EduCSW: Building a Mandarin-English Code-Switched Generation Pipeline for Computer Science Learning

Ruishi Chen*
Stanford University
ruishich@stanford.edu

Yiling Zhao*
Stanford University
ylzhao@stanford.edu

Abstract

This paper presents *EduCSW*, a novel pipeline for generating Mandarin-English code-switched text to support AI-powered educational tools that adapt computer science instruction to learners' language proficiency through mixed-language delivery. To address the scarcity of code-mixed datasets, we propose an encoder-decoder architecture that generates natural code-switched text using only minimal existing code-mixed examples and parallel corpora. Evaluated on a corpus curated for computer science education, human annotators rated 60–64% of our model's outputs as natural, significantly outperforming both a baseline fine-tuned neural machine translation (NMT) model (22–24%) and the DeepSeek-R1 model (34–44%). The generated text achieves a Code-Mixing Index (CMI) of 25.28%, aligning with patterns observed in spontaneous Mandarin-English code-switching. Designed to be generalizable across language pairs and domains, this pipeline lays the groundwork for generating training data to support the development of educational tools with dynamic code-switching capabilities.

1 Introduction

Code-switching (CSW), the practice of alternating between two or more languages within an utterance or conversation, is prevalent across diverse settings and multilingual communities (Gardner-Chloros, 2009; Poplack, 2001). Prior research has shown that CSW enables language learners to express their perspectives, convey culturally specific ideas, and build social relationships (Bhatia and Ritchie, 2006). In educational contexts, CSW has been found to enhance student engagement and help teachers clarify complex concepts, making it a valuable pedagogical strategy in multilingual classrooms (Sakaria and Priyana, 2018).

Despite its demonstrated benefits, support for code-mixing in educational tools remains limited (Yong et al., 2023). This gap is particularly pronounced in computer science education, where much of the terminology originates in English (Foote, 2023). For English-as-a-second-language learners, especially Chinese students pursuing studies abroad, this creates a dual challenge: mastering both general English and domain-specific vocabulary needed to comprehend technical content and participate in academic discourse.

Recent advances in large language models (LLMs) and speech recognition have shown potential in addressing challenges in CSW research (Giattino et al., 2023). While efforts have been made in speech translation for code-switched recognition (Alastruey et al., 2023; Wang and Li, 2023) and decoding code-mixed text (Sterner and Teufel, 2023), progress remains hindered by several issues. Studies reveal that even advanced multilingual LLMs struggle to produce natural code-switched text, often defaulting to full translation instead of authentically mixing languages (Kaji and Shah, 2023). This limitation stems from training predominantly on monolingual datasets, rather than natural code-switched corpora (Zhang et al., 2023). Moreover, challenges such as limited availability of code-mixed textual data, grammatical complexity, and domain mismatch further restrict development (Hussein et al., 2023). In particular, the lack of publicly available Mandarin-English code-mixed datasets impedes the creation of LLM-powered educational tools that support CSW.

To address these challenges, our work makes two primary contributions to CSW research. First, we introduce a generalizable pipeline for code-mixed data generation that can be adapted to various language pairs and subject domains. Second, we demonstrate its effectiveness by curating a domain-specific dataset for computer science education, focused on Chinese students studying at English-

*Equal contribution.

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 908–919

medium universities. This implementation lays the foundation for developing AI-powered tutoring systems that dynamically incorporate code-switching to support learners' acquisition of English technical language.

2 Related Work

2.1 Code-switching Background

CSW research has a relatively long history, dating back to the early 1900s (Winata et al., 2023). As the field evolved, advancements in machine learning, particularly deep learning (Gupta et al., 2020), have enabled more effective methods for both curating CSW datasets and managing various CSW tasks (Yong et al., 2023). However, the field faces considerable challenges, notably the scarcity of publicly available CSW datasets (Pratapa et al., 2018; Winata et al., 2023). Additionally, formal records of CSW texts are limited, and a significant portion of existing data is private or restricted, making it difficult to evaluate models and expand CSW research into new languages and contexts. These limitations hinder the diversification of CSW tasks and slow progress in generating comprehensive multilingual datasets.

Studies have attempted to identify key linguistic features in CSW with the goal of generating synthetic CSW data to address various challenges. Prior research has highlighted the Equivalence Constraint theory, which suggests that CSW occurs when the grammatical rules of all involved languages are maintained in a given sentence (Winata et al., 2023; Deuchar, 2020). Other works have identified the Matrix Language Frame (MLF) model (Myers-Scotton, 2001), which posits the existence of a dominant "matrix" language providing the grammatical structure, while the "embedded" language contributes additional content. This model has been proven successful in preserving syntactic features and grammatical structures from the matrix language (Callahan, 2002; Deuchar, 2006; Rahimi and Dabaghi, 2013).

2.2 Code-switching in Education

Most of the research has focused on the use of CSW in bilingual-classroom settings, suggesting its potential in enhancing instruction across subjects and improving classroom engagement. Sakaria and Priyana have identified that the use of code-switched instructional language can increase the efficiency in delivering lesson objectives and provide a theoretical framework (Sakaria and Priyana, 2018). Meanwhile, Milroy et al. also proposed that the use of code-switching can help teachers shape classroom culture, fostering different teacher-student relationships in the classroom environment (Milroy and Muysken, 1995). For instance, when teachers use the students' first language in instruction, it creates a playful and less formal environment. When the teachers switch back to the language the students are learning in that session, they reassert their authority and thus redefine the situation to be more formal.

These studies reveal the multifaceted benefits of code-switching, providing greater motivation for us to empower education by addressing the data scarcity issues in this field.

2.3 Algorithmic Solutions to Generating Code-mixed Data

Prior studies have adopted various linguistic theories and advanced language models to address the challenges in generating code-mixed texts, each reflecting distinct emphases.

For instance, Pratapa et al. (Pratapa et al., 2018) employed equivalence constraint theory, focusing on syntactic compatibility at switch points where language structures coincide. They used projections of parallel monolingual sentences to generate grammatically valid code-mixed sentences. Gupta et al. (Gupta et al., 2020) applied the Matrix Language Frame (MLF) theory, emphasizing the role of a dominant language in structuring code-mixed sentences. Tarunesh et al. (Tarunesh et al., 2021) utilized the Embedded Matrix Theory (EMT), a variation of MLF, applying clause substitution methods to create code-mixed text that satisfies Hindi-English grammatical structures.

For code-mixed data evaluation, prior scholars have proved the efficiency in various methods when assessing the naturalness of code-mixed data. Pratapa et al. (Pratapa et al., 2018) primarily assessed perplexity reductions on real code-mixed test sets using their RNN language model, which was trained on various combinations of monolingual, synthetic, and real code-mixed data. In contrast, Gupta et al. (Gupta et al., 2020) employed more direct metrics such as BiLingual Evaluation Understudy (BLEU) (Papineni et al., 2001), ROUGE (Lin and Hovy, 2002), and METEOR (Lavie and Agarwal, 2007), along with human evaluation to assess the syntactic and semantic correctness, and natural-

ness of the generated code-mixed sentences. These diverse approaches guided our team in developing appropriate validation methods for our generated synthetic texts.

3 Method

This section outlines the data source utilized for this project and then presents the generalizable code-switched generation pipeline (see Appendix A.1 for more details). The repository is publicly available[1].

3.1 Data

A representative Mandarin-English code-mixed dataset for computer science education must possess two essential characteristics.

First, the dataset should accurately represent instructional language and encompass educational materials in computer science. This provides a domain-specific context that can shape the generated code-mixed corpus to offer effective and specific support for computer science instruction.

Second, the corpus should align with how bilingual users naturally develop and use code-mixed content in educational and daily contexts. This naturalness is critical as it ensures the code-mixed text authentically reflects the language patterns observed in real-world bilingual educational settings.

Accordingly, we utilize two datasets that satisfy the above criteria in our project: a Mandarin dataset capturing domain-specific computer science instructional content, and a second dataset reflecting spontaneous code-mixing patterns in Mandarin-English speakers' daily communication.

3.1.1 Computer Science Instruction Dataset

This study incorporates the Hugging Face dataset 2imi9/llama2_7B_data_10G, which contains ten gigabytes of bilingual text data sourced from Hugging Face and the Chinese Software Developer Network (CSDN), covering technical instructions in computer science. The dataset was carefully curated to support the development of AI-powered educational tools for personalized learning in Shenzhen University's *University Computer* course. It includes a column of conceptual questions ("Instruction") and serves as the primary input for generating code-mixed representations in this study. Due to computational constraints, we used a subset of this dataset containing 744 technical instructions for computer science (file name: `data_alpaca_standardized_data`), which captures common questions and explanations of key computer science terminology.

Instruction
什么是计算机? (what is computer?)
如何解释人工智能在不同领域（如医疗、金融、教育）中的应用及其带来的影响? (How to explain the application of artificial intelligence in various fields (such as healthcare, finance, education) and the impacts it brings?)

Table 1: Sample Data Entries (The parentheses contain translations, not part of the data.)

3.1.2 Spontaneous Mandarin-English Code-Mixed Dataset

To train our model on real CSW data, we incorporated the speech transcription dataset CAiRE/ASCEND (Lovenia et al., 2022)[2] into our pipeline. We filtered the original training dataset to retain only code-mixed text, resulting in 2,739 code-mixed utterances used in this study. This subset provides a Mandarin-English code-switching corpus that reflects authentic code-switched language patterns in bilingual speakers' habits. Sample code-mixed transcriptions from this dataset are shown in Table 2.

Code-mixed Data in ASCEND
快快要期末考试了他可能觉得非常stress 非常nervous (It's getting close to the final exam. He might feel very stressed and nervous.)
放在剧情上的focus on the script but not the action but not the 特效 (Focus on the script rather than the action or the special effects.)

Table 2: Sample Mandarin-English code-mixed data (The parentheses contain translations, not part of the data.)

3.2 Pipeline

Overall, the code-mixed text generation included three key stages: the *Preparation* stage, the *Code-mixed generation* stage, and the *Evaluation* stage.

[1] https://github.com/RuishiCh-git/EduCSW/tree/main

[2] ASCEND (A Spontaneous Chinese-English Dataset) is a spontaneous multi-turn conversational dialogue recorded in Hong Kong.

3.2.1 Preparation Stage

The preparation stage includes three major steps: parallel corpus preparation, language alignment, and hard-coded code-mixed data generation.

Firstly, the machine translation model Helsinki-NLP/opus-mt-zh-en (Tiedemann and Thottingal, 2020)[3] was used to obtain the corresponding English corpus for the Mandarin computer science instruction dataset.

Secondly, the word aligner awesome-align (Dou and Neubig, 2021) was employed to create an alignment matrix for the parallel corpus. The input consists of parallel sentences separated by "|||", and the output is in the i-j Pharaoh format. A pair i-j indicates that the i-th word (zero-indexed) of the source sentence (Mandarin) is aligned to the j-th word of the target sentence (English). An example is shown in Appendix A.2.

Thirdly, a BERT-based Named Entity model and the *jieba* module were used to tokenize and extract linguistic features and tags from English and Mandarin corpus. The Matrix Language Frame (MLF)[4] was followed to generate code-switched text (Myers-Scotton, 2002). In this study, Mandarin serves as the matrix language dominating the sentence, while English is the embedded language inserted into the sentence. Named entities (NE), noun phrases (NP), and adjectives (ADJ) in the English sentence were identified as candidate words/phrases for insertion into the Chinese sentence.

For each candidate word/phrase, the language switch-point was determined based on the POS tag and position in the sentence. Insertion probabilities were set to 20% - 30% to achieve an observed code-mixing index (CMI) consistent with natural code-mixed utterances, based on prior literature (Li et al., 2012). If a switch was decided, the English word/phrase was inserted into the corresponding position in the Mandarin sentence. The resulting dataset was used as the first round of "hard-coded" CSW data.

[3]This model was developed by the Language Technology Research Group at the University of Helsinki and is designed to translate from Chinese (source language) to English (target language).

[4]MLF, proposed by Myers-Scotton, introduced the "asymmetry principle," where the language providing the morphosyntactic structure is the "matrix language," while the "embedded language" contributes elements that switch into the matrix language (Myers-Scotton, 2002)

3.2.2 Code-Mixed Generation Model

We experimented with three approaches for code-mixed data generation. The first approach extends a neural machine translation (NMT) model, serving as a baseline for comparison. The second uses the DeepSeek-R1 model to establish a benchmark performance. The third, and our primary contribution, is a custom encoder-decoder architecture designed specifically for generating natural code-switched text.

Approach 1: Fine-tuning NMT For the Neural Machine Translation (NMT) fine-tuning approach, we used the `Helsinki-NLP/opus-mt-zh-en` model[5], originally designed for Chinese-to-English translation. This model served as our baseline for generating code-switched text. It consists of approximately 77 million parameters and features an architecture with 6 encoder layers and 6 decoder layers, offering a robust foundation for capturing the complexities of both Chinese and English, as well as the nuances of code-switching patterns.

To adapt the model to our specific task, we used two primary sources of training data: the first round of "hard-coded" CSW data and code-mixed transcriptions from the ASCEND dataset (Lovenia et al., 2022). This combination was selected to balance domain-specific accuracy with the naturalness of authentic code-switching.

The model was fine-tuned over 3 epochs, using a learning rate of 2e-5 and a batch size of 16 per device. These parameters were chosen to ensure adequate adaptation to the code-switching task while minimizing the risk of overfitting. The fine-tuning concluded with a final training loss of 0.709, indicating a solid trade-off between specialization and generalization. The resulting model is publicly available[6].

Approach 2: DeepSeek-R1 Benchmark To benchmark our code-mixed text generation pipeline against a strong pre-trained baseline, we utilized Distilled DeepSeek-R1 7B, based on Qwen—a large language model trained on both Chinese and English corpora (DeepSeek-AI, 2025). DeepSeek has demonstrated remarkable performance across a range of Chinese natural language understanding and generation tasks, making it a valuable reference point for evaluating code-switching capabili-

[5]https://huggingface.co/Helsinki-NLP/opus-mt-zh-en

[6]https://huggingface.co/yl31/code-mixed-cs-edu-model

ties. Although DeepSeek is not explicitly trained for code-switching, it offers insight into how well general-purpose, state-of-the-art language models can handle code-mixing in the absence of domain-specific supervision. As such, this benchmark serves as a reasonable point of comparison for our customized generation pipeline.

We adopted a few-shot prompting strategy to guide DeepSeek toward producing Mandarin-English code-switched output. Each prompt included two illustrative examples demonstrating how to naturally integrate English computer science terminology into Mandarin instructional sentences. These examples showcased both noun phrase-level and verb-level switches—patterns commonly observed in bilingual academic discourse. The complete prompt is provided in Appendix A.5. This prompt was applied to all 744 Mandarin instructional sentences in our dataset. Model outputs were collected without any post-processing to preserve their authenticity for subsequent evaluation.

