyond pairwise preferences, authors conducted manual sanity checks during prompt development and post-generation spot checks to verify ECT/MLF adherence and naturalness.

C Dataset Characterization and Controls

C.1 Foreign-Token Proportions (Noun-Token, EN as Matrix)

We report the average percentage of tokens replaced by embedded-language items per instance (means across test splits). Minor length fluctuations (≈5%) arise from morphological realization.

Dataset	EN→AR	EN→FR	EN→DE	EN→ZH
Belebele	36.92	31.49	32.00	31.83
MMLU	39.98	34.18	34.02	34.91
XNLI	33.92	29.43	29.53	29.20

Table 9: Average proportion (%) of embedded tokens per instance (noun-token CSW).

C.2 Ratio vs. Noun-Token Replacement Rates

The *ratio-token* variant fixes the replacement rate at **20%**, whereas the *noun-token* variant yields ≈30–40% depending on dataset/language (Table 9).

D Instructional Prompts for Prompt-Based Mitigation

We use short instructions instantiated per language (AR/FR/DE/ZH) and per task. Prompts avoid translation instructions and cue the model to *reason over mixed text*.

Belebele

```
You will be given a passage and a question written in code-switched English and {
    LANGUAGE}.
Do not translate. Read and reason over the mixed-language text, then answer with A/B
    /C/D.
```

Figure 9: Mitigation instruction for *Belebele*. Instantiate {LANGUAGE} as AR/FR/DE/ZH.

MMLU

```
You will be given a question written in code-switched English and {LANGUAGE}.
Do not translate. Read and reason over the mixed-language text, then answer with A/B
    /C/D.
```

Figure 10: Mitigation instruction for *MMLU*.

XNLI

```
You will be given a premise and a hypothesis written in code-switched English and {
    LANGUAGE}.
Do not translate. Read and reason over the mixed-language text, then answer with
    0/1/2.
```

Figure 11: Mitigation instruction for *XNLI*.

E Instruction Tuning for Model-Based Mitigation

We instruction-tuned *LLaMA-3.1-8B-Instruct* on CSW *generation* tasks (to expose the model to CSW structure), then evaluated comprehension (zero-shot) on our CSW benchmarks.

E.1 Dataset Preparation

We used parallel TED Talks (Qi et al., 2018). For each English sentence >70 words, we built 4 CSW variants (EN as matrix; AR/FR/DE/ZH embedded) using the LLM-Centric noun-token pipeline (§B.3). Total size: **14.6k** instruction–response pairs.

E.2 Prompt Templates (Five Styles)

Each instance sampled one of five equivalent instruction templates to avoid overfitting surface form; examples below.

```
Take this English sentence and infuse it with <LANGUAGE> code-switching:
English: "<ENGLISH_SENTENCE>"
<LANGUAGE>: "<TRANSLATION_SENTENCE>"
```

Figure 12: Infusion-style template (one of five).

```
Convert the following English line into a code-switched mix with <LANGUAGE>:
English: "<ENGLISH_SENTENCE>"
<LANGUAGE>: "<TRANSLATION_SENTENCE>"
```

Figure 13: Conversion-style template.

```
Blend English and <LANGUAGE> in the sentence below:
English text: "<ENGLISH_SENTENCE>"
<LANGUAGE> equivalent: "<TRANSLATION_SENTENCE>"
```

Figure 14: Blending-style template.

```
Generate a code-switched rendition by swapping in <LANGUAGE>:
English original: "<ENGLISH_SENTENCE>"
<LANGUAGE> snippet: "<TRANSLATION_SENTENCE>"
```

Figure 15: Rendition-style template.

```
Switch parts of this English sentence into <LANGUAGE>:
English: "<ENGLISH_SENTENCE>"
<LANGUAGE>: "<TRANSLATION_SENTENCE>"
```

Figure 16: Switching-style template.

E.3 Training Setup

Full-parameter instruction tuning, single epoch, BF16 mixed precision, dynamic packing to 4096 context length, per-device batch size 4.

Optimizer & schedule: AdamW (β_1=0.9, β_2=0.95), weight decay 0.1, grad clip 1.0, linear decay with 5% warmup.

Validation and selection. We held out a stratified dev split for early stopping and checkpoint selection on CSW perplexity.

Example instance (abridged). Example instance (IFT, EN↔FR).

```
[
  {
    "content": "Take this English sentence and infuse it with French code-switching
        :\nEnglish: This is back in 1964 , when the technical managers have suits
        and ties and NRK rolled all its equipment on board a ship , and 200 meters
        out of the shore , transmitting the signal back , and in the machine room ,
        they talked to the machine guy , and on the deck , they have splendid
        entertainment .\nFrench: En 1964 par exemple , - les directeurs techniques
        portaient encore des cravates . La NRK avait embarqué son matériel à bord ,
        qui , jusque à 200 m des côtes , transmettait le signal . Dans la salle des
        machines , ils ont parlé aux machinistes . Et sur les ponts , ils avaient de
         fabuleuses activités récréatives .\n",
    "role": "user"
  },
  {
    "content": "This is back in 1964, when the directeurs techniques have cravates
        and costumes and NRK rolled all its matériel on board a navire, and 200 mè
        tres out of the côtes, transmitting the signal back, and in the salle des
        machines, they talked to the machiniste guy, and on the ponts, they have
        splendid activités récréatives.",
    "role": "assistant"
  }
]
```

Figure 17: IFT example instance (French).

Model	EN→AR	EN→DE	EN→FR	EN→ZH
LLaMA 3B	−0.16	−0.13	−0.15	−0.18
Qwen 3B	+0.02	+0.03	+0.02	+0.02
Allam 7B	+0.01	+0.01	+0.01	0.00
Mistral 7B	−0.01	−0.02	−0.02	−0.01
Qwen 7B	+0.02	+0.01	+0.02	+0.02
LLaMA 8B	−0.07	−0.04	−0.04	−0.04
LLaMA 70B	−0.13	−0.13	−0.03	−0.17
Qwen 72B	+0.05	+0.05	+0.06	+0.07

Table 10: ICL mitigation deltas: (ICL accuracy) − (unprompted CSW accuracy), weighted across tasks.

F ICL vs. Unprompted: Accuracy Deltas (Noun-Token)

Table 10 reports absolute ICL versus CSW accuracy deltas (weighted averages) for EN as matrix; positive values indicate that mitigation helps. As discussed in the main text, gains concentrate in multilingual-heavy families (*Qwen*, *Allam*), while *Llama* and *Mistral* often see neutral or negative effects.