Optimizing Code-Switching in Conversational Tutoring Systems: A Pedagogical Framework and Evaluation

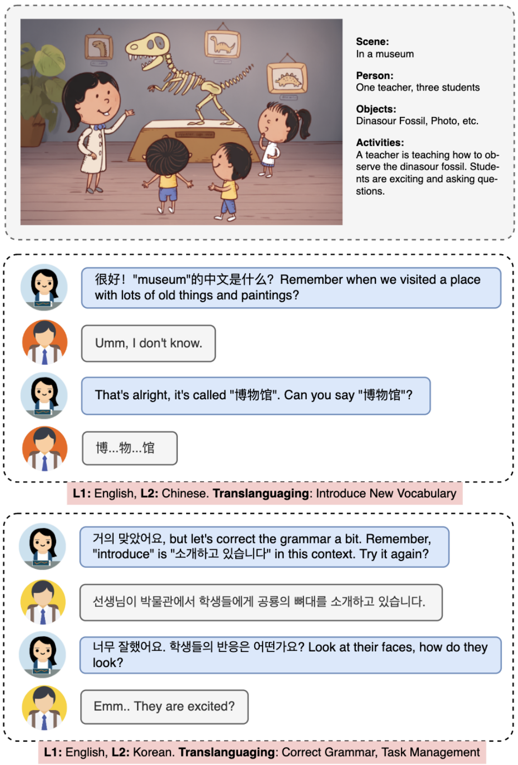
Zhengyuan Liu[†*], **Stella Xin Yin**[‡*], **Nancy F. Chen**[†]
[‡]Nanyang Technological University, Singapore
[†]Institute for Infocomm Research (I^{2}R), A*STAR, Singapore
`liu_zhengyuan@i2r.a-star.edu.sg` `nfychen@i2r.a-star.edu.sg`

Abstract

Large language models demonstrate remarkable proficiency in various tasks across multiple languages. However, their potential in code-switching remains underexplored, particularly in cultural and educational contexts. Code-switching or translanguaging plays a crucial role in bilingual education, facilitating comprehension and engagement among students with varied language proficiency. In this work, we present a pedagogy-inspired framework that introduces traditional classroom practices of code-switching to intelligent tutoring systems. Specifically, we develop fine-grained instructional strategies tailored to multilingual and educational needs. We conduct experiments involving both LLM-based evaluation and expert analysis to assess the effectiveness of translanguaging in tutoring dialogues. Our experimental results indicate that strategic code-switching can significantly enhance the learning experience. This work not only advances dialogic tutors in language learning but also extends LLMs to better accommodate multilingual interaction.

1 Introduction

Large Language Models (LLMs) excel in diverse tasks and various languages (Ouyang et al., 2022; Wang et al., 2023a). While their task-solving capabilities in monolingual scenarios are well-studied (Zheng et al., 2024), their potential in code-switching - the practice of alternating languages within an utterance - is still less explored (Zhang et al., 2023b). In multilingual communications, people sometimes switch languages during the conversation to convey context-specific concepts and reinforce social connections. However, current LLMs aren't specifically trained for translanguaging scenarios, highlighting an emerging research interest in their ability to understand and utilize code-switching (Doğruöz et al., 2023).



Scene:
In a museum

Person:
One teacher, three students

Objects:
Dinasour Fossil, Photo, etc.

Activities:
A teacher is teaching how to observe the dinasour fossil. Students are exciting and asking questions.

很好！"museum"的中文是什么？Remember when we visited a place with lots of old things and paintings?

Umm, I don't know.

That's alright, it's called "博物馆". Can you say "博物馆"?

博...物...馆

L1: English, **L2:** Chinese. **Translanguaging**: Introduce New Vocabulary

거의 맞았어요, but let's correct the grammar a bit. Remember, "introduce" is "소개하고 있습니다" in this context. Try it again?

선생님이 박물관에서 학생들에게 공룡의 뼈대를 소개하고 있습니다.

너무 잘했어요. 학생들의 반응은 어떤가요? Look at their faces, how do they look?

Emm.. They are excited?

L1: English, **L2:** Korean. **Translanguaging**: Correct Grammar, Task Management

Figure 1: Examples of conversation segments in language learning using pedagogical code-switching.

Code-switching is not only relevant in the realm of natural language processing but also has significant implications in educational settings (Lin, 2013). For instance, bilingual children often have an imbalanced exposure to their first languages (**L1**) at home, and show less proficiency in vocabulary, grammar, and sentence structures of the target language (**L2**) at school. As a result, they tend to switch between **L1** and **L2** in classrooms. Accordingly, as shown in Figure 1, teachers apply code-switching strategies to clarify instructions, introduce new concepts, provide necessary encouragement, and facilitate the learning process (Rabbidge, 2019). These efforts are termed as “translanguaging” (Williams, 1994) or “code-switching” in pedagogical theories, referring to the planned and sys-

---

[*] Equal contribution.

Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 500–515
September 18–20, 2024. ©2024 Association for Computational Linguistics

tematic use of two languages in the classroom, and it includes both intra-sentential and inter-sentential switching (Lin, 2016).

While code-switching in tutoring has been studied extensively in the last decades (De La Cruz, 2019; Turnbull and Arnett, 2002; De La Campa and Nassaji, 2009; Littlewood and Yu, 2011), most previous works focus on traditional classrooms. On the other hand, while Intelligent Tutoring Systems (ITSs) have shifted adaptive and personalized education from traditional classrooms to online learning, they are often limited to monolingual contexts and lack adaptability to mixed bilingual communication. In particular, one major challenge of improving code-switching of dialogue systems is the scarcity of data (Ahn et al., 2020; Doğruöz et al., 2023). Due to the highly colloquial nature of translanguaging, existing resources for specific domains are limited, and collecting data at a large scale requires considerable annotation efforts. Furthermore, previous studies evaluate multilingual models in general-purpose code-switching scenarios (Tan and Joty, 2021; Adilazuarda et al., 2022) and they often simulate mixed generation by replacing words in parallel corpora. However, this approach fails to capture the complexity of effective pedagogical code-switching, which requires strategic and purposeful integration that considers learners' proficiency levels and educational objectives. Thus, pedagogical code-switching should go beyond mere word swapping to facilitate comprehension of complex concepts and provide instructional scaffolding.

In this work, we aim to improve the generated code-switching in conversational tutoring systems. We first gain insights from translanguaging theories and empirical dialogue studies to construct a pedagogical code-switching framework, where each dimension combines relevant scaffolding strategies to enhance language teaching through targeted translanguaging interventions. Our framework contributes to language learning by facilitating vocabulary acquisition, grammatical understanding, and conversational fluency. It also supports content mastery through concept clarification and emotional support.

To anchor a practical application, we conduct a case study on image description for language learning, and leveraging LLMs as tutoring agents. We utilize our proposed framework in both instructing LLMs for fine-grained code-switching generation and assessment. We conduct experiments on two representative translanguaging cases (e.g., Chinese-English, Korean-English), and deploy automated evaluation and qualitative analysis to assess the effectiveness of pedagogical code-switching generation in LLM-based tutoring systems. Experimental results indicate that state-of-the-art models are capable of tailoring scaffolding actions and code-switching to learners' language proficiency levels and teaching content, and strategic code-switching can significantly enhance the learning experience.

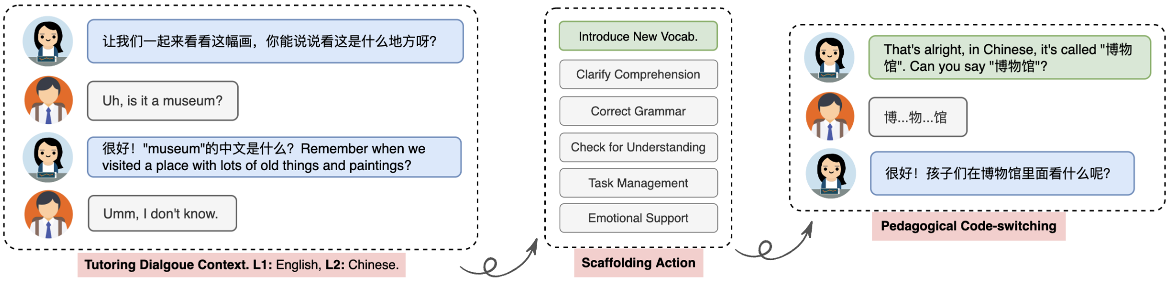
2 Related Work

2.1 Code-switching in Dialogue Systems

Dialogue systems (conversational agents) are designed to imitate various human linguistic and behavioral patterns (Chen et al., 2017), including the language mixing patterns of multilingual users (Parekh et al., 2020; Bawa et al., 2020). Rule-based approaches use linguistic features such as discourse makers and templates to produce bilingual utterances via word replacement (Ahn et al., 2020). Data-driven methods can achieve higher flexibility, beyond simple lexical borrowing to blending of languages at syntactic, grammatical, and morphological levels (Doğruöz et al., 2023; Liu et al., 2022), but they heavily rely on well-annotated data. While recent LLMs show strong multilingual capability and share certain knowledge across various languages (Wang et al., 2023a), their potential for coherent translanguaging is still underdeveloped (Zhang et al., 2023b), and the exploration to dialogic tutoring remains limited (Choi et al., 2023).

2.2 Code-switching in Education

Research on code-switching in classrooms dates back to the 1970s. Jacobson (1981) proposed New Concurrent Approach (NCA), as the first bilingual pedagogy, introducing flexible bilingual language practices for children and youth. The pedagogical aspects of language mixing affirm code-switching as a viable approach to bilingual teaching and learning (Hornberger and Link, 2012). The practice of code-switching (intentional instructional strategies that integrate two or more languages in real classrooms) has been well-studied (Lin, 2013; Sinclair and Fernández, 2023; García, 2009; Bon, 2021), which reveals that code-switching can significantly enhance learning and can be used as a strategic teaching method (Moore, 2002; Cenoz and Gorter, 2022; Barahona, 2020; Vaish and Subhan, 2015). Studies in language education further confirmed



让我们一起来看看这幅画，你能说说看这是什么地方呀?

Uh, is it a museum?

很好！"museum"的中文是什么？Remember when we visited a place with lots of old things and paintings?

Umm, I don't know.

That's alright, in Chinese, it's called "博物馆". Can you say "博物馆"?

博...物...馆

很好！孩子们在博物馆里面看什么呢?

Figure 2: Decomposed pedagogical code-switching process, where the translanguaging is featured on scaffolding.

that the use of **L1** can facilitate language acquisition, improve student engagement, and establish rapport between teacher and students in **L2** learning (Pan and Pan, 2010). Besides, these studies emphasized the key role of teachers in managing teacher-centered Initiation-Response-Feedback (IRF) (Wells, 1999) sequences that promote the use of code-switching in the classroom.

2.3 Intelligent Tutoring Systems

The advancement of ITSs has marked a significant step forward in education practice (Graesser et al., 2018; Demszky and Hill, 2023; Wang et al., 2023b). These systems provide personalized learning experiences and instant feedback (Chaffar and Frasson, 2004; Harley et al., 2015; Grivokostopoulou et al., 2017), tailored to learners' characteristics and needs (Dzikovska et al., 2014; Grawemeyer et al., 2016; Nihad et al., 2017), and are shown to positively influence students' engagement in learning and academic performance (Kulik and Fletcher, 2016; Xu et al., 2019).

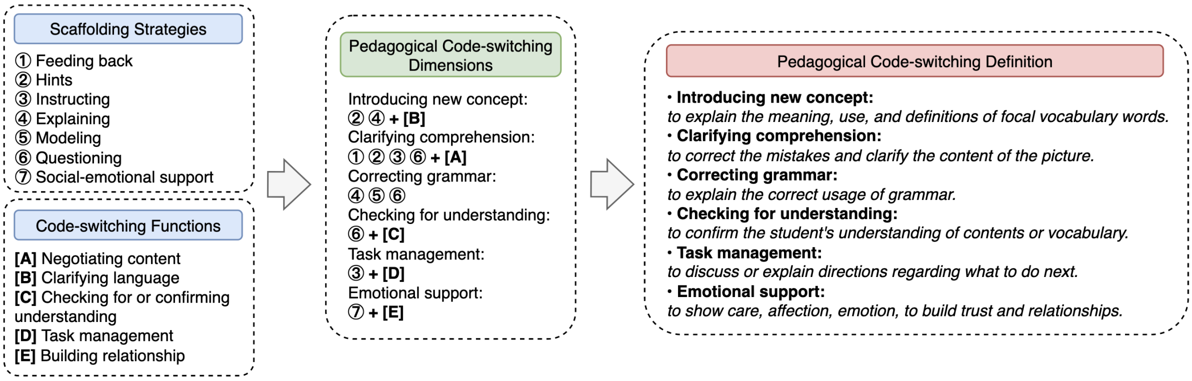
Dialogue tutor is a particular type of intelligent tutoring system that interacts with students via natural language conversation (Nye et al., 2014; Ruan et al., 2019). In STEM domains, conversational ITSs can facilitate university students in problem-solving by providing real-time feedback and hints in text formats (Nye et al., 2023; Paladines and Ramirez, 2020; Arnau-González et al., 2023). However, prior work has widely relied on rule-based systems with human-crafted domain knowledge (Nye et al., 2014; Graesser et al., 2018), or data-driven approaches that require a certain amount of human annotation for supervised learning (MacLellan and Koedinger, 2022). Recently, LLMs show strong potential to build dialogue tutors with less data supervision and higher coherence (Afzal et al., 2019; Demszky and Hill, 2023; Macina et al., 2023b), and they can be further improved by integrating LLMs with pedagogical and learning science principles (Stasaski et al., 2020; Sonkar et al., 2023; Macina et al., 2023a).

3 Pedagogical Code-switching

We conceptualize "pedagogical code-switching" as a combination of two aspects: scaffolding and translanguaging, as shown in Figure 3. Scaffolding is a dynamic intervention finely tuned to the learner's ongoing progress. The support given by the teacher during scaffolding strongly depends on the patterns of teacher-student interactions (Vygotsky and Cole, 1978; van de Pol et al., 2010). Therefore, the scaffolding abilities of tutors become the key criteria of effective tutoring systems. There are seven dimensions of scaffolding strategies to facilitate teaching and learning, which are **(1)** Feeding back, **(2)** Hints, **(3)** Instructing, **(4)** Explaining, **(5)** Modeling, **(6)** Questioning, **(7)** Social-emotional Support (Gibbons, 2015).

On the other hand, the multifaceted functions of code-switching in teaching practices highlight its strategic use in pedagogical contexts. DiCamilla and Antón (2012) identified four major functions of code-switching in classroom discussions: 1) creating, discussing, and translating content, 2) negotiating grammatical, lexical, and stylistic choices, 3) planning, defining, and managing tasks, and 4) maintaining and developing interpersonal relationships. Building on this classification, Tigert et al. (2019) expanded the framework to include five major functions of code-switching in bilingual classrooms: **[A]** Negotiating content, **[B]** Clarifying language, **[C]** Checking for or confirming understanding, **[D]** Task management, **[E]** Building relationships. These findings reveal how multiple ways that teachers could employ code-switching strategies to scaffold learning.

Language learning is one of the typical applications of code-switching in classrooms. The inherent nature of students' limited proficiency in the target language requires extensive scaffolding through



- **Introducing new concept:** *to explain the meaning, use, and definitions of focal vocabulary words.*
- **Clarifying comprehension:** *to correct the mistakes and clarify the content of the picture.*
- **Correcting grammar:** *to explain the correct usage of grammar.*
- **Checking for understanding:** *to confirm the student's understanding of contents or vocabulary.*
- **Task management:** *to discuss or explain directions regarding what to do next.*
- **Emotional support:** *to show care, affection, emotion, to build trust and relationships.*

Figure 3: Conceptualizing pedagogical code-switching. Each dimension can be used as instructions for dialogic tutors, as well as the evaluation rubrics.

the use of **L1** for explaining, modeling, and providing instructions. Based on scaffolding strategies and Tigert et al. (2019)'s framework, we analyze dialogues between teachers and students in bilingual classrooms from previous work DiCamilla and Antón (2012); Tigert et al. (2019); Vaish and Subhan (2015). We build a pedagogical code-switching framework with six dimensions (as shown in Figure 3): 1) *Introducing New Concept (Vocabulary)*, 2) *Clarifying Comprehension*, 3) *Correcting Grammar*, 4) *Checking for Understanding*, 5) *Task Management*, 6) *Emotional Support*. For example, we integrate scaffolding strategies **(2)** Hints and **(4)** Explaining to code-switching function **[B]** Clarifying language. We rename it as Introducing New Concept and define it as "to explain the meaning, use, and definitions of focal vocabulary words."

4 Improving Tutoring Systems with Pedagogical Code-Switching

LLMs pave a new way to build dialogue tutors with less data supervision and higher customization (Macina et al., 2023b). They can be shaped along desired dimensions to mimic human conversational characteristics such as tone and personality traits to deliver better user experience (Safdari et al., 2023; Shao et al., 2023; Liu et al., 2024b). We thus leverage LLMs as a tutoring agent for language learning, and adopt pedagogical code-switching based on our proposed framework. Previous work shows that in task-specific settings, the general and coarse instruction may lead to inconsistent generation (Liu et al., 2024a). Therefore, we set up fine-grained instructions to improve and evaluate the LLM-based tutoring systems.

Bi-lingual Setting: The conversational tutoring system is designed for **L2** learning (**L1** is English), and particularly focuses on the image description task. In each tutoring session, the student is presented with a picture and asked to describe the incidents in **L2** (e.g., Chinese, Spanish). From a syntactic perspective, languages can be classified into three structural patterns: 1) Subject-Verb-Object (e.g., English, Chinese); 2) Verb-Subject-Object (e.g., Arabic, Irish); 3) Subject-Object-Verb (e.g., Japanese, Korean). Given that code-switching is a structured and syntax-related phenomenon in linguistics, in our study, we select two representative translanguaging combinations (**Chinese-English**, **Korean-English**) in our language learning case study, to represent the applicability and generalizability of code-switching in different language contexts.

Tutoring Setting: Teaching and improving students' L2 acquisition through image description is a dynamic and engaging approach. In the image description tasks, the learning objective is using target sentences to describe the given image that includes a particular place or setting, people or animals, items and actions, etc. The teacher uses scaffolding and code-switching strategies to guide students step by step until they can independently complete the image description task. We build a multi-agent communication environment following previous work (Zhang et al., 2023a; Wu et al., 2023; Liu et al., 2024b).

Teacher Role Instruction: The teacher guides students to describe the items, emotions of people, and incidents depicted in the images, following teacher-centered IRF (Wells, 1999) sequences to promote interaction. In this process, the teacher applies scaffolding strategies, such as questioning, reformulation, and elaboration to assist learners in knowledge construction and expression (Gibbons,

Dimension	Definition	Chinese-English Code-switching Example
Introducing New Concept	use code-switching to explain the meaning, use, and definitions of focal vocabulary words.	“hungry”的中文是“饿”。我们可以说“小男孩有点饿了”, it means “The boy is hungry” (The Chinese word for “hungry” is “e”. We can say “The little boy is a bit hungry”, it means “The boy is hungry”)
Clarifying Comprehension	use code-switching to correct the mistakes and clarify the content of the picture	再仔细看看图片，they are not just standing. Think about how we stand in line when we wait for food at school. We stand one behind the other, right? (Take a closer look at the picture, they are not just standing. Think about how we stand in line when we wait for food at school. We stand one behind the other, right?)
Correcting Grammar	use code-switching to explain the correct usage of grammar	不过在一个句子的开头，我们通常会加上主语，we say "小朋友在玩游戏", instead of "玩游戏”。 (However, at the beginning of a sentence, we usually add the subject, we say "The children are playing games", instead of "playing games".)
Checking for Understanding	use code-switching to confirm students’ understanding of contents or vocabulary	用中文我们怎么说 "reading books"? (How do we say "reading books" in Chinese?)
Task Management	use code-switching to discuss or explain directions regarding what to do next	OK, now look at the right part of the picture. 你能看到有两个男孩吗? (OK, now look at the right part of the picture. Can you see two boys?)
Emotional Support	use code-switching to show care, affection, and emotion, to build trust and relationships	Great! 你做得真棒！ (Great! You did a great job!)

Table 1: Definition and examples of pedagogical code-switching in bilingual language learning (L1: English, L2: Chinese). See Table 4 in the Appendix for Korean-English code-switching examples.

C1: Teacher Role Instruction

[Role & Task Definition] *You are a primary school language teacher. You teach the student to describe the picture. The student’s first language (L1) is English, and target language (L2) is Chinese/Korean.*
[Pedagogical Instruction] *You apply scaffolding and code-switching of L1 and L2 during tutoring.*
– Detailed pedagogical code-switch description. –
[Behavior Constraint] *Ask the student only one question at a time. Always wait for input before proceeding to the next step. Correct the student’s answers if they are inaccurate.*

C2: Student Role Instruction

[Role & Task Definition] *You are a primary school student. You are taking a language learning class, and describing the given pictures.*
[Language Capability] *Your first language is English, and your Chinese/Korean proficiency is limited. You make some grammar errors in your responses to the teacher.*

2015). Due to the students’ limited proficiency in **L2**, we instruct the teacher agent (as in Codebox C1) to apply six dimensions of pedagogical code-switching during the tutoring process, and add reference examples of each dimension (as in Table 1).

Student Role Instruction: We follow the learning process via human-machine interaction, where the tutoring system (i.e., teacher) leads the conversation, and we feed responses from a student simulator instead of the human participants. To trigger scaffolding and code-switching strategies, we set the student role (as in Codebox C2) to include both **L1** and **L2**, while **L2** (i.e., Chinese, Korean) proficiency is low, and the student occasionally makes grammar mistakes. With the support and guidance from the teacher agent, the student is able to complete the given task, and improve **L2** skills including vocabulary, organization, and fluency (de Oliveira et al., 2023).

5 Experimental Setup

In our preliminary study, open LLMs such as Mistral-7B and Llama-3-8B cannot follow the pedagogical instructions well, they tend to produce monolingual responses and fail to generate coherent tutoring dialogues. Therefore, our experiments are conducted on two state-of-the-art representative models: Gemini (Team et al., 2023) and GPT-4-turbo (Achiam et al., 2023).[1] Following previous work (Touvron et al., 2023), we adjust instructions to the chat template of each model. For tutoring dialogue generation, both teacher and student roles use the same model, and we feed the concatenated utterances for dialogue generation. We randomly sample 50 open-sourced cartoon images and use one sentence of image description as a learning target to generate 400 tutoring dialogues. The total utterance number is 9K.

---

[1] The experimented versions are Gemini-Pro-1.5 and GPT-4-turbo-0125-Preview.

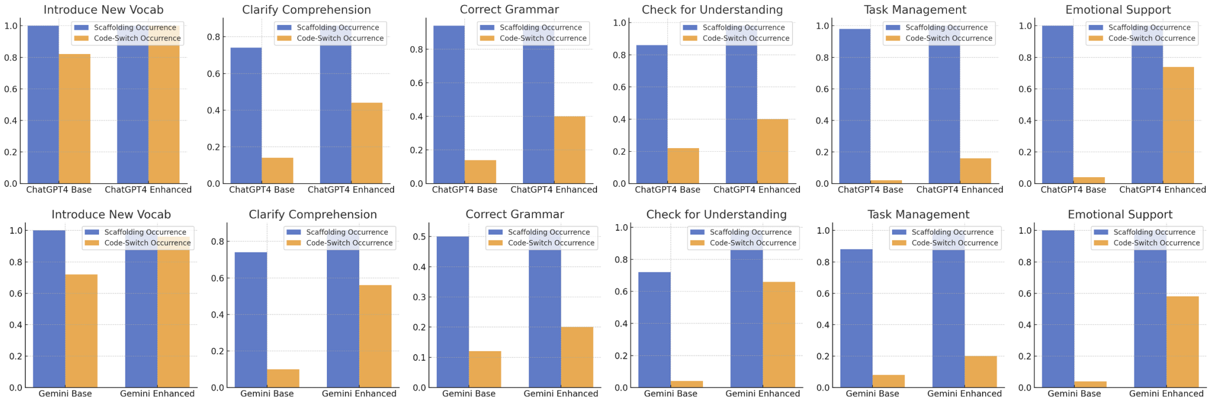


Figure 4: Quantitative results on Chinese-English pedagogical code-switching via automated evaluation.

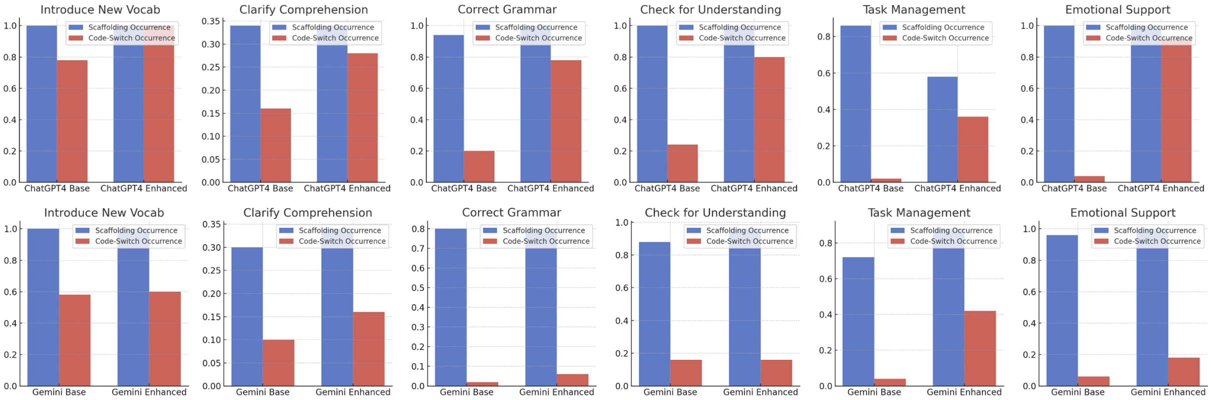


Figure 5: Quantitative results on Korean-English pedagogical code-switching via automated evaluation.

For quality assessment and data analysis, we conduct both LLM-based automated evaluation and manual rating by human experts.

Automated Evaluation: We adopt the LLM-as-a-judge (Saha et al., 2023) with GPT-4 to validate the effectiveness of pedagogical code-switching according to our proposed framework. Given a dialogue, the evaluator is to predict the occurrence of each dimension (e.g., introduce new vocabulary, clarify comprehension, etc.). Next, we assess whether code-switching is applied to these functions. In each dimension, one point will be added if the dialogue meets the criteria (as shown in Appendix Table 5). To build the reference label, two bilingual experts annotate the generated dialogues. Their teaching languages are Chinese and Korean. We provide each expert with an overview of the study, its objectives, and code-switching evaluation rubrics.

Qualitative Analysis: We invite two language teachers to complete a survey rating the pedagogical ability of the tutoring systems. The survey items are adapted from Tack and Piech (2022). We randomly select 15 dialogue segments from each system and in both Chinese-English and Korean-English. For each dialogue segment, we ask three questions: To what extent do you think the teacher 1) speaks like a human tutor, 2) applies code-switching effectively, and 3) scaffolds the student's learning? The annotators are asked to indicate their agreement with these three statements on a 5-point scale, ranging from 1 (strongly disagree) to 5 (strongly agree). The survey example can be found in Appendix Figure 7.

6 Results and Analysis

6.1 Code-switching Generation Evaluation

We first measure the model's performance in generating pedagogical code-switching. In Chinese-English tutoring dialogue evaluation, GPT-4-turbo ($p = 0.021$) and Gemini ($p = 0.007$) show significant improvement when applying pedagogical code-switching across six dimensions, compared to the systems with the base instruction. Not surprisingly, all tested systems show a high utilization of *Introducing new vocabulary*, which is often involved with simple word replacement within a

Model	Scaffolding Label		Pedagogical CSW		Scaffolding Label		Pedagogical CSW	
	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
LLaMA-3-Chat 8B	83.33	81.83	74.33	73.14	80.67	76.55	51.33	50.18
Mistral-7B-Instruct-v0.2	85.33	84.52	71.50	70.09	80.33	75.10	51.17	46.07
GPT-3.5-turbo-1106	89.00	85.46	76.33	76.55	82.50	74.59	62.67	62.63
GPT-4-turbo-1106	90.83	88.79	78.50	78.33	85.16	80.49	71.17	70.33

Table 2: Model comparison of automated evaluation with LLM-as-a-judge. Scaffolding labeling is only to predict the scaffolding types regardless of translanguaging. CSW is short for code-switching. Columns in pink denote the results of English (L1) and Chinese (L2). Columns in blue denote the results of English (L1) and Korean (L2).

sentence. However, LLMs with the base instruction cannot produce diverse code-switching in other pedagogical types (e.g., *Check for Understanding*, *Task Management*, *Emotional Support*).

Moreover, LLMs perform differently across different language mixing of code-switching. For instance, in Korean-English tutoring dialogue evaluation, only GPT-4-turbo (p = 0.022) shows significant improvement in applying code-switching strategies across six dimensions when compared to no pedagogical code-switching instructions; The Gemini does not show the same trend in improvement. Particularly, the performance on *Introducing new vocabulary*, *Correct Grammar*, and *Checking for Understanding* remains the same. We speculate that this is because LLMs' multilingual capability differs across languages, and the code-switching beyond simple word replacement (e.g., *Correct Grammar*, *Checking for Understanding*) relies more on cross-lingual knowledge.

6.2 LLM-as-a-judge Model Comparison

To validate the efficacy of automated evaluation and compare the performance across open and closed-source LLMs, we use manual annotation as a reference, and results are presented in terms of correlation with human judgments (in accuracy and F1 scores). We selected and tested a list of representative models (e.g., LLaMA-3, Mistral, GPT-3.5, and GPT-4). As shown in Table 2, all tested models can provide reasonable results on labeling Chinese-English dialogues, where GPT-4 performs slightly better in the code-switching labeling. However, they all achieve a lower performance on labeling Korean-English dialogues; the scores of LLaMA and Mistral are sub-optimal. We speculate that this is due to the models' inconsistent multilingual capability (Wang et al., 2023a).

6.3 Qualitative Analysis

We conduct qualitative analysis in each dimension of code-switching strategies on models' generation (see examples in Appendix Table 6 and Table 7):

Introducing New concept: The tutor alternates between **L1** and **L2** to explain the meanings and uses of target words. For example, when introducing "春节/설날" (Chinese New Year), it uses both Chinese and Korean to introduce the meaning of the word, "the beginning of the year according to the Chinese calendar," and English to explain the cultural significance of Chinese New Year, as "*It's a time for families to get together and have a big feast.*" By offering these complementary meanings across languages, students are able to construct a better understanding of the target vocabulary.

Clarifying Comprehension: When students make some misunderstanding, the tutor uses **L1** to provide hints and guidance to help them co-construct knowledge of the picture. In addition, the tutor encourages them to show their own understanding and modify answers in **L2**. For example, when students misinterpret the setting of the image as a birthday party when it actually represents a Chinese New Year celebration, the tutor gives positive feedback in **L2**, like "你的观察很有趣， 生日派对也会有很多人在一起庆祝。/너의 관찰이 재미있어요. 생일 파티에도 많은 사람들이 모여서 축하할 거예요". Next, the tutor provides some hints in **L1** to guide the students toward an accurate description, "*However, this is not a birthday party. Do you notice the red lanterns and Chinese characters on the wall in the picture?*".

Correcting Grammar: When students make grammar mistakes, the tutor follows the Subject-Verb-Object linguistic feature in Chinese/Subject-Object-Verb in Korean to explain the grammar usage. It also uses the modeling strategy to provide structured examples for the students to imitate. For instance, the tutor provides sentence structure to facilitate students to complete sentences by filling in the blanks, "We can say [Subject]이 [Action]을 위해 [Body Part]을 들고 있습니다. / [主语]将[身体部位]举起来[谓语+宾语] (e.g., *She is holding*

Translanguaging Type	GPT-4-turbo Mean	 SD	Gemini Pro Mean	 SD	*t-test* *t* value	 *p* value
Chinese-English	11.4	1.453	14.133	0.819	-8.976	< .001
Korean-English	11.6	2.283	10.767	1.006	1.829	0.073

Table 3: Comparison of GPT-4-turbo and Gemini on pedagogical ability.

up her hand to receive a gift.)”

Checking for Understanding: The tutor uses two ways to check or confirm students’ understanding. First, the “repeat-after-me” strategy is to ask students to repeat new vocabulary or sentences in **L2**. The verbal repetition process is beneficial to language acquisition. It is believed that learning occurs as a result of repeated exposure to a given stimulus (Watson, 2017). Another method is to ask students to translate vocabulary from **L1** into **L2**. For example, the tutor asks, “*Do you know ‘Dinosaur’ in Chinese / Korean?*”

Task Management: We observe that the tutor applies code-switching when giving instructions, explaining what students going to do next, or drawing attention to the objective of the task. This function serves to raise awareness of the focal words and learning tasks. For instance, “*Now look at the right part of the picture.* 你能看到有两个男孩吗? /두 소년을 봤어요?”

Emotional Support: The tutor gives a lot of positive affirmation and encouragement to learners, such as “잘 했어요! / 太棒了！ *You did a great job!*” Both GPT-4-turbo and Gemini perform very well in this function because LLMs are designed to understand and respond to human emotions effectively. No matter what the student’s answer is, these models primarily respond with emotional support first. While this design enhances user experience and encourages continued engagement, its consistently stable structure makes interactions feel more like with a machine rather than a human teacher.

6.4 Pedagogical Ability Evaluation Results

The pedagogical ability evaluation results (shown in Table 3) indicate that both GPT-4-turbo and Gemini with code-switching strategies demonstrate strong pedagogical effectiveness.

In Chinese-English tutoring, Gemini (*Mean* = 14.113, *SD* = 0.819) outperformed GPT-4-turbo (*Mean* = 11.400, *SD* = 1.453) by generating teacher-student dialogues that are very natural and fluent ($p < 0.001$). For example, in tutoring example 23, when the student observes a person in the picture wearing a white coat, the tutor provides hints, “非常好！你观察到图片中的男人穿着一件白大褂，手里拿着一些工具 *to check people’s teeth. Can you guess his job*?” This example shows that the tutor builds on students’ ideas and expands their knowledge. In contrast, the GPT-4-turbo tutor is more target-driven. The conversation focuses on teaching vocabulary and guiding students in completing learning tasks. When students become distracted and answer with irrelevant words, the tutor directly corrects them and asks them back to the tasks. For example, in tutoring example 30, when the student responds incorrectly and provides irrelevant information about the picture by saying “回家了 (They are back home)”, the tutor corrects them with, “*No, they are looking for shells,* ‘贝壳’.”

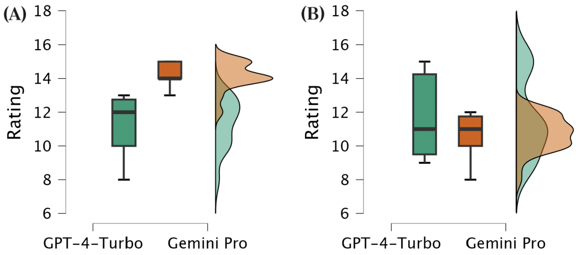


Figure 6: Descriptive statistics of human ratings on code-switching dialogue segments. (A) Chinese-English result. (B) Korean-English result.

In Korean-English tutoring, the performance of Gemini (*Mean* = 10.767, *SD* = 1.006) and GPT-4-turbo (*Mean* = 11.600, *SD* = 2.283) does not show significant difference ($p = 0.073$). GPT-4-turbo follows the instructions very well, especially when correcting grammar and providing explanations. For example, in tutoring example 43, when the student makes a grammatical error by confusing the order of verbs in a sentence, the tutor advises, “*Remember, in Korean, we often place the verb at the end of the sentence. You should say,* ‘소년이 청소기를 사용해 방을 청소하고 있습니다’.” However, its code-switching ability is inconsistent, with ratings varying from 5 to 15 (see Figure 6). This suggests that the multilingual ability requires further enhancement to improve its reliability. In contrast, Gemini’s performance is more consistent, although sometimes it mixes other languages (e.g., Russian, Japanese, Arabic) in utterances. For example, "아니요, these are not presents. Look closely at the box the woman is holding. '通常、誕生日にプレゼントを渡します。'(*Usually we give presents on birthdays*)."

7 Conclusion

In this work, we combined scaffolding strategies and translanguaging functions to propose a pedagogical code-switching framework. In a theory-inspired practice, we developed fine-grained instructional strategies tailored to multilingual learners and bilingual education needs, and leveraged LLMs as the tutoring agent and automated evaluator. Our experimental results revealed that state-of-the-art LLMs demonstrated reasonable code-switching and pedagogical ability in bilingual learning contexts. Moreover, we observed that English-centric LLMs show imbalanced performance in scaffolding, translanguaging, and pedagogical abilities across different languages, and improving cross-lingual consistency can be one of the future work. Aside from language learning, our proposed code-switching framework can also be extended to broader multilingual interactions.

Limitations

In our experimental settings, we set the dialogic interaction in primary school language learning context, which focus mainly on basic vocabulary, grammar, and sentence structure. The translanguaging usage with advanced words and in complex syntax may pose other challenges. However, the proposed code-switching framework can be adapted to different contexts upon further refinement.

In addition, we are aware that it remains an open problem to mitigate hallucinations and biases in large language models, which may cause communication issues in human-machine interaction and computer-assisted education. Of course, current models and laboratory experiments are always limited in this or similar ways. We do not foresee any unethical uses of our proposed methods or their underlying tools, but hope that it will contribute to reducing incorrect system outputs.

Ethics and Impact Statement

We acknowledge that all of the co-authors of this work are aware of the provided ACL Code of Ethics and honor the code of conduct. In our experiments, models are applied under proper license. All data used in this work are only for academic research purposes and should not be used outside of academic research contexts. Our proposed methodology in general does not create a direct societal consequence and is intended to be used to improve the performance, robustness, and safety of the intelligent tutoring systems.

Acknowledgments

This research is supported by the Agency for Science, Technology and Research (AI4EDU Programme), and the National Research Foundation, Singapore under its AISG Programme (AISG2-GC-2022-005). We thank Carolyn Lee and Geyu Lin for the research discussions. We also thank the anonymous reviewers for their precious feedback to help improve and extend this piece of work.

References

2021. Translanguaging in education. *Language Teaching*, 54(4):439–471.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Muhammad Farid Adilazuarda, Samuel Cahyawijaya, Genta Indra Winata, Pascale Fung, and Ayu Purwarianti. 2022. IndoRobusta: Towards robustness against diverse code-mixed Indonesian local languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 25–34, Online. Association for Computational Linguistics.

Shazia Afzal, Tejas Dhamecha, Nirmal Mukhi, Renuka Sindhgatta, Smit Marvaniya, Matthew Ventura, and Jessica Yarbro. 2019. Development and deployment of a large-scale dialog-based intelligent tutoring system. In *Proceedings of the NAACL 2019*, pages 114–121, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Ahn, Cecilia Jimenez, Yulia Tsvetkov, and Alan Black. 2020. What code-switching strategies are effective in dialogue systems? *Society for Computation in Linguistics*, 3(1).

Pablo Arnau-González, Miguel Arevalillo-Herráez, Romina Albornoz-De Luise, and David Arnau. 2023. A methodological approach to enable natural language interaction in an intelligent tutoring system. *Computer Speech & Language*, 81:101516.

Malba Barahona. 2020. The potential of translanguaging as a core teaching practice in an EFL context. *System*, 95:102368.

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! *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–23.

Jasone Cenoz and Durk Gorter. 2022. Pedagogical Translanguaging and Its Application to Language Classes. *RELC Journal*, 53(2):342–354.

Soumaya Chaffar and Claude Frasson. 2004. Inducing optimal emotional state for learning in intelligent tutoring systems. In *International Conference on Intelligent Tutoring Systems*, pages 45–54. Springer.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35.

Yunjae J Choi, Minha Lee, and Sangsu Lee. 2023. Toward a multilingual conversational agent: Challenges and expectations of code-mixing multilingual users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Juliane C De La Campa and Hossein Nassaji. 2009. The amount, purpose, and reasons for using l1 in l2 classrooms. *Foreign language annals*, 42(4):742–759.

Terri De La Cruz. 2019. *Codeswitching in the Classroom*, volume 15. Routledge.

Luciana C. de Oliveira, Loren Jones, and Sharon L. Smith. 2023. Interactional scaffolding in a first-grade classroom through the teaching–learning cycle. *International Journal of Bilingual Education and Bilingualism*, 26(3):270–288.

Dorottya Demszky and Heather Hill. 2023. The ncte transcripts: A dataset of elementary math classroom transcripts. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 528–538.

Frederick J DiCamilla and Marta Antón. 2012. Functions of l1 in the collaborative interaction of beginning and advanced second language learners. *International Journal of Applied Linguistics*, 22(2):160–188.

A Seza Doğruöz, Sunayana Sitaram, Barbara E Bullock, and Almeida Jacqueline Toribio. 2023. A survey of code-switching: Linguistic and social perspectives for language technologies. *arXiv preprint arXiv:2301.01967*.

Myroslava Dzikovska, Natalie Steinhauser, Elaine Farrow, Johanna Moore, and Gwendolyn Campbell. 2014. BEETLE II: Deep Natural Language Understanding and Automatic Feedback Generation for Intelligent Tutoring in Basic Electricity and Electronics. *International Journal of Artificial Intelligence in Education*, 24(3):284–332.

Ofelia García. 2009. *Chapter 8 Education, Multilingualism and Translanguaging in the 21st Century*, pages 140–158. Multilingual Matters, Bristol, Blue Ridge Summit.

Pauline Gibbons. 2015. *Scaffolding language, scaffolding learning*. Heinemann.

Arthur C Graesser, Xiangen Hu, and Robert Sottilare. 2018. Intelligent tutoring systems. In *International handbook of the learning sciences*, pages 246–255. Routledge.

Beate Grawemeyer, Manolis Mavrikis, Wayne Holmes, Sergio Gutierrez-Santos, Michael Wiedmann, and Nikol Rummel. 2016. Affecting off-task behaviour: how affect-aware feedback can improve student learning. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, LAK '16, page 104–113, New York, NY, USA. Association for Computing Machinery.

Foteini Grivokostopoulou, Isidoros Perikos, and Ioannis Hatzilygeroudis. 2017. An Educational System for Learning Search Algorithms and Automatically Assessing Student Performance. *International Journal of Artificial Intelligence in Education*, 27(1):207–240.

Jason M. Harley, François Bouchet, M. Sazzad Hussain, Roger Azevedo, and Rafael Calvo. 2015. A multi-componential analysis of emotions during complex learning with an intelligent multi-agent system. *Computers in Human Behavior*, 48:615–625.

Nancy H Hornberger and Holly Link. 2012. Translanguaging and transnational literacies in multilingual classrooms: A biliteracy lens. *International journal of bilingual education and bilingualism*, 15(3):261–278.

R. Jacobson. 1981. The implementation of a bilingual instructional model: The new concurrent approach. In R. V. Padilla, editor, *Ethnoperspectives in Bilingual Education Research, Vol. 3: Bilingual Education Technology*, pages 14–29. Eastern Michigan University, Ypsilanti, MI.

James A. Kulik and J. D. Fletcher. 2016. Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review. *Review of Educational Research*, 86(1):42–78.

Angel Lin. 2013. Classroom code-switching: three decades of research. *Applied Linguistics Review*, 4(1):195–218.

Angel Lin. 2016. Code-Switching in the Classroom: Research Paradigms and Approaches. In Kendall A. King, Yi-Ju Lai, and Stephen May, editors, *Research Methods in Language and Education*, May, pages 1–15. Springer International Publishing, Cham.

William Littlewood and Baohua Yu. 2011. First language and target language in the foreign language classroom. *Language teaching*, 44(1):64–77.

Zhengyuan Liu, Shikang Ni, Ai Ti Aw, and Nancy F. Chen. 2022. Singlish message paraphrasing: A joint task of creole translation and text normalization. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3924–3936, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Zhengyuan Liu, Stella Xin Yin, Carolyn Lee, and Nancy F Chen. 2024a. Scaffolding language learning via multi-modal tutoring systems with pedagogical instructions. *arXiv preprint arXiv:2404.03429*.

Zhengyuan Liu, Stella Xin Yin, Geyu Lin, and Nancy F Chen. 2024b. Personality-aware student simulation for conversational intelligent tutoring systems. *arXiv preprint arXiv:2404.06762*.

Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023a. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of EMNLP 2023*, pages 5602–5621.

Jakub Macina, Nico Daheim, Lingzhi Wang, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023b. Opportunities and challenges in neural dialog tutoring. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2357–2372.

Christopher J MacLellan and Kenneth R Koedinger. 2022. Domain-general tutor authoring with apprentice learner models. *International Journal of Artificial Intelligence in Education*, 32(1):76–117.

Danièle Moore. 2002. Code-switching and Learning in the Classroom. *International Journal of Bilingual Education and Bilingualism*, 5(5):279–293.

Elghouch Nihad, En-naimi El Mokhtar, and Yassine Zaoui Seghroucheni. 2017. Analysing the outcome of a learning process conducted within the system als_corr(lp). *International Journal of Emerging Technologies in Learning (iJET)*, 12(03):pp. 43–56.

B Nye, Dillon Mee, and Mark G Core. 2023. Generative large language models for dialog-based tutoring: An early consideration of opportunities and concerns. In *AIED Workshops*.

Benjamin D Nye, Arthur C Graesser, and Xiangen Hu. 2014. Autotutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education*, 24:427–469.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

José Paladines and Jaime Ramirez. 2020. A systematic literature review of intelligent tutoring systems with dialogue in natural language. *IEEE Access*, 8:164246–164267.

Yi-chun Pan and Yi-ching Pan. 2010. The use of l1 in the foreign language classroom. *Colombian Applied Linguistics Journal*, 12(2):87–96.

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

Michael Rabbidge. 2019. *Translanguaging in EFL contexts: A call for change*. Routledge.

Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L Murnane, Emma Brunskill, and James A Landay. 2019. Quizbot: A dialogue-based adaptive learning system for factual knowledge. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–13.

Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. 2023. Branch-solve-merge improves large language model evaluation and generation. *arXiv preprint arXiv:2310.15123*.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.

Arabella J. Sinclair and Raquel Fernández. 2023. Alignment of code switching varies with proficiency in second language learning dialogue. *System*, 113(November 2022):102952.

Shashank Sonkar, Naiming Liu, Debshila Mallick, and Richard Baraniuk. 2023. Class: A design framework for building intelligent tutoring systems based on learning science principles. In *Findings of EMNLP 2023*, pages 1941–1961.

Katherine Stasaski, Kimberly Kao, and Marti A Hearst. 2020. Cima: A large open access dialogue dataset for tutoring. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–64.

Anaïs Tack and Chris Piech. 2022. The AI Teacher Test: Measuring the Pedagogical Ability of Blender and GPT-3 in Educational Dialogues. *Proceedings of the 15th International Conference on Educational Data Mining, EDM 2022*.

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

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

Johanna Tigert, James Groff, Melinda Martin-Beltrán, Megan Madigan Peercy, and Rebecca Silverman. 2019. Exploring the pedagogical potential of translanguaging in peer reading interactions. In *Codeswitching in the Classroom*, pages 64–87. Routledge.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Miles Turnbull and Katy Arnett. 2002. Teachers' uses of the target and first languages in second and foreign language classrooms. *Annual review of applied linguistics*, 22:204.

Viniti Vaish and Aidil Subhan. 2015. Translanguaging in a reading class. *International Journal of Multilingualism*, 12(3):338–357.

Janneke van de Pol, Monique Volman, and Jos Beishuizen. 2010. Scaffolding in Teacher–Student Interaction: A Decade of Research. *Educational Psychology Review*, 22(3):271–296.

Lev Semenovich Vygotsky and Michael Cole. 1978. *Mind in society: Development of higher psychological processes*. Harvard university press.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, Ai Ti Aw, and Nancy F Chen. 2023a. Seaeval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. *arXiv preprint arXiv:2309.04766*.

Rose E Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2023b. Step-by-step remediation of students' mathematical mistakes. *arXiv preprint arXiv:2310.10648*.

John B Watson. 2017. *Behaviorism*. Routledge.

Gordon Wells. 1999. Language and education: Reconceptualizing education as dialogue. *Annual Review of Applied Linguistics*, 19:135–155.

C Williams. 1994. An evaluation of teaching and learning methods in the context of secondary education. *Unpublished Doctoral Dissertation, University of Bangor*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

Zhihong Xu, Kausalai Wijekumar, Gilbert Ramirez, Xueyan Hu, and Robin Irey. 2019. The effectiveness of intelligent tutoring systems on K-12 students' reading comprehension: A meta-analysis. *British Journal of Educational Technology*, 50(6):3119–3137.

Jintian Zhang, Xin Xu, and Shumin Deng. 2023a. Exploring collaboration mechanisms for llm agents: A social psychology view. *arXiv preprint arXiv:2310.02124*.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Dimension	**Definition**	**Korean-English Code-switching Example**
Introduce New Vocabulary	use code-switching to explain the meaning, use, and definitions of focal vocabulary words.	There is a special celebration in Korea called “설날” which is kind of like New Year’s Eve. It’s a time for families to get together and have a big feast. (There is a special celebration in Korea called “Seollal” which is kind of like New Year’s Eve. It’s a time for families to get together and have a big feast.)
Clarify Comprehension	use code-switching to correct the mistakes and clarify the content of the picture	아니요, 생일은 아니에요. Do you see the red decorations and Chinese characters in the picture? (No, it’s not the birthday party. Do you see the red decorations and Chinese characters in the picture?)
Correct Grammar	use code-switching to explain the correct usage of grammar	거의 맞았어요, but let’s correct the grammar a bit. Remember, sentences in Korean usually follow the Subject-Object-Verb structure. For example, “가족이 저녁 식사를 합니다.” (You almost got it, but let’s correct the grammar a bit. Remember, sentences in Korean usually follow the Subject-Object-Verb structure. For example, "The family is having dinner.")
Check for Understanding	use code-switching to confirm students’ understanding of contents or vocabulary	너무 잘했어요. Can you try saying it one more time?“ (You did very well. Can you try saying it one more time?)
Task Management	use code-switching to discuss or explain directions regarding what to do next	오른쪽 부분을 좀 보세요. Did you see two boys? (Please look at the right part of the picture. Did you see two boys?)
Emotional Support	use code-switching to show care, affection, emotion, to build trust and relationships	잘 했어요! You did a great job! (Great! You did a great job!)

Table 4: Definition and examples of pedagogical code-switching in bilingual language learning (L1: English, L2: Korean). See Table 1 for Chinese-English code-switching examples.

I will give you tutoring dialogues for teaching students to learn Chinese via image description.
Please evaluate the code-switching of the Chinese teacher based on the following rubrics.
Add 1 point of each dimension only when the teacher properly uses English for code-switching.
Code-switching Evaluation Rubrics:
1. Introduce New Vocabulary:
- Definition: adopt code-switching to explain the meaning, use, and definitions of focal vocabulary words.
- Example: "hungry的中文是’饿’。我们可以说’小男孩有点饿了’。"
2. Clarify Comprehension:
- Definition: adopt code-switching to correct the mistakes and clarify the content of the picture.
- Example: "再看看图片，再仔细看看图片. Think about how we stand in line when we wait for food at school. We stand one behind the other, right?"
3. Correct Grammar:
- Definition: adopt code-switching to correct grammar or explain the usage of grammar.
- Example: "我们通常会加上主语，we say ’Children are playing games’, instead of ’playing games’."
4. Check for Understanding:
- Definition: adopt code-switching to confirm the student’s understanding of ideas or vocabulary.
- Example: "用中文我们怎么说’reading books’?"
5. Task Management:
- Definition: adopt code-switching to discuss or explain directions regarding what to do next.
- Example: "OK, now look at the right part of the picture. 你能看到有两个男孩吗?"
6. Emotional Support:
- Definition: adopt code-switching to show care, affection, emotion, to build trust and relationships.
- Example: "Great!你做得真棒！", "That’s ok,我们一起来看看."

Table 5: The instruction example of leveraging LLMs for automated pedagogical code-switching evaluation.

Generated Tutoring Conversation Example (GPT-4-0125-Preview):

Teacher: 小朋友，看这张图片，这是在什么地方呢?

Student: I see a big room with books.

Teacher: 很好！这个地方有很多书，我们中文叫它什么？Do you remember we talked about a place where we borrow books?

Student: Library? 书...店?

Teacher: 是的，书店里也有很多书，But this is a place where the books are for borrowing instead of for sale. 中文里我们说“图书馆”。你能试着用中文说吗?

Student: 图书… 馆?

Teacher: 非常好！图书馆。现在看看图片里的人都在做什么?

Student: They are quiet.

Teacher: 对，他们都很安静。记得“安静”这个词吗?

Student: (No response.)

Teacher: 没关系。安静就是没有声音，very quiet. Can you say “安静”?

Student: 安静。

Teacher: 太好了！现在我们试着用中文描述老师在图书馆里做的事情?

Student: 老师… quiet in 图书馆。

Teacher: 接近了，Let’s put it into a full sentence. 我们用“让”来表示老师让孩子们保持安静 The teacher asked the children to keep quiet。试试看，用完整的句子。

Student: 老师在图书馆让孩子们安静。

Teacher: 非常接近了！但记得我们要用“保持安静” keep quiet。这样句子会更准确。再试一次?

Student: 老师在图书馆里让孩子们保持安静。

Teacher: 完美！You did very well，我很高兴看到你的进步。我们继续学习，好吗?

Student: 好的，谢谢老师！

Generated Tutoring Conversation Example (Gemini Pro):

Teacher: 小朋友，看看这张图片，上面画了什么呀?

Student: I see a boy in a restaurant.

Teacher: 很好！你能用中文描述一下他吗?

Student: 他在餐厅。

Teacher: 非常好！他看起来怎么样?

Student: He is hungry.

Teacher: 还记得我们学过的“hungry”的中文吗?

Student: 我不知道。

Teacher: No worries！“hungry”的中文是“饿”。我们可以说“小男孩饿了”, it means "The boy is hungry"。Now, look closely. 他现在在做什么呢?

Student: He is waiting in line.

Teacher: “waiting in line”的中文怎么说?

Student: I don’t know.

Teacher: Think about how we stand in line when we wait for food at school. We stand one behind the other, right?

Student: Yes.

Teacher: In Chinese, we say "排队". Now, can you say "小男孩有点饿了，他在餐厅排队"?

Student: 小男孩有点饿了，他在餐厅排队。

Teacher: 太棒了！但是，你看图片中，小男孩站在队伍的前面，而不是后面。He is standing in front of the line instead of the behind. 这是什么意思呢?

Student: I don’t know.

Teacher: When we want to get something quickly, but we don’t wait our turn, we "cut in line". In Chinese, we say "插队". Can you use "插队" in your sentence?

Student: 小男孩太饿了，他在餐厅插队。

Teacher: 非常好！You said very clear。你觉得小男孩插队好么?

Student: No.

Teacher: 为什么呢?

Student: Because it’s not fair.

Teacher: 没错！插队是不好的行为。我们应该遵守规则，排队等候。Now, how would you describe this picture?

Student: 小男孩有点饿了，他在餐厅插队，但他不应该这样做。

Teacher: 你做得真棒！You did great job! 你不仅学会了用中文描述图片，而且还表达了自己的观点。继续努力，好吗?

Student: 好的，谢谢老师！

Table 6: Two generated dialogue examples with pedagogical code-switching (L1: English, L2: Chinese).

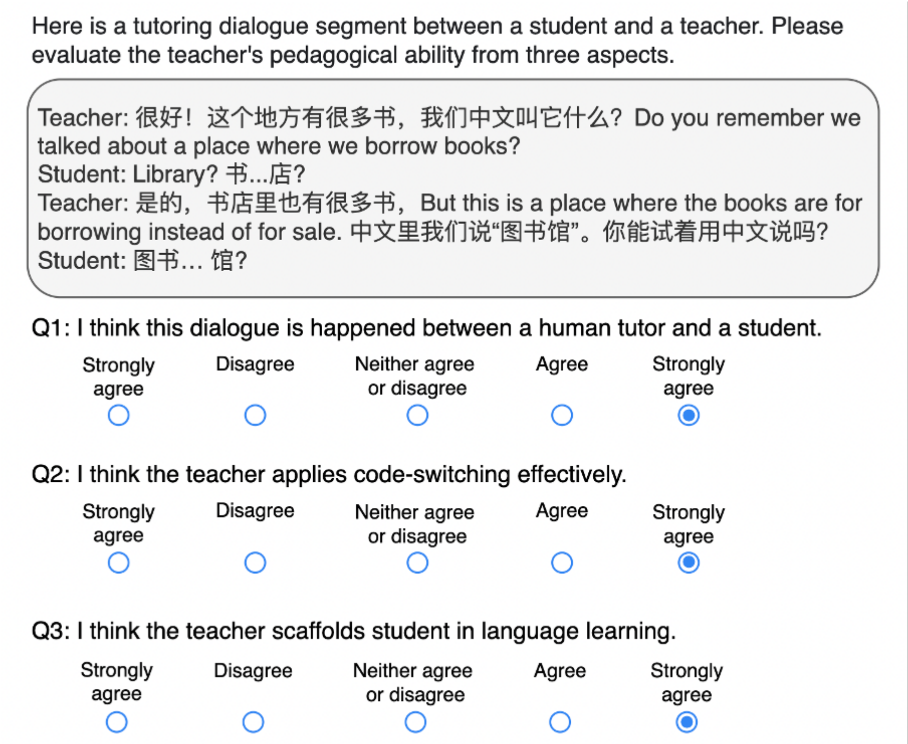
Generated Tutoring Conversation Example (GPT-4-0125-Preview):

Teacher: 안녕하세요! 오늘 우리는 이 사진을 한국어로 설명할 거예요. Take a look at this picture and see if you can tell me what's happening.
Student: (Looks at the picture) ...Umm... I'm not sure.
Teacher: 오른쪽 부분을 좀 보세요. Look at the right part of the picture. Do you see the person in the white coat?
Student: Yes, he looks like a doctor.
Teacher: 네, 정확해요! Yes, exactly! The man in the white coat is most likely a dentist. Do you know how to say "dentist" in Korean?
Student: Dentist... Uh... 치과의사?
Teacher: 네, 아주 좋아요! "치과의사" is exactly the word we use for dentist. It is the name for "tooth doctor." What do you think he might be doing?
Student: He's checking something... maybe?
Teacher: 네, 맞아요! Yes, that's right! He's examining something. Can you guess what he might be examining?
Student: (Hesitates) Uh... hair?
Teacher: 아니요, hair 아니에요. Look at the boy's mouth. What do you see there?
Student: His mouth is open... and there's a tool in it.
Teacher: Exactly! The dentist is using a tool to look at the boy's teeth. What Korean word do we use for "teeth?"
Student: (Stays silent)
Teacher: We use "치아" for teeth. It's similar to the English word "teeth" but pronounced a little differently.
Teacher: So, if we put it all together, what sentence can we make to describe what's happening in the picture?
Student: (Looks at the picture again) ... I don't know.
Teacher: No problem! Let's try building the sentence together. The dentist is the 치과의사 and he's looking at the boy's 치아. Where are they now?
Student: At the clinic, I guess?
Teacher: 네, 좋아요! Yes, good! We can say "클리닉" for clinic. So how can we put it all together in a sentence?
Student: The dentist... is looking at... the boy's teeth... at the clinic?
Teacher: Almost there! Can you try making it a complete sentence in Korean? For example, we can say "가다 (gada)" which means "to go" to make a sentence like "학교에 가요" which means "I go to school."
Student: 치과의사가... 클리닉에서... 치아를... I'm not sure how to say "looking at" in Korean.
Teacher: That's okay! There are a few ways to say "looking at." One way is "검사합니다" which means "to examine." So, how could we complete the sentence.
Student: 치과의사가 클리닉에서 어린이의 치아를 검사합니다.
Teacher: 너무 잘했어요! Very good! You described the picture perfectly! Keep practicing like this and you'll be a Korean description pro in no time!

Generated Tutoring Conversation Example (Gemini Pro):

Teacher: 오늘 우리는 이 사진을 한국어로 설명할 거예요. 사진을 먼저 보세요. What do you see in the picture?
Student: Uh, I see a... a big animal and a car?
Teacher: 맞아요, very good! 그 큰 동물, what is it?
Student: Oh, is it... giraffe?
Teacher: what is 'giraffe' in Korean?
Student: (No response.)
Teacher: It's OK, we call it '기린'! And how about that machine near '기린'? Do you remember that word?
Student: I know it! It is 'Car'!
Teacher: Perfect! How to say it in Korean?
Student: Is it '자동차'?
Teacher: 맞아요! Now, let's try to make a sentence by using '기린' and '자동차'. Can you try saying it?
Student: 기린이... 해변에서... 자동차를... umm, 운전... 습니다.
Teacher: 거의 맞았어요, but let's correct the grammar a bit. You were very close. We say "운전하고 있습니다." It means "is driving." Can you try it again?
Student: 기린이 해변에서 자동차를 운전하고 있습니다.
Teacher: 완벽해요! Very good! Do you want to describe any other details in the picture? Maybe what the weather looks like? 날씨가 어떻게 보이나요?
Student: It looks... sunny? No clouds?
Teacher: 정확해요, 맑은 날씨네요. Sunny day! 이제 '맑은 날씨'라고 말해 볼까요?
Student: 맑은... 날씨...
Teacher: 잘 했어요! You're doing really well. It's important to keep practicing like this. Remember, the more you speak, the better you will get.
Student: I will try, thank you!

Table 7: Two generated dialogue examples with pedagogical code-switching (L1: English, L2: Korean).



Here is a tutoring dialogue segment between a student and a teacher. Please evaluate the teacher's pedagogical ability from three aspects.

Teacher: 很好！这个地方有很多书，我们中文叫它什么？Do you remember we talked about a place where we borrow books?
Student: Library? 书...店?
Teacher: 是的，书店里也有很多书，But this is a place where the books are for borrowing instead of for sale. 中文里我们说“图书馆”。你能试着用中文说吗?
Student: 图书… 馆?

Q1: I think this dialogue is happened between a human tutor and a student.

Strongly agree	Disagree	Neither agree or disagree	Agree	Strongly agree
○	○	○	○	●

Q2: I think the teacher applies code-switching effectively.

Strongly agree	Disagree	Neither agree or disagree	Agree	Strongly agree
○	○	○	○	●

Q3: I think the teacher scaffolds student in language learning.

Strongly agree	Disagree	Neither agree or disagree	Agree	Strongly agree
○	○	○	○	●

Figure 7: The survey form used for manual pedagogical ability evaluation.