Approach 3: Encoder-Decoder Architecture For the encoder-decoder architecture model, the rationale is to use the encoder to provide context while the decoder generates target sequences with a copy mechanism, improving model performance through a combination of translation and copying from input text.

We first leverage **transfer learning** to initiate our code-mixed generation model. This approach aims to reduce the required training data for code-mixed generation while ensuring high-quality bilingual representations essential for natural code-switching data generation. Specifically, we fine-tune the neural machine translation model Helsinki-NLP/opus-mt-zh-en on our curated parallel corpus of computer science educational content and dialogue. The fine-tuning process enables the model to capture language-specific features, including domain-specific terminology and language patterns unique to computer science education in both Mandarin and English, as well as cross-lingual mappings, such as semantic equivalences and contextual relationships between the language pairs.

The weights learned during this fine-tuning phase provide monolingual understanding and capture cross-lingual feature characteristics. The next step is to use an **encoder-decoder architecture** that builds on the fine-tuned weights to integrate additional components extracted from the preliminary code-mixed dataset to build the code-mixed text generation model.

The **encoder**, built on the transformer layers of the MarianMTModel, processes the sequences of tokens in Chinese texts to produce hidden states that capture sequential dependencies and generate contextual representations for the sentences. These representations are then received by the attention mechanism in the decoder, allowing the model to have more focused access to relevant source information. This enables the preservation of both language-specific features and cross-lingual relationships.

Subsequently, the **decoder** uses a processing mechanism to adopt a standard decoder path for translation logits and a dedicated gate mechanism for copy probability calculation. With the attention mechanism, the encoder's representations are processed to produce hidden states, which inform both generation and copying decisions. When copying from the input texts is decided, the model computes copy probabilities for the input tokens. Subsequently, the model expands input tokens to align with the target sequence length and then maps the tokens into the known vocabulary space using scatter operations, locating the vocabulary tokens in the input text. Such a mechanism is important to preserve technical terminology for conversational corpus related to Computer Science, where many words tend to co-occur for domain-specific meanings. For example, with the term 'neural network,' the model can directly copy these tokens rather than regenerate words for "network" or "neural" to maintain precise technical accuracy.

With the encoder-decoder architecture built, we optimize our operation with a specialized loss function that combines loss with a mixing ratio penalty. In particular, we incorporate a Code-Mixing Loss function to calculate the ratio of Chinese to English tokens and penalize the outputs that deviate from a ratio of 0.5 (set for a minimal mixing ratio). This approach preserves semantic accuracy within the code-mixed dataset while encouraging the model to learn from the trained dataset and generate balanced code-mixing data.

During training, the model processes both the hard-coded CSW data and the transcriptions from the ASCEND dataset (Lovenia et al., 2022). The training setup uses parallel data: the original Mandarin text serves as input, while the corresponding code-mixed versions (both hard-coded and AS-

CEND transcripts) serve as the target outputs. The generation strategy employs beam search with a beam width of 5, meaning it maintains the top 5 most probable sequences at each decoding step. Then, the model uses a 2-gram prevention strategy to prevent two consecutive tokens from appearing more than once in the generated sequence. These parameters were chosen to maintain output diversity and technical accuracy while preventing common generation issues like repetitive text.

4 Results and Evaluation

4.1 Description of Generated CSW Data

The 744 Mandarin text entries from the 2imi9/llama2_7B_data_10G dataset were used as input for all three of our code-mixed generation models: the fine-tuned NMT approach, DeepSeek-R1, and the encoder-decoder architecture. This parallel processing enabled the generation of three distinct sets of code-switched (CSW) data, facilitating a comparative analysis across methods.

The generated CSW text preserves the educational content and structure of the original Mandarin entries while incorporating English elements in a way that reflects natural code-switching patterns commonly observed in bilingual educational contexts.

4.2 Evaluation

4.2.1 Code-Mixing Index

The Code-Mixing Index (CMI) (Das and Gambäck, 2014) is a widely used metric for measuring the complexity of code-mixed text (Srivastava and Singh, 2021). It quantifies the fraction of tokens or words that differ from the matrix language[7]. In our study, we calculated the sentence-level CMI [8] by dividing the number of English tokens by the total word count in each CSW sentence.

The overall CMI for each generated CSW dataset was computed as the average of all sentence-level CMIs within that dataset. As presented in Table 3, the CMI for the hard-coded first round of generated CSW data is 26.98%. The CMIs for the NMT fine-tuning, DeepSeek-R1, and encoder-decoder approaches are 23.05%, 9.95%, and 25.28%, respectively.

