Can Code-Switched Texts Activate a *Knowledge Switch* in LLMs? A Case Study on English-Korean Code-Switching

Anonymous ACL submission

Abstract

Code-switching (CS), a phenomenon where multilingual speakers alternate between languages in a discourse, can convey subtle cultