CS-Sum: A Benchmark for Code-Switching Dialogue Summarization and the Limits of Large Language Models

Anonymous ACL submission

Abstract

Code-switching (CS) poses a significant challenge for Large Language Models (LLMs), yet its comprehensibility remains underexplored in LLMs. We introduce **CS-Sum**, to evaluate the comprehensibility of CS by the LLMs through CS dialogue to English summarization. CS-Sum is the first benchmark for CS dialogue summarization across Mandarin-English (EN-ZH), Tamil-English (EN-TA), and Malay-English (EN-MS), with 900-1300 human-annotated dialogues per language pair. Evaluating ten LLMs, including open and closed-source models, we analyze performance across few-shot, translate-summarize, and fine-tuning (LoRA, QLoRA on synthetic data) approaches. Our findings show that though the scores on automated metrics are high, LLMs make subtle mistakes that alter the complete meaning of the dialogue. To this end, we introduce 3 most common type of errors that LLMs make when handling CS input. Error rates vary across CS pairs and LLMs, with some LLMs showing more frequent errors on certain language pairs, underscoring the need for specialized training on code-switched data.

1 Introduction

Code-switching (CS) is the practice of alternating between two or more languages within a single conversation or utterance. Bilingual and multilingual speakers frequently engage in CS. This form of communication has also become prevalent in social media and is the norm in multilingual societies. Large Language Models (LLMs) (Hurst et al., 2024; Team et al., 2023; Dubey et al., 2024), are trained on huge volumes of data majority of which is written in English. Despite the prevalence of CS in real-world communication, existing benchmarks (Huzaifah et al., 2024; Zhang et al., 2023; Yong et al., 2023) fail to assess how well LLMs process CS dialogues. Improved CS comprehension would enable LLMs to interpret multilingual prompts more effectively, enhancing accessibility for CS speakers (Bawa et al., 2020).

CS-Dialogue
Adele: Naa innum antha song ah ketuttu irukan.
Kode: Ethu
Adele: Nee recommended pannathu
Kode: Ama athu nalla irukkum
Adele: Ama nalla irukku
Kode: Atha ennoda music library la 3 years ah vechirkan
Adele: Omg athu avlo old ah?
Kode: hmm
Adele: Athey singer kitta irunthu vera ethavthu suggest panriya
Kode: Naa unakku whole album ah USB la kudukkren
Adele: Eppo nee en veetukku varuva?
Kode: Naa ippo free ah irukken
Adele:Apo naa wait panren
Kode: Varen.

EN-Summary
Adele is still listening to the song which Kode recommended. Kode will come to Adele's home today and will give her the whole album of the same singer.

Figure 1: An instance from the benchmark

Summarization is an ideal task to evaluate CS understanding, as it requires both grasping the dialogue's main idea and generating a concise, coherent summary across languages. The only existing CS dialogue summarization dataset, Gupshup (Mehnaz et al., 2021), covers only one language pair (English-Hindi). Other CS benchmarks like LinCE (Aguilar et al., 2020) and GLUECos (Khanuja et al., 2020) focus on tasks such as language identification and POS-tagging, which assess word- or sentence-level understanding, whereas summarization requires discourse-level comprehension, a much harder test of CS ability.

To truly evaluate the ability of LLMs on CS, we propose the **CS-Sum** benchmark, which contains 900-1300 CS dialogue-summary pairs for three language pairs: Mandarin-English (EN-ZH), Tamil-English (EN-TA) and Malay-English (EN-MS). The CS dialogues were created by native speakers of the respective languages. Instead of generating a dialogue from scratch, the speakers were asked to translate the English dialogues in test sets of DialogSum (Chen et al., 2021) and SAMSum (Gliwa et al., 2019) - two popular English dialogue summarization datasets - to CS dialogues in their respective languages.

We evaluate 9 open source and 1 proprietary LLMs under few-shot, translate-summarize, LoRA (Hu et al., 2022) and QLoRA (Dettmers et al., 2023) on synthetic data and provide a detailed analysis on the struggles of current SOTA LLMs(<=9B). For LoRA and QLoRA fine-tuning, the training data was generated using Gemini-2-flash (Team et al., 2023), as it is a large-scale LLM with a free-to-use API, making it a practical choice for our experiments.

Our qualitative analysis of the summaries generated by LLMs showed that tranditional summarization metrics like ROUGE (Lin, 2004), BERTScore (Zhang et al., 2019) are not sufficient for measuring the quality of the summaries since the LLMs are capable of generating summaries that contain subtle errors that change the entire meaning of the CS dialogues. To this end, we identify the three most common errors (CSL, MST, SMA Section 5) that the LLMs make when summarizing CS dialogues and suggest an LLM-driven approach to analyse their summaries with respect to the errors.

Our main contributions are:

- We introduce CS-Sum, the first CS dialogue summarization benchmark for EN-ZH, EN-TA and EN-MS.
- We conduct an in-depth evaluation of state-of-the-art LLMs, identifying critical failure patterns in their ability to process and summarize CS dialogues.
- We release CS-Sum, along with the full synthetic training dataset.

2 Related Work

While LLMs have demonstrated strong performance in many multilingual tasks, they still face significant challenges when dealing with CS. Studies (Zhang et al., 2023; Yong et al., 2023) have shown that LLMs, including GPT-4 and GPT-3.5, struggle with generating code-switched text, often resulting in language collapse, where the model fails to mix languages properly, and in task-specific failures such as poor performance in summarization and machine translation for CS data. These issues arise because LLMs are typically trained on monolingual data and lack the capacity to handle the intricacies of CS discourse. Additionally (Huzaifah et al., 2024), their performance is inconsistent, particularly when translating low-resource language pairs, with models showing better results for high-resource languages but underperforming when confronted with languages that are underrepresented in training datasets. These findings highlight a big problem as people prefer to interact with LLMs in code-switch (Bawa et al., 2020).

Recent benchmarks such as LinCE (Aguilar et al., 2020) and GLUECoS (Khanuja et al., 2020) have been instrumental in advancing the understanding of CS phenomena by focusing on tasks like LID, NER, and POS tagging. While these datasets have provided valuable insights, it's important to recognize that LLMs continue to encounter challenges when dealing with the complexities inherent in code-switched discourse. To truly evaluate the CS comprehension of LLMs, other benchmarks are necessary—ones that go beyond word-level tasks and assess deeper linguistic understanding. The first step in this direction is CS-Sum, a benchmark that evaluates LLMs on a more complex task, such as summarization, which requires a nuanced comprehension of code-switched text.

3 CS-Sum Benchmark

3.1 Dataset Construction

The goal of building the CS-Sum benchmark, was to address the lack of a benchmark that evaluates the CS comprehension of LLMs. We selected Mandarin (zh), Bahasa Melayu (ms), and Tamil (ta) because these languages have large bilingual speaker communities that frequently code-switch with English. To efficiently create a diverse CS benchmark within resource constraints, we translated a subset of the combined test sets from DialogSum and SAMSum. We recruited 7 native speakers for each language to translate the English dialogues to CS dialogues in their respective languages. The translators were native speakers, all university students (bachelor's or master's level). They were instructed to translate the dialogues depending on how they would have conversed with their peers in CS. The entire translation process lasted for about 5 months.

Language	M-Index	I-Index	Burstiness	Span Entropy	Memory
EN-ZH	0.40	0.36	-0.79	0.13	-0.08
EN-MS	0.41	0.38	-0.65	0.64	-0.24
EN-TA	0.42	0.40	-0.68	0.58	-0.22

Table 1: CS metrics measured on the CS-Sum dataset

	EN-ZH	EN-TA	EN-MS
Number of instances	1320	1000	918
EN utterances	1140	344	517
Lang utterances	2079	983	265
EN as matrix language	4699	2340	2984
Lang as matrix language	4779	6790	1919
Avg. monolingual utterance length	2.344	3.616	5.515
Avg. CS utterance length	5.816	9.642	13.885

Table 2: CS statistics of CS-Sum. 'Lang' refers to the language other than English

CS pair	KL Div.	JS Div.
EN-ZH	2.4786	0.4918
EN-MS	0.5481	0.1454
EN-TA	0.5089	0.1017

Table 3: Divergence between CS-Sum and CS-Sum-Syn

3.2 Corpus Overview and Analysis

Table 2 presents key statistics of the CS-Sum benchmark across three language pairs. We observe that non-English utterances outnumber English ones in EN-ZH and EN-TA, while the reverse is true for EN-MS, suggesting stronger local language dominance in the former two. The matrix language distribution further highlights this variation: Tamil dominates as the matrix language in EN-TA, whereas EN-ZH and EN-MS are more balanced. This indicates differing CS behaviors—embedding in EN-TA versus alternation in EN-ZH and EN-MS.

Utterance lengths also vary significantly. EN-MS exhibits the longest average CS utterances (13.88 tokens), compared to 5.81 in EN-ZH, suggesting more complex span-level dependencies in Malay-English dialogues. The longer monolingual spans and varied matrix language roles imply that LLMs must manage different CS dynamics across language pairs.

3.3 Quantifying Code-Switching

We quantify the structural properties of code-switching in CS-Sum using five established metrics (Guzmán et al., 2017), reported in Table 1. The **M-Index** measures the distributional balance between languages, with values near 1 indicating equal usage. The **I-Index** captures the token-level likelihood of language alternation. **Burstiness** characterizes the temporal irregularity of switching, where negative values imply regular alternation and positive values indicate clustering. **Span Entropy** quantifies the unpredictability in monolingual segment lengths, while **Memory** reflects the autocorrelation of consecutive spans—positive values suggest persistence, negative values suggest alternation.

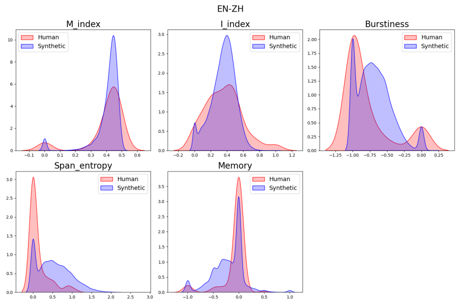
Across the three CS pairs, CS-Sum exhibits moderate switching behavior. The M-Index values (0.40–0.42) suggest relatively balanced bilingual distributions, and I-Index scores (0.36–0.40) indicate frequent intra-sentential switching, most prominently in EN-TA. Burstiness values are consistently negative, reflecting structured rather than chaotic switching. Span Entropy is lowest for EN-ZH (0.13), indicating predictable switch boundaries, while EN-TA (0.58) and EN-MS (0.64) exhibit higher variability. Memory values are near zero or negative across all pairs, implying weak temporal consistency in span lengths, and highlighting the need for LLMs to dynamically track and adapt to switching boundaries during summarization.

3.4 Gemini-Generated Training Data

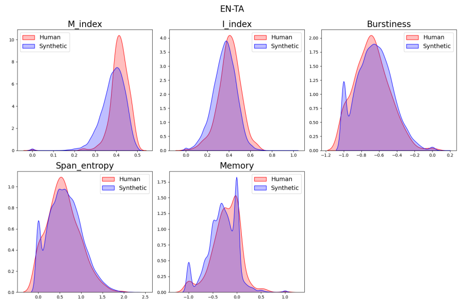
To support fine-tuning for evaluating LLMs on CS-Sum, we construct a synthetic training set, **CS-Sum-Syn**, by converting 19,014 English dialogue-summary pairs from DialogSum and SAMSum into code-switched dialogues using Gemini-2[1] (Team et al., 2023). This augmentation is not a contribution in itself, but a means to enable parameter-efficient adaptation (e.g., LoRA, QLoRA). To validate the distributional alignment of synthetic and human CS data, we compute CS metrics across both sets and visualize them for each language pair in Figures 2a, 2b, and 2c.

We find that Gemini-generated data broadly approximates human CS patterns for EN-TA and EN-

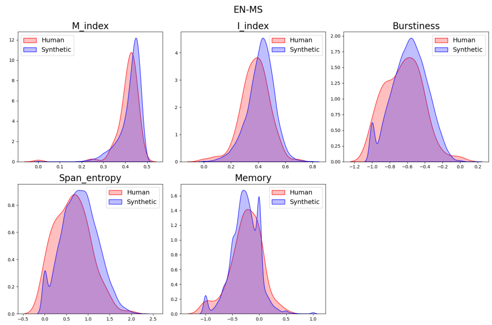
[1] https://ai.google.dev/, gemini-2-flash-exp



(a) ZH Distribution



(b) TA Distribution



(c) MS Distribution

MS. In EN-TA, distributions over I-Index, burstiness, and span entropy align closely, suggesting that Gemini captures intra-sentential switching well. EN-MS also shows good overlap, though with slightly more bursty and irregular switching in the synthetic set. In contrast, EN-ZH exhibits substantial divergence: synthetic dialogues show flatter memory and burstiness curves, as well as a long-tailed span entropy distribution, indicating reduced switching regularity and predictability. This mismatch is quantified in Table 3, where the KL and JS divergence for EN-ZH are 4–5× higher than for the other pairs. These discrepancies may affect the quality of fine-tuning and are revisited in Section 5 during error diagnosis.

4 LLM performance on CS-Sum

In this section, we present the performance of ten LLMs—nine open-source[2] and one proprietary—on CS-Sum across four settings: Few-Shot, Translate-Summarize, LoRA, and QLoRA. In LoRA and QLoRA, open-source models are fine-tuned on CS-Sum-Syn. The evaluated open-source LLMs include LLAMA-3-8B (Dubey et al., 2024), MISTRAL-7B (Jiang et al., 2023), MINISTRAL-8B[3], GEMMA-2-2B and GEMMA-2-9B (Team et al., 2024), QWEN2.5-2B and QWEN2.5-7B (Yang et al., 2024), SEA-LION-GEMMA-2-9B[4], and SEALLM-7B (Nguyen et al., 2024). The proprietary model used is GPT-4O (Hurst et al., 2024). These models were selected based on their multilingual pretraining exposure and computational feasibility, with sizes ranging from 2B to 9B parameters. Models like LLAMA-3-8B, MISTRAL-7B, and GEMMA-2-9B have shown strong generalization across NLP tasks, while QWEN2.5-7B, SEA-LION-GEMMA-2-9B, and SEALLM-7B were chosen for their targeted multilingual capabilities, especially in languages like Bahasa Melayu, Tamil, and Mandarin. GPT-4O serves as a strong proprietary baseline due to its SOTA performance in cross-lingual comprehension.