Notably, the CMIs for most of our generated CSW datasets fall within the 20% to 30% range,

Method	**Matrix Lang.**	**CMI**
Hard Code	/	26.89%
NMT Fine-tuning	Chinese	23.05%
Deepseek R1	Chinese	9.95%
Encoder/Decoder	Chinese	25.28%

Table 3: CMIs for Different Methods

which aligns with values observed in spontaneous Chinese-English code-switching utterances from prior studies (see Appendix A.3). This suggests that our generated CSW data—excluding the output from DeepSeek-R1—closely mirrors natural code-mixing patterns, reinforcing the credibility and authenticity of the synthetic text. The substantially lower CMI of DeepSeek-R1 (9.95%) indicates limited code-switching behavior, which may reduce its effectiveness for simulating natural bilingual communication.

4.2.2 Human Labeling

To comprehensively evaluate the quality of the generated data, we recruited two bilingual annotators to label the CSW outputs from the NMT model, DeepSeek-R1, and the encoder-decoder framework. Both annotators were proficient in Mandarin-English code-mixing and had familiarity with domain-specific computer science terminology. They were instructed to rate the naturalness of each sentence using a standardized 3-point Likert scale (Joshi et al., 2015): unnatural (1), acceptable (2), and natural (3). If a sentence contained nonsensical segments that severely disrupted its meaning, annotators could label it as "wrong," in which case it was excluded from the naturalness evaluation.

Each annotator labeled 50 entries from each of the three models. These entries were derived from 50 randomly sampled Chinese input sentences. To assess annotation consistency, we calculated inter-rater reliability using Cohen's kappa coefficient (Blackman and Koval, 2000). The resulting κ values were 0.6739 for the fine-tuned NMT model, 0.6793 for the encoder-decoder model, and 0.7622 for DeepSeek-R1—indicating moderate to strong agreement between annotators.

We then compared the performance of the three models in generating natural CSW outputs. Table 4 presents the percentage of outputs rated as natural. The encoder-decoder approach significantly outperformed both the fine-tuned NMT and DeepSeek-R1

[7] https://tech.skit.ai/Code-Mixing-Metrics/
[8] See Appendix A.4 for CMI formula.

models. Annotators consistently rated a higher proportion of encoder-decoder outputs as natural (64% and 60%) compared to those from the NMT model (22% and 24%) and DeepSeek-R1 (34% and 44%).

Labeler	Fine-tuned NMT	DeepSeek R1-Distill	Encoder Decoder
1	22%	44%	64%
2	24%	34%	60%

Table 4: Comparison of Natural Output Percentages by Annotators

Sentences annotated as natural typically demonstrated preservation of the grammatical rules of the matrix language (Mandarin) and exhibited switches at technical terms and language-sensitive words (words more commonly used in English). For instance, in the example shown in Table 5, the technical terms "自然语言处理" and "机器学习" in the input Mandarin sentence were switched to English expressions "language processing" and "machine learning" respectively, and the resulting sentence was labeled as natural.

Conversely, sentences labeled as unnatural often disobeyed Mandarin grammar and displayed issues such as incomplete semantic segments, mistranslations, or unbalanced proportions of Mandarin and English segments. Examples of such cases are also provided in Table 5.

Input	Output and Label	
在自然语言处理(NLP)中，如何利用机器学习进行情感分析？请描述其过程和应用场景。 (In NLP, how to utilize machine learning for sentiment analysis? Please describe the process and application scenarios.)	在language processing(nlp)中，如何利用machine learning进行emotional analysis?请描述其过程and application场景. (In NLP, how can machine learning be utilized for emotional analysis? Please describe the process and application scenarios.)	*Natural*
	在language processing(nlp)中，如何用machine learning进analysis? please deplecation of中的processing and application processing. (In NLP, how can we use machine learning for analysis? Please clarify the meanings of "processing" and "application processing" in "depletion of".)	*Unnatural*

Table 5: Comparison of Natural and Unnatural Labels (The parentheses contain translations, not part of the data.)

4.2.3 Qualitative Evaluation

To further assess the quality of the generated code-switched text, we conducted a qualitative evaluation of outputs from the fine-tuned NMT approach, DeepSeek-R1, and the encoder-decoder framework. This analysis revealed clear differences in code-switching quality among the three methods.

The encoder-decoder framework demonstrated a superior ability to generate natural and coherent code-switched text. As shown in Appendix A.6, its outputs exhibit several favorable characteristics. The code-switched segments primarily consist of noun phrases and computer science-related terms in English, reflecting authentic bilingual speech patterns. Language switch points appear more natural and intuitive, and grammatical structures in both languages are better preserved, resulting in higher overall linguistic quality.

In contrast, the fine-tuned NMT model showed notable limitations. As illustrated in Appendix A.6, its outputs often exhibit grammatical inconsistencies when transitioning between English and Chinese. Additionally, it occasionally produces nonsensical or incoherent English terms (e.g., "converence," "protology," "diploration"), leading to awkward transitions and a lower degree of naturalness compared to the encoder-decoder output.

DeepSeek-R1, a large language model trained on Chinese text, also displayed weaknesses in generating natural code-switching. Many outputs defaulted to full English translations rather than producing genuine code-switched language, resulting in a low Code-Mixing Index (CMI) and limited alignment with real-world bilingual discourse. While DeepSeek-R1 occasionally produced natural-sounding examples, its performance was inconsistent, and it was outperformed overall by the encoder-decoder framework.

In summary, the qualitative evaluation shows that the encoder-decoder model consistently generates more natural, coherent, and contextually appropriate code-switched text than both the fine-tuned NMT and DeepSeek-R1 approaches. Its outputs closely mimic authentic bilingual communication, particularly in technical domains, and exhibit a balanced and grammatically sound integration of

English terminology.

5 Discussion & Conclusion

In this study, we developed a comprehensive, effective, and reusable pipeline for generating synthetic code-mixed data, with the goal of supporting the training of human-centered tutoring large language models (LLMs) and chatbots that communicate using a code-mixed approach. This work is motivated by the pedagogical value of code-mixed instruction for bilingual learners adapting to second-language environments. At the same time, existing publicly available LLMs show limited proficiency in handling code-switching, often focusing narrowly on topic-related nouns (Yong et al., 2023). In addition to proposing a general pipeline, we apply it to create a Mandarin-English code-mixed dataset specifically curated for computer science education.

We accomplished two key objectives:

First, we successfully developed a generalizable pipeline for generating code-mixed data across language pairs (with English as one of the languages). The pipeline consists of three main steps: (1) generating preliminary synthetic code-switched data using the Matrix Language Frame (MLF) theory and BERT-based Named Entity Recognition to prepare the non-English monolingual data; (2) passing the text through an encoder-decoder architecture initialized with weights from an NMT model fine-tuned on a parallel corpus, and training it using both the synthetically generated and real code-mixed data; and (3) iteratively annotating and retraining to enhance the naturalness of the generated outputs.

To adapt the pipeline for other language pairs, users only need to modify two components: (1) the Matrix Language Frame to match the grammatical structure of the target language, and (2) the code-switched speech transcription dataset, which is often more readily available than textual resources. With these changes, users can input their own monolingual data and generate suitable code-mixed datasets for downstream tasks.

Second, we successfully curated a domain-specific code-mixed dataset for computer science education that can support downstream training of LLMs or chatbots. This dataset was validated through three evaluation methods: the Code-Mixing Index (CMI), human ratings, and qualitative analysis. Across all measures, our encoder-decoder architecture outperformed both the state-of-the-art DeepSeek LLM and a traditional fine-tuned neural machine translation model in generating natural code-switched text.

We offer two suggestions based on our findings. First, given the success of our pipeline in the computer science domain, we recommend applying this approach in other STEM fields where technical vocabulary creates challenges for bilingual learners (Bhatia and Ritchie, 2006). Second, we encourage the development of interactive tutoring systems and LLM-powered chatbots using our curated dataset and pipeline, with the capacity to dynamically adjust the degree of code-mixing based on learners' language proficiency. As supported by prior work (Milroy and Muysken, 1995), flexible language use in educational settings can greatly enhance learner engagement and comprehension.

5.1 Limitation and Future Work

We identify two limitations in this study.

First, although we use transcriptions from a code-mixed audio dataset to fine-tune the naturalness of our model's outputs, the ASCEND training dataset occasionally contains spelling errors, incomplete sentences, and casual conversational utterances. These issues may affect the quality of the generated code-mixed text. Future researchers may improve results by further cleaning and curating a high-quality subset of the transcription data or by sourcing data from more professional or domain-relevant contexts.

Second, due to the nature of the fine-tuned NMT model being primarily designed for translation tasks, it occasionally produces fully translated English output. This indicates that the model's control over the language mixing ratio is not yet optimal. Future work could explore increasing the number of training iterations and implementing a feedback loop to monitor and dynamically adjust the language balance during generation, thereby enhancing the consistency and naturalness of code-switching.

6 Ethical Consideration

Our primary data source, the Mandarin instructional dataset for computer science learning, is open-sourced on Hugging Face and explicitly designed to improve AI model performance in educational settings. Our use aligns with this stated purpose, and we have properly cited the source. Similarly, the ASCEND dataset, used for code-

mixing patterns, is open-sourced and appropriately cited. For annotation, we engaged voluntary participants, ensuring ethical practices in data labeling.

The primary application of our work is developing AI-powered tutoring chat bots for personalized computer science learning, bridging the gap for bilingual learners transitioning from Mandarin to English-language education. We acknowledge the need to preserve language integrity, respect cultural nuances, and avoid exacerbating educational disparities.

References

Belen Alastruey, Matthias Sperber, Christian Gollan, Dominic Telaar, Tim Ng, and Aashish Agarwal. 2023. Towards real-world streaming speech translation for code-switched speech. *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, page 14–22.

Tej K. Bhatia and William C. Ritchie. 2006. *The Handbook of Bilingualism*. Blackwell Pub.

Nicole J-M Blackman and John J Koval. 2000. Interval estimation for cohen's kappa as a measure of agreement. *Statistics in medicine*, 19(5):723–741.

Laura Callahan. 2002. The matrix language frame model and spanish/english codeswitching in fiction. *Language Communication*, 22(1):1–16.

Amitava Das and Björn Gambäck. 2014. Identifying languages at the word level in code-mixed indian social media text.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Margaret Deuchar. 2006. Welsh-english code-switching and the matrix language frame model. *Lingua*, 116(11):1986–2011. Celtic Linguistics.

Margaret Deuchar. 2020. Code-switching in linguistics: A position paper. *Languages*, 5(2):22.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Keith D. Foote. 2023. A brief history of large language models.

Penelope Gardner-Chloros. 2009. *Code-switching*. Cambridge University Press.

Charlie Giattino, Edouard Mathieu, Veronika Samborska, and Max Roser. 2023. Artificial intelligence. *Our World in Data*. Https://ourworldindata.org/artificial-intelligence.

Deepak Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2020. A semi-supervised approach to generate the code-mixed text using pre-trained encoder and transfer learning. *Findings of the Association for Computational Linguistics: EMNLP 2020*, page 2267–2280.

Amir Hussein, Shammur Absar Chowdhury, Ahmed Abdelali, Najim Dehak, Ahmed Ali, and Sanjeev Khudanpur. 2023. Textual data augmentation for arabic-english code-switching speech recognition. *2022 IEEE Spoken Language Technology Workshop (SLT)*.

Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396–403.

Arshad Kaji and Manan Shah. 2023. Contextual code switching for machine translation using language models. *arXiv preprint arXiv:2312.13179*.

Alon Lavie and Abhaya Agarwal. 2007. Meteor. *Proceedings of the Second Workshop on Statistical Machine Translation - StatMT '07*, page 228–231.

Ying Li, Yue Yu, and Pascale Fung. 2012. A mandarin-english code-switching corpus. In *LREC*, pages 2515–2519.

Chin-Yew Lin and Eduard Hovy. 2002. Manual and automatic evaluation of summaries. *Proceedings of the ACL-02 Workshop on Automatic Summarization* - , 4:45–51.

Holy Lovenia, Samuel Cahyawijaya, Genta Indra Winata, Peng Xu, Xu Yan, Zihan Liu, Rita Frieske, Tiezheng Yu, Wenliang Dai, Elham J Barezi, et al. 2022. Ascend: A spontaneous chinese-english dataset for code-switching in multi-turn conversation. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*.

Dau-Cheng Lyu, Tien Ping Tan, Engsiong Chng, and Haizhou Li. 2010. Seame: a mandarin-english code-switching speech corpus in south-east asia. In *Interspeech*, volume 10, pages 1986–1989.

L. Milroy and P.C. Muysken. 1995. *One speaker, two languages: Cross-disciplinary perspectives on code-switching*. Cambridge University Press.

Carol Myers-Scotton. 2001. The matrix language frame model: Development and responses. *Trends in Linguistics Studies and Monographs*, 126:23–58.

Carol Myers-Scotton. 2002. *Contact linguistics: Bilingual encounters and grammatical outcomes*. Oxford University Press, USA.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, page 311.

S. Poplack. 2001. Code switching: Linguistic. *International Encyclopedia of the Social amp; Behavioral Sciences*, page 2062–2065.

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. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 1543–1553.

Meisam Rahimi and Azizollah Dabaghi. 2013. Persian–english codeswitching: A test of the matrix language frame (mlf) model. *System*, 41(2):322–351.

Severinus Sakaria and Joko Priyana. 2018. Code-switching: A pedagogical strategy in bilingual classrooms. *American Journal of Educational Research*, 6(3):175–180.

Vivek Srivastava and Mayank Singh. 2021. Challenges and limitations with the metrics measuring the complexity of code-mixed text. *arXiv preprint arXiv:2106.10123*.

Igor Sterner and Simone Teufel. 2023. Tongueswitcher: Fine-grained identification of german-english code-switching. *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, page 1–13.

Ishan Tarunesh, Syamantak Kumar, and Preethi Jyothi. 2021. From machine translation to code-switching: Generating high-quality code-switched text. *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)*, page 3154–3169.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Qinyi Wang and Haizhou Li. 2023. Text-derived language identity incorporation for end-to-end code-switching speech recognition. *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, page 33–42.

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

Zheng-Xin Yong, Ruochen Zhang, Jessica Zosa Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Indra Winata, Lintang Sutawika, Jan Christian Blaise Cruz, Yin Lin Tan, 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.

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.

A Appendix

A.1 Pipeline Flowchart

Note: The pipeline flowchart shown on the next page (Appendix: 1) illustrates our overall approach.

A.2 Example of input and output for word alignment using awesome-align

Type	Content
Input	我 喜欢 吃苹果 (zh) ||| I like to eat apples (en)
Output	0-0 1-1 2-3 2-4

A.3 Reference CMI values from literature

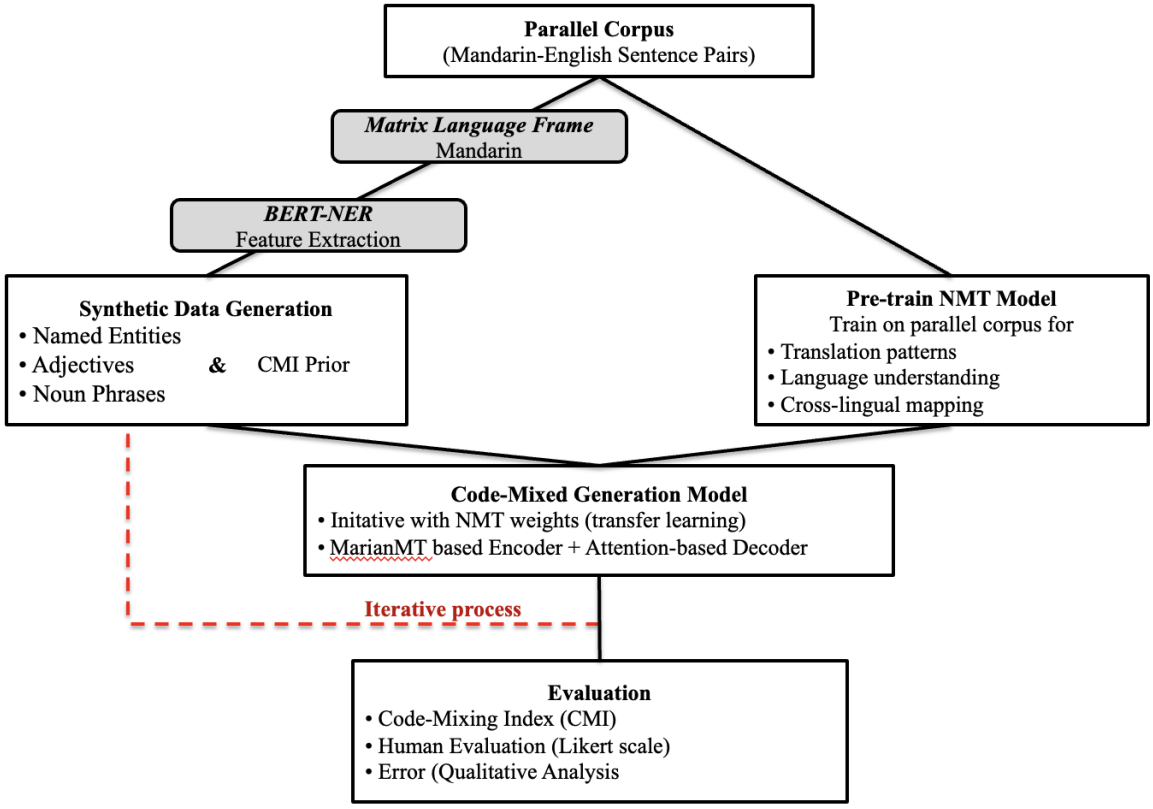
Reference	Matrix Language	CMI
(Li et al., 2012)	Chinese	21.15%
(Lyu et al., 2010)	Chinese	25%

A.4 CMI Formula

The CMI is calculated using the following formula:

$$CMI = 100 * \left(1 - \frac{\max(w_i)}{n-u}\right) \text{ if } n > u \quad (1)$$

where w_i is the number of words in language i (English), n is the total number of words, and u is the number of language-independent words.



Appendix A.1: Overall Pipeline. This flowchart shows the steps involved in the code-mixed generation model.

A.5 DeepSeek Prompt Template

Prompt

<system>
You are a helpful assistant. Your job is to convert Mandarin computer science questions into Mandarin-English code-switched sentences that sound natural to bilingual learners.
Only output the sentence. Do not explain or comment.
<user>
Input: 在深度学习中，如何训练卷积神经网络?
Output: 在deep learning中，如何train convolutional neural network?
Input: 什么是计算机网络的拓扑结构?
Output: 什么是computer network 的 topology 结构?
Now process the following:
Input: {text}
Output:

A.6 Comparison of Encoder-Decoder and NMT Generated Outputs (The parentheses contain translations, not part of the data.)

Output	Label
Encoder-Decoder Generated	
什么是data consistence? (What is data consistence?)	Natural
深度learning 中curly network (cnn) 如何实现image 分类and 对象检测? 请详细解释其working principles and technologies. (How does the curly network (CNN) in deep learning achieve image classification and object detection? Please elaborate on its working principles and technologies.)	Natural
Fine-tuned NMT Generated	
什么是data converence? (What is data converence?)	Wrong
深度学习中的blough network (CNN) 如何实现image diagation and operation processing? please process processing work chrinkings and key processings. (How does the blough network (CNN) in deep learning achieve image diffusion and operation processing? please process processing work chrinkings and key processings.)	Acceptable
DeepSeek-R1-Distill	
什么是data consistency? (What is data consistency?)	Natural
How to train a deep learning model to recognize cats and dogs in images?	Wrong