By evaluating these models across different parameter scales and varying degrees of multilingual exposure, we investigate how model size and pretraining diversity affect performance. LoRA and QLoRA fine-tuning on CS-Sum-Syn result in higher scores on automated metrics, suggesting improved alignment with reference summaries. However, it remains unclear whether these gains reflect actual comprehension of code-switched dialogue or simply the ability to replicate the distributional patterns of Gemini-generated synthetic data. To explore this, we conduct an in-depth error anal-

[2]Instruction-tuned and latest versions
[3]https://huggingface.co/mistralai/Ministral-8B-Instruct-2410
[4]https://sea-lion.ai/

Model	Lang	ROUGE-L	BERTScore	SBERT-COSINE	JACCARD	METEOR
Gemma-2-2B	EN-ZH	0.2330 / -4	0.8876 / -0.35	0.7268 / -1.40	0.1658 / -7.64	0.2699 / -0.9
	EN-TA	0.2202 / -7	0.8812 / -0.38	0.6755 / -0.85	0.1523 / -9.19	0.2453 / -5.23
	EN-MS	0.2421 / -7.18	0.8898 / -0.98	0.7352 / -4.35	0.1739 / -13.52	0.2902 / -12.43
Qwen2.5-3B	EN-ZH	0.2393 / 9.88	0.8824 / 0.93	0.7143 / 1.70	0.1643 / 5.82	0.2974 / 2.05
	EN-TA	0.2227 / -1.46	0.8805 / 0.44	0.6880 / -2.35	0.1468 / -0.13	0.2591 / -4.82
	EN-MS	0.2408 / 6.21	0.8868 / 0.60	0.7308 / -2.25	0.1694 / 1.14	0.3100 / -4.57
Qwen2.5-7B	EN-ZH	0.2482 / 27.10	0.8106 / 11.20	0.4927 / 54.12	0.0783 / 162.62	0.1118 / 188.89
	EN-TA	0.2800 / -5.11	0.8937 / -0.55	0.7081 / -2.14	0.1837 / -9.62	0.2723 / -9.20
	EN-MS	0.3117 / -1.65	0.9041 / -0.46	0.7520 / -0.67	0.2122 / -8.37	0.3372 / -10.26
SEALLM-7B	EN-ZH	0.2115 / 19.25	0.8122 / 9.15	0.4836 / 45.48	0.0834 / 89.27	0.1297 / 104.29
	EN-TA	0.2270 / 0.99	0.8686 / 1.21	0.6470 / 2.53	0.1336 / 3.44	0.2287 / 1.55
	EN-MS	0.2575 / 8.41	0.8801 / 1.28	0.6956 / 4.91	0.1633 / 9.04	0.2800 / 5.95
Mistral-7B	EN-ZH	0.2568 / -3.10	0.8902 / -0.42	0.7208 / -0.34	0.1754 / -8.01	0.3010 / -7.82
	EN-TA	0.2198 / -2.15	0.8792 / 0.04	0.6847 / -0.95	0.1470 / -6.91	0.2657 / -10.45
	EN-MS	0.2421 / 1.59	0.8897 / -0.46	0.7293 / -2.15	0.1796 / -11.49	0.3084 / -8.40
Ministral-8B	EN-ZH	0.2560 / 4.23	0.8875 / 0.52	0.7037 / 1.06	0.1588 / 6.13	0.2517 / 1.15
	EN-TA	0.2468 / -5.61	0.8858 / -0.18	0.6752 / -2.95	0.1561 / -3.66	0.2387 / -8.53
	EN-MS	0.2700 / 6.97	0.8917 / 0.37	0.7262 / 0.03	0.1794 / 2.93	0.2843 / -0.09
LLaMA-3-8B	EN-ZH	0.2868 / -2.29	0.8845 / 0.76	0.7090 / 4.09	0.1918 / -6.99	0.2916 / 0.88
	EN-TA	0.2453 / 0.47	0.8769 / 0.99	0.6707 / 3.28	0.1598 / 1.07	0.2570 / -2.13
	EN-MS	0.2599 / 8.06	0.8724 / 2.04	0.6538 / 11.21	0.1804 / 0.02	0.2833 / 5.50
Gemma-2-9B	EN-ZH	0.2995 / -5.62	0.8987 / -0.27	0.7603 / -0.91	0.2054 / -5.83	0.3246 / -3.06
	EN-TA	0.2761 / -3.85	0.8911 / 0.57	0.7203 / 4.42	0.1832 / 4.63	0.3107 / 1.25
	EN-MS	0.3009 / -3.00	0.8975 / 0.01	0.7573 / -0.94	0.2096 / -3.96	0.3279 / -0.07
SEA-LION-Gemma-2-9B	EN-ZH	0.2799 / -3.96	0.8968 / -0.34	0.7542 / -0.41	0.1820 / -4.29	0.2944 / -4.46
	EN-TA	0.2703 / 0.26	0.8898 / 0.48	0.7138 / 3.79	0.1727 / -0.12	0.2884 / -2.38
	EN-MS	0.2860 / -1.56	0.8963 / -0.12	0.7404 / 1.21	0.1837 / -4.16	0.2983 / -4.75
GPT-4o	EN-ZH	0.2965 / -1.03	0.8816 / 2.21	0.7082 / 9.63	0.1792 / 11.66	0.2853 / 16.52
	EN-TA	0.3157 / -4.92	0.8981 / 0.39	0.7570 / 2.19	0.2041 / -0.55	0.3419 / 0.75
	EN-MS	0.3126 / -1.70	0.9037 / -0.15	0.7750 / 0.12	0.2045 / 0.33	0.3365 / 1.09

Table 4: Few-Shot Performance/Translate-Summarize % improvement on CS-Sum

ysis in Section 5, which reveals that fine-tuning on synthetic data does not improve—and often degrades—the model's ability to accurately interpret and summarize code-switched content.

4.1 Result Analysis

In this subsection, we present the results of the LLMs' performance on the CS-Sum in Few-Shot, Translate-Summarize, LoRA and QLoRA settings. Appendix B provides an explanation on the different settings. We use standard summarization metrics like ROUGE (Lin, 2004), BERTScore (Zhang et al., 2019), SBERT-Cosine (Reimers and Gurevych, 2019), Jaccard and METEOR (Banerjee and Lavie, 2005). This comprehensive set of metrics evaluate both word-level matching and semantic similarity.

The few-shot performance of LLMs across the three CS language pairs is reported in Table 4. While semantic similarity metrics such as BERTScore and SBERT-Cosine report high values, these can be misleading. As illustrated in Figure 3, the summaries often fail to accurately capture the core content or intent of the original dialogue, despite appearing semantically plausible. This reflects a common failure mode of LLMs in CS summarization: generating fluent, high-overlap summaries that omit or distort key factual elements. In this example, Gemma-2-9B misrepresents who confirms the location of the conference and fabricates speaker opinions, despite achieving a high semantic similarity score. ROUGE-L scores further corroborate this gap, with LLMs scoring nearly 50% lower than the best-performing models on the monolingual DialogSum[5] and SAMSum[6] benchmarks.

From Table 4, we observe that GPT-4o is the best-performing model overall. However, due to

[5] https://paperswithcode.com/sota/text-summarization-on-dialogsum

[6] https://paperswithcode.com/sota/text-summarization-on-samsum-corpus

the lack of transparency in its training and architecture, we cannot provide deeper analysis on its performance characteristics. Among open-source models, Gemma-2-9B consistently outperforms others across most metrics and language pairs. This aligns with findings from the Gemma-2 paper (Team et al., 2024), which attributes its strong multilingual performance to extensive exposure to diverse languages and an optimized tokenizer. Interestingly, SEA-Lion-Gemma-2-9B lags slightly behind, despite its regional specialization in Southeast Asian languages, suggesting that broader multilingual exposure may be more beneficial than region-specific pretraining for CS comprehension.

One notable observation is the underperformance of most models on the EN-ZH pair, which is surprising given that all models were exposed to Mandarin during pretraining. Manual inspection reveals that some LLMs generate summaries partially or entirely in Chinese, despite being explicitly prompted to produce English outputs. This mismatch leads to lower scores on automated metrics, particularly those that penalize non-English outputs. However, as discussed in our fine-grained evaluation (Section 5), these summaries often contain fewer semantic or structural errors than those in EN-TA or EN-MS, indicating that the low metric scores may not reflect actual comprehension quality.

We also find that smaller models, such as Gemma-2-2B and Qwen2.5-3B, perform significantly worse than their larger counterparts, reinforcing the impact of model scale on CS summarization. Surprisingly, SEALLM-7B performs comparably to these smaller models, despite being explicitly trained on Southeast Asian languages. Qualitative analysis of its outputs reveals several failure modes in the few-shot setting: **(a)** copying utterances directly from the dialogue, **(b)** generating descriptive paraphrases rather than true summaries, and **(c)** producing outputs in the wrong language. These behaviors suggest that SEALLM struggles with in-context learning and generalization, even within its intended linguistic domain.

Translate-Summarize % Improvement Table 4 reports percentage improvements under the Translate-Summarize setting. Contrary to expectations, most models show little to no improvement, and in many cases, performance declines across all metrics. This suggests that translation introduces additional challenges, likely due to LLMs' inability

LLM	**Human**
Emily is asking where the ASEES conference is held. Chloe thinks it's probably in San Francisco or in Hawaii, but James thinks it's held in California. Chloe thinks that's sad.	James informs that the next ASEEES conference will be held in San Francisco. Chloe was hoping for Hawaii. James doesn't think there will be a conference in Hawaii soon due to the high cost of travel.

Figure 3: Wrong summary with high BERTScore 0.903

to preserve the structural and discourse-level nuances of code-switched input. Notably, Qwen2.5-7B and SEALLM-7B improve significantly for EN-ZH (27.1% and 19.25% in ROUGE-L), reflecting benefits from Mandarin or SEA-specific pretraining. However, similar gains are absent for EN-TA and EN-MS, indicating that CS translation remains unreliable outside of high-resource language pairs.

LoRA and QLoRA Results We fine-tune LLMs on the synthetic CS-Sum-Syn dataset using LoRA and QLoRA to evaluate their ability to handle CS summarization beyond just metric gains. As shown in Table 5 [7], EN-ZH sees the largest improvements (e.g., SEALLM-7B and Qwen2.5-7B exceed 200% in METEOR), EN-TA shows consistent gains, while EN-MS exhibits mixed results across metrics.

These improvements are surprising given the distributional mismatch between CS-Sum and CS-Sum-Syn (Figure 2a, Table 3). A qualitative review reveals that models often produce summaries with subtle semantic errors, indicating they may have learned surface-level summary structure rather than true code-switching comprehension. This motivates the deeper analysis in Section 5.

5 Error Analysis

In this section, we present an in-depth analysis of the common error types exhibited by LLMs when summarizing CS dialogues. Our qualitative analysis across the EN-ZH, EN-TA, and EN-MS language pairs revealed three major failure modes:

- **Code-Switching Loss (CSL)**: The summary primarily utilizes the English parts of the dialogue, ignoring or missing critical information from non-English segments.
- **Meaning Shift from Poor Translation**

[7] Due to space constraints, rest of the scores are presented in Table 8

Model	Lang	ROUGE-L	BERTScore	SBERT-COSINE	JACCARD	METEOR
SEALLM-7B	EN-ZH	50.45 / 51.94	11.09 / 11.21	60.22 / 60.69	170.47 / 160.94	163.13 / 142.51
	EN-TA	47.99 / 43.82	4.49 / 4.52	19.84 / 19.86	75.18 / 69.14	47.14 / 35.42
	EN-MS	34.20 / 17.03	3.29 / 2.53	12.01 / 27.78	47.49 / 4.22	13.08 / 27.78
LLaMA-3-8B	EN-ZH	23.51 / 25.21	3.14 / 3.12	12.88 / 12.79	29.85 / 34.26	19.70 / 29.68
	EN-TA	51.45 / 47.07	4.32 / 4.00	18.91 / 18.28	66.19 / 60.27	39.72 / 42.35
	EN-MS	46.81 / 21.73	4.85 / 3.27	22.19 / 16.94	50.56 / 23.01	39.34 / 17.83
Gemma-2-9B	EN-ZH	-1.45 / 10.66	0.12 / 0.82	4.10 / 2.32	8.56 / 11.61	1.68 / 7.95
	EN-TA	6.54 / 19.37	0.63 / 1.56	7.14 / 9.12	18.12 / 27.89	7.28 / 13.01
	EN-MS	7.81 / -6.27	0.71 / -0.61	3.60 / -1.41	8.48 / -3.45	5.31 / -7.06
SEA-LION-Gemma-2-9B	EN-ZH	-22.93 / -6.64	-3.10 / -0.71	-3.27 / 1.73	-4.13 / 8.67	6.29 / 14.39
	EN-TA	-21.48 / 3.59	-2.61 / 0.24	1.74 / 7.08	-3.49 / 20.38	7.38 / 17.21
	EN-MS	-21.98 / -8.08	-2.95 / -0.99	-3.53 / -0.77	-8.16 / 3.18	9.09 / 3.03

Table 5: LoRA / QLoRA % Improvement over Few-Shot on CS-Sum

(MST): The model misunderstands the code-switched segments, resulting in summaries that deviate from the dialogue's true meaning.

- **Speaker Misattribution (SMA)**: Summaries incorrectly assign statements to speakers, thus distorting intended meanings or misrepresenting participants' views.

Figure 4 illustrates CSL, demonstrating how the summary incorrectly emphasizes English segments and misinterprets speaker intentions. Further examples across categories are provided in Tables 9, 10, and 11. Errors were identified automatically via GPT-4o, detailed further in Appendix C.

To quantify these errors and compare model behaviors, we report the percentage of summaries exhibiting each error type across 9 LLMs under both Few-Shot and LoRA settings (Table 6). The analysis reveals the following:

CSL persists as the dominant failure mode in Few-Shot evaluation. Across *all* 9 models and the three language pairs, CSL exceeds 50% in the Few-Shot setting (Table 6). Even the best performer, SEA-LION-Gemma-2-9B, attains a minimum of 53.73% CSL on **EN-ZH**, confirming that LLMs systematically ignore non-English spans when summarizing code-switched dialogues, regardless of architecture or pre-training recipe.

Scaling within a model family does not reliably reduce errors. Comparing parameter-matched variants shows no monotonic gains: Gemma-2-9B lowers CSL on **EN-ZH** (77.61→59.70) yet *raises* Speaker Misattribution (SMA) on **EN-TA** by 7 pp over its 2B sibling, while Qwen2.5-7B improves CSL on **EN-MS** but leaves MST virtually unchanged. These inconsistencies indicate that sheer scale is not a substitute for task-specific multilingual evaluation.

EN-TA exhibits consistently higher error rates across models. All evaluated models show their highest rates of CSL and SMA on EN-TA dialogues (e.g., Qwen2.5-7B: 91.81% CSL, 57.31% SMA). This suggests that certain language-specific properties, such as morphological complexity and syntactic divergence from English, may contribute to increased summarization difficulty in this setting.

SEA-oriented models excel in error analysis despite mediocre automatic scores. Regional models such as SEALLM-7B and SEA-LION-Gemma-2-9B achieve the lowest CSL and SMA within their parameter class (e.g., 53.73% CSL on **EN-ZH**), yet trail larger English-specific models models on ROUGE and BERTScore. This divergence reinforces that conventional metrics do not capture CS comprehension, underscoring the need for better metrics.

Synthetic fine-tuning amplifies errors under distribution shift. Fine-tuning on the Gemini-generated CS-Sum-Syn corpus degrades performance when the training distribution diverges from CS-Sum, most notably for **EN-ZH**, whose KL divergence is 2.48 versus $\leq$0.55 for the other pairs (Table 3). After adaptation, SEA-LION-Gemma-2-9B's CSL jumps from 53.73% to 83.94% and MST from 11.19% to 76.64% (Table 6). These results illustrate that even synthetic data generated using a high-performing LLM does not improve the model's ability to comprehend code-switched input, and the failure to produce data that enhances CS understanding also underscores the limitations of current LLMs comprehensibility of CS.

Model	Lang	CSL		MST		SMA	
		Few-Shot	LoRA	Few-Shot	LoRA	Few-Shot	LoRA
Gemma-2-2B	EN-ZH	77.61	88.32	60.45	90.51	40.30	77.37
	EN-TA	94.15	92.69	86.26	90.06	69.88	74.85
	EN-MS	77.19	94.87	62.28	84.62	35.09	67.52
Qwen2.5-3B	EN-ZH	55.22	83.94	52.24	69.34	36.57	39.42
	EN-TA	91.81	88.01	89.47	75.15	76.02	50.58
	EN-MS	60.53	84.62	49.12	60.68	31.58	40.17
Qwen2.5-7B	EN-ZH	71.64	81.75	32.84	48.91	9.70	26.28
	EN-TA	91.81	79.23	76.32	58.64	57.31	32.58
	EN-MS	78.95	82.05	35.96	42.74	16.67	24.79
SEALLM-7B	EN-ZH	64.18	84.67	41.79	70.80	21.64	33.58
	EN-TA	91.23	93.57	81.87	71.64	63.16	47.95
	EN-MS	68.42	85.47	45.61	69.23	23.68	40.17
Mistral-7B	EN-ZH	55.97	83.21	37.31	69.34	22.39	40.88
	EN-TA	88.30	89.47	81.58	79.82	66.37	53.22
	EN-MS	67.54	77.78	44.74	70.09	28.07	44.44
Ministral-8B	EN-ZH	76.12	85.40	52.24	74.45	31.34	48.91
	EN-TA	95.03	90.94	88.80	79.82	66.37	55.56
	EN-MS	78.95	95.73	49.12	82.91	20.18	57.26
LLaMA-3-8B	EN-ZH	74.63	71.17	35.82	45.07	16.41	21.35
	EN-TA	90.94	**56.92**	80.70	37.89	55.85	19.93
	EN-MS	73.68	62.91	40.35	33.38	14.04	17.33
Gemma-2-9B	EN-ZH	59.70	85.40	18.66	57.66	8.21	35.77
	EN-TA	71.35	82.75	35.09	67.84	22.22	46.20
	EN-MS	59.65	79.60	18.42	52.53	8.77	29.45
SEA-LION-Gemma-2-9B	EN-ZH	**53.73**	83.94	**11.19**	76.64	**3.73**	52.55
	EN-TA	60.53	84.50	**27.78**	76.61	**17.54**	49.71
	EN-MS	**48.25**	86.32	**14.04**	64.96	**7.02**	46.15

Table 6: Comparison of error rates (%) in Few-Shot vs. LoRA fine-tuned models across different CS language pairs.

Dialogue:
Lauren: Hi kau masih perlukan aku tak esok
Pam: Yes please!!
Lauren: Kau ada lagi rota?
Pam: No, tapi Manager aku balik esok so she may do some more then. Aku akan call kau in the morning and let you know.
Lauren: ok that's great
Pam: Did you have a good holiday?
Lauren: Yes, esok aku akan share ngan kau
Pam: Look forward to it!

Qwen2.5-7B summary:
Pam needs Lauren's help tomorrow. Lauren will be contacted by Pam's manager in the morning to confirm the arrangements. Pam asks Lauren if her holiday was good, and Lauren confirms it was, and she will share the details with Pam tomorrow.

Figure 4: CSL error example

6 Conclusion

In this paper, we introduced CS-Sum, the first CS dialogue-to-English summarization benchmark across three language pairs: EN-TA, EN-ZH, and EN-MS. Our experiments revealed that LLMs perform inconsistently across CS pairs, with some open-source models outperforming SEA-specific models, highlighting that CS data must be present in pretraining, not just added through fine-tuning. Furthermore, our analysis identified three major error types—CSL, MST, and SMA—persisting across all models, demonstrating that current LLMs lack intrinsic CS comprehension. The quantitative results reinforce the severity of these failures, showing that fine-tuning alone does not fully resolve CS-specific challenges. To drive further research in multilingual NLP, we will release CS-Sum along with our codebase and evaluation framework.

7 Limitations

While this work provides a comprehensive evaluation of CS dialogue summarization through the

CS-Sum benchmark, there are some limitations that should be addressed in future work.

- **Focus on Summarization:** Our experiments primarily focus on the task of summarization. However, code-switching also poses significant challenges for other long-context NLP tasks, such as machine translation, dialogue generation, and question answering. Extending the CS-Sum benchmark to these tasks would provide a broader understanding of LLMs' capabilities in handling CS data across different domains.

- **Finetuning with Synthetic Data:** The finetuning experiments in this paper are limited to synthetic code-switched data generated using Gemini-2. While this approach allowed us to assess the potential of fine-tuning for CS comprehension, it does not capture the complexities of real-world CS data.

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.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Anshul Bawa, Pranav Khadpe, Pratik Joshi, Kalika Bali, and Monojit Choudhury. 2020. Do multilingual users prefer chat-bots that code-mix? let's nudge and find out! *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1).

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

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

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Gualberto Guzmán, Joseph Ricard, Jacqueline Serigos, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2017. Metrics for modeling code-switching across corpora. In *Interspeech 2017*, pages 67–71.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. GLUECoS: An evaluation benchmark for code-switched NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Laiba Mehnaz, Debanjan Mahata, Rakesh Gosangi, Uma Sushmitha Gunturi, Riya Jain, Gauri Gupta, Amardeep Kumar, Isabelle G. Lee, Anish Acharya, and Rajiv Ratn Shah. 2021. GupShup: Summarizing open-domain code-switched conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6177–6192, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2024. SeaLLMs - large language models for Southeast Asia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 294–304, Bangkok, Thailand. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Zheng Xin Yong, Ruochen Zhang, Jessica Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Winata, Lintang Sutawika, Jan Christian Blaise Cruz, Yin Lin Tan, Long Phan, Long Phan, Rowena Garcia, Thamar Solorio, and Alham Fikri Aji. 2023. Prompting multilingual large language models to generate code-mixed texts: The case of south East Asian languages. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 43–63, Singapore. Association for Computational Linguistics.

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.

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

A Synthetic Data Generation Process

To create a synthetic CS-dialogue summarization dataset, we combined the training sets of DialogSum and SAMSum, resulting in 19,014 English dialogue-summary pairs. We used Gemini-2-flash to translate the English dialogues to code-switched dialogues in three CS pairs, as it is a large-scale SOTA LLM with a free-to-use API, making it a practical choice for our experiments.

We used the following prompt to convert English dialogues into CS dialogues. Each prompt was augmented with few-shot examples that were human generated.

Example Prompt for Malay-English Code-Switching

You are a {Malay/Tamil/Chinese} person in your 20s.
You are recruited for translating English dialogues to English-{Malay/Tamil/Chinese} code-switched dialogues.
The code-switched dialogues should follow the same structure as the English dialogue.
This is for educational purposes, so DO NOT include swear words in your translation.
Return just the code-switched dialogue.

B Experiment settings

B.1 Few-shot and Translate-Summarize

The Few-Shot and Translate-Summarize settings evaluate the inherent CS comprehension of LLMs. Evaluating the LLMs in this setting gives us an idea of their performance without any additional training.

Few-shot In the Few-Shot setup, we provide each LLM with three human-crafted examples containing code-switched dialogues and their corresponding summaries. The results from this setup, will quantify the inherent CS understanding of the LLMs for the three language pairs.

Translate-Summarize In the Translate-Summarize approach, each CS dialogue is first translated into English before being summarized. This serves as a baseline to determine whether the presence of code-switching affects summary quality. The translation step is performed by the same LLM, ensuring that the summarization model receives only monolingual input.

B.2 LoRA and QLoRA

In this subsection, we describe the fine-tuning process using *LoRA* and *QLoRA* to adapt open-source LLMs for CS dialogue summarization. Given that full fine-tuning is computationally expensive, we use parameter-efficient fine-tuning (PEFT) techniques to assess whether task-specific adaptation improves CS comprehension. The models are fine-tuned on CS-Sum-Syn.

Model	r	α	LR
Ministral-8B	32	32	5e-5
LLaMA-3-8B	16	16	5e-5
Qwen2.5-3B	64	64	3e-5
Gemma-2-9B	32	32	5e-5
SEA-Lion-Gemma-9B	32	32	5e-5
Gemma-2-2B	8	8	5e-5
Qwen-2.5-7B	64	64	3e-5
Mistral-7B	32	32	5e-5
SEALLM-7B	64	64	3e-5

Table 7: Hyperparameters used for LoRA and QLoRA fine-tuning.

The fine-tuning process follows model-specific hyperparameter configurations, as shown in Table 7. The main hyperparameters include the rank parameter (r), scaling factor (α), and learning rate (lr). These configurations were chosen based on empirical observations to balance training stability and performance.

All the models were trained for 4 epochs with a batch size of 8 and gradient accumulation for 8 steps. We followed the cosine learning rate scheduling after warming up for the first 3% of steps. We used bf16 and gradient checkpointing for memory efficient training. All the LLMs were trained on an H100 GPU and the training lasted between 1 to 4 hours depending on the architecture and size of the LLM.

C Error Analysis Process

The summaries generated by the LLMs typically exhibit the three major errors listed in Section 5. In this section, we describe the process we followed to detect these errors in the generated summaries.

We used GPT-4o to analyse the summary given the CS-dialogue for different errors. For each error type, we used a separate prompt to guide the classification process. The LLM was asked to analyse the summary with respect to the particular error type before providing a verdict on presence of the error.

Below, we present the prompts used for each category.

Prompt for CSL

You will be given a code-switched (CS) dialogue between two speakers and its generated summary. Identify whether the summary has ignored or removed key information from the non-English parts of the dialogue.
The summary has an error if it meets one or more of the following criteria:

- Does not include important content from the CS part of the dialogue.
- The summary only relies on the English part of the dialogue.

Prompt for SMA

You will be given a code-switched (CS) dialogue and its generated summary. Analyze whether the summary has mixed up who said what.
The summary has an error if it meets one or more of the following criteria:

- Has swapped roles or perspectives between speakers.
- Statements are wrongly assigned.

Prompt for MST

You will be given a code-switched (CS) dialogue and its generated summary. Analyze whether the summary distorts the meaning of the original conversation due to incorrect understanding of the code-switching.
The summary has an error if it meets one or more of the following criteria:

- The summary misinterprets the dialogue due to poor translation.
- Intent of the speakers has changed because of poor understanding of a CS phrase.
- Key points of the dialogue are misrepresented.

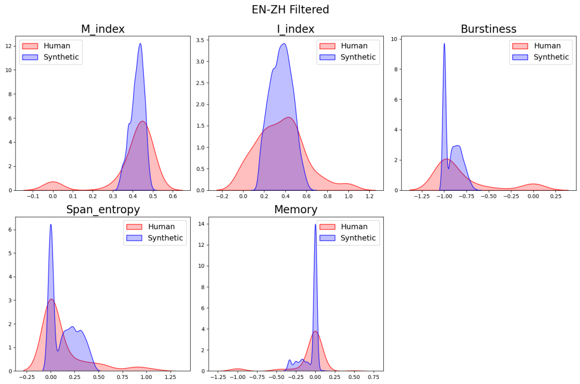


Figure 5: Distribution b/w filtered CS-Sum-Syn and CS-Sum for EN-ZH

D Training on filtered data

The difference in the distribution of CS-Sum and CS-Sum-Syn for EN-ZH (see Figure 2a) is much more than the differences in distribution for EN-MS and EN-TA. We hypothesize that this might be due to the tokenization strategy of the LLMs which is more robust for Roman script compared to Chinese script. The % improvement gains for EN-ZH after LoRA and QLoRA finetuning on EN-ZH is also less compared to the other two CS pairs (Table 5).

To check if filtering CS-Sum-Syn to match CS-Sum's distribution would improve the finetuning results, we treat the CS metrics shown in Table 1 as a multivariate gaussian. We calculated the mean $\boldsymbol{\mu}$ and the covariance matrix Σ of CS-Sum and used the Mahalanobis distance (eqn. 1) to calculate the distance between the instances in the training set and the test set.

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})} \quad (1)$$

We filtered the training data aggressively by selecting only the top 20 percentile of the training instances with the shortest distances. The filtered dataset had 3801 instances. The resulting distribution is shown in Figure 5 which matches CS-Sum's distribution better than the unfiltered one (see Figure 2a).

We finetuned (QLoRA) all the 9 open-source LLMs on the filtered data for 3 epochs (4 for the unfiltered one) with all the other parameters staying the same. Figure 6 shows the % improvement of LLMs finetuned on filtered data over the LLMs finetuned on the unfiltered one.

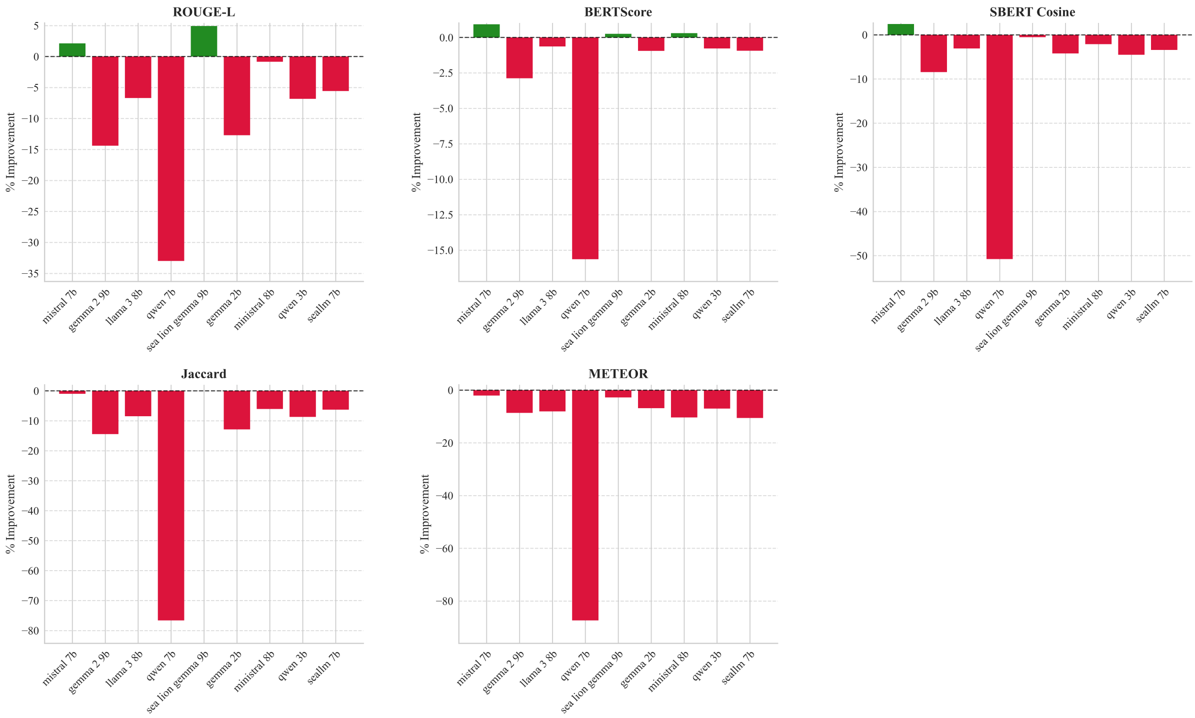


Figure 6: Filtered data % improvement over CS-Sum-Syn for EN-ZH

It can be seen that only 2 LLMs - Mistral 7B and SEA-Lion-Gemma-9B show actual improvement while other LLMs either don't improve much or have worse performances. Qwen2.5-7B suffers the worst when finetuned on filtered data. This suggests that filtering using Mahalanobis distance may only be beneficial for certain architectures while being detrimental to others. Further investigation is required to determine if alternative filtering thresholds or strategies could yield better generalization across LLMs.

Model	Lang	ROUGE-L	BERTScore	SBERT-COSINE	JACCARD	METEOR
Gemma-2-2B	EN-ZH	9.13 / 22.26	0.15 / 1.21	3.52 / 6.23	19.28 / 21.69	15.47 / 15.78
	EN-TA	21.52 / 23.29	1.62 / 1.79	12.10 / 11.86	23.88 / 29.04	19.16 / 20.98
	EN-MS	16.59 / 7.98	0.76 / 0.34	4.82 / 1.26	14.51 / 2.31	8.16 / -4.92
Qwen2.5-3B	EN-ZH	38.32 / 36.44	2.72 / 2.80	10.99 / 11.20	44.42 / 41.06	21.43 / 21.29
	EN-TA	39.08 / 35.50	2.73 / 2.55	12.30 / 11.96	47.58 / 45.49	30.90 / 33.84
	EN-MS	34.00 / 14.60	2.11 / 1.03	5.97 / 3.56	33.91 / 17.70	16.09 / 10.70
Qwen2.5-7B	EN-ZH	40.89 / 39.07	11.95 / 12.20	59.88 / 60.92	217.69 / 208.36	234.54 / 239.63
	EN-TA	22.86 / 20.85	1.83 / 1.72	11.50 / 11.10	32.72 / 30.17	35.66 / 36.26
	EN-MS	9.71 / 2.06	0.74 / -0.10	6.27 / 3.64	12.16 / 1.82	11.38 / 5.36
Mistral-7B	EN-ZH	17.91 / 2.18	0.73 / -0.80	7.78 / 2.43	28.10 / 15.19	20.10 / 9.41
	EN-TA	-1.79 / -3.62	-2.20 / -2.85	3.90 / 1.90	8.93 / 6.68	13.76 / 11.95
	EN-MS	-11.88 / 15.12	-2.93 / 0.56	-1.21 / 2.93	-8.56 / 8.96	3.83 / -2.09
Ministral-8B	EN-ZH	9.92 / 13.16	1.15 / 0.69	9.58 / 9.85	30.37 / 34.98	29.41 / 33.61
	EN-TA	18.38 / 20.88	1.41 / 1.41	13.73 / 14.13	34.19 / 36.83	35.43 / 39.83
	EN-MS	-24.12 / 7.51	-2.73 / 0.63	-1.64 / 4.51	-8.07 / 9.89	7.89 / 2.76

Table 8: LoRA / QLoRA % Improvement over Few-Shot on CS-Sum

EN Dialogue	CS Dialogue	LLM Summary	Error Explanation
Matthew: I have a cold. Need y'all support. It's painful Lisa: Fight it. Drink orange juice. Make some soup Matthew: I'll try Lisa: I believe in u ...	Matthew: Enakku cold irukku. Unga elllaroda support um vennum. Athu romba painful Lisa: Atha fight pannu. Orange juice kudi. Soup pannu Matthew: Naa try panren Lisa: Naa unna believe panran ...	Matthew is feeling unwell due to a cold and an allergy. Lisa suggests that he drink orange juice and soup to feel better. Matthew tries it and feels a bit better.	Matthew says that he will try the pill but the summary mentions that he had tried it and feels better
Anna: is anyone going to pick Mark from the airport? Marcus: i could but when and where from? Anna: Sydney, Thursday at 3 Marcus: am or pm? :D Leslie: haha fortunately pm:D Marcus: hmm i have a meeting at 1. I don't think i can make it Leslie: well i guess it will take him some time after landing, reclaiming luggage etc Anna: yeah I reckon it's fine if you're there at 4 Marcus: oh well ok then ...	Anna：有人去机场接Mark吗? Marcus：我可以，但是什么时候去哪里呢? Anna：Sydney，星期四3点 Marcus：am 还是pm？:D Leslie：haha，幸运的是下午:D Marcus：hmm 我1点有个会议。我想我去不了了 Leslie：well 我想after his landing 他需要一些时间，reclaiming luggage etc Anna：是的，我认为如果你4点到那里没问题了 Marcus：oh 那好吧 ...	Marcus originally plans to pick up Mark at the airport but has a meeting at 1 PM and can't make it. Leslie suggests Marcus arrive after 4 PM to give Mark enough time to reclaim luggage. Anna will inform Mark about Marcus's new plan.	Anna was the one who suggest Marcus to go at 4 PM which is wrongly attributed to Leslie
Ken: You are an idiot. Greg: What? Ken: Yeah you are, I want my money back Greg: Was your account hacked by some prankster? Ken: No, I'm the prankster, just having a laugh at your expense Greg: Well, you are an idiot too then XD	Ken：操你，你pimp Greg：什么？ Ken：操你，我要回我的钱 Greg：你的账户被恶作剧者hacked了吗？ Ken：不，我才是恶作剧者，只是开个玩笑以你为代价 Greg：Well，那你也操XD	Ken and Greg were joking with each other. Ken played a prank on Greg but now he wants his money back. Greg was not happy about it	The dialogue is just a friendly banter and there's actually no transaction of money.

Table 9: Error examples

EN Dialogue	CS Dialogue	LLM Summary	Error Explanation
... Person2: Large deposits of coal were found nearly and so many industries located themselves here. The village quickly grew into a key industrial centre. Person1: As the city grew, it must have absorbed many village nearby. Person2: Yes, it did. The names of those village survive as the names of parts of the city. Person1: I see. Are there any building more than 200 years old in your city? Person2: Oh, yes. Several of the buildings from the villages still survive. Many of them were inns for travelers and today survive as pubs. There was a castle near one village, so our city has a castle too. Person1: Really? So your city does have some old history after all.	... Person2: 附近发现了大量coal. 许多工业都设在这里。这个村庄迅速发展成为一个重要的industrial centre。 Person1: 随着城市发展，它一定absorbed了附近的许多村庄。 Person2: 是的。这些村庄的名字作为城市部分地区的名字而留存下来。 Person1: I see. 你的城市里有超过200 年历史的building吗? Person2: 村庄里的一些buildings仍然保存完好。其中许多都是旅客的inns，如今作为pubs保留了下来。一个村庄附近有一座castle，所以我们的城市也有一座castle。 Person1: 真的吗？原来你们城市也有悠久的history 啊。	Person2 tells Person1 how their city grew and has some buildings over 200 years old. Person2 also introduces a castle nearby	The summary is based on only the English words of the dialogue and also Person2 talks about a castle, does not introduce it
Alan: <file_photo> Alan: look what I just found Robert: dude, that's just nasty and you know it Robert: it has no sugar, no taste, and additional cinnamon flavoring Alan: yeah, I know - that's awesome Robert: you sir have a very strange tastes :P Alan: well, and I found a perfect company for it <file_photo> Robert: oh, that's more like it! Robert: but does the whiskey go well with the cinnamon? flavored whiskey is the worst... Alan: Actually it does taste surprisingly well. The cinnamon is not overpowering. If you put enough whiskey that is ...	Alan: <file_photo> Alan: tengok ape aku jumpe Robert: dude, that's just nasty and you know it Robert: it has no sugar, no taste, and additional cinnamon flavoring Alan: yeah, I know - itulah yang awesome Robert: you sir ada tastes pelik :P Alan: well, and I found a perfect company for it <file_photo> Robert: oh, that's more like it! Robert: Tapi does the whiskey go well with the cinnamon? flavored whiskey rabak... Alan: Actually it does taste surprisingly well. The cinnamon tak terlalu kuat. If you put enough whiskey that is ...	Alan shares a photo of a cinnamon-flavored whiskey with Robert, who initially finds it disgusting. Alan defends the unique taste and shares another photo of him enjoying it with a companion. Robert jokingly points out Alan's unusual taste preferences but admits the cinnamon taste isn't overpowering when paired with enough whiskey. They reminisce about a previous cherry flavor and agree it was better than the current sugar-free version.	The summary wrongly states that Alan has a companion when he was actually referring to eating cinnamon along with it

Table 10: Error examples

EN Dialogue	CS Dialogue	LLM Summary	Error Explanation
Callan: Something's wrong with my Samsung S8. Wade: What, exactly? Callan: All I get is a black screen. Wade: Did you drop it or something? Callan: I think I know, this has happened before. Wade: It did? What happened before? Callan: It overheats and then it takes a while for it to restart again.. Wade: That sucks.. Be carefull with the battery, it can be dangerous. ...	Callan: Samsung S8 aku ada problem. Wade: What, exactly? Callan: Aku asyik dapat black screen. Wade: Kau terjatuhkan ke apa? Callan: Aku rasa aku tahu kot, ni pernah happened kat aku. Wade: It did? What happened before? Callan: Dia overheat pastu it takes a while for it to restart again.. Wade: That sucks.. Be carefull with the battery, boleh jadi bahaya. ...	The speaker, Callan, is having issues with his Samsung S8 as it keeps getting black screens. Wade suggests it might be overheating. Callan intends to take it to a store as it's still under warranty	Callan was the one who suggested the phone might be overheating
Person1: What do you know about our company? Person2: Well, as I know this company is one of the largest and best in this field of business. It mainly produces toys for children. It employs more than 10, 000 people throughout the world. The president now is Mr. Jackson. The Shanghai branch was founded five years ago with a staff of more than 2, 000. Person1: Very good. If you enter this company, what department would you like to work in? Person2: May I work in sales department? Person1: OK. Oh, how much do you make at your current job, if you don't mind my asking? Person2: I'm paid 2, 000 yuan per month plus some allowances and bonus as that come to about 3, 000 yuan a month. How much will you pay me? Person1: Well, the starting salary for a clerk in our company is 2, 500 yuan monthly and after three months he would get a raise if his work is satisfactory. Person2: Do you have any fringe benefits? ...	Person1: Enga company pathi enna therium ungaluku? Person2: Indha field indha company dhaan largest and best nu therium. Main focus vandhu children toys produce panradhu. Current president Mr. Jackson. Shanghai la 5 year munnadi oru branch start panirkeenga. Anga 2000 employees irukanga. Total ah 10000 employees. Person1: Very good. Inga job kedaicha, endha department la work pananum ungaluku? Person2: Sales department la? Person1: OK. Current job la evlo earn panreenga? If you don't mind my asking? Person2: Current salary 2000 yuan plus konjam allowances and bonus ku apram 3000 yuan per month. Neenga evlo pay panuveenga? Person1: Clerks ku starting salary 2500 yuan. 3 months ku apram raise varum. Person2: Fringe benefits edhavadhu iruka? ...	Person2 is describing their current company which is the largest and best in the field, focusing on children's toys. They have a branch in Shanghai with 2000 employees, totaling 10,000 employees worldwide. Person1 is interested in a sales position and asks about the salary and benefits, learning that Person2 currently earns 3000 yuan per month with allowances and bonuses. Person1 explains that the starting salary for clerks is 2500 yuan, with a raise after three months, plus benefits like semi-annual bonuses, a small Spring Festival bonus, and 4 weeks of	LLM has completely switched Person1 and Person2 in the first half of the summary

Table 11: Error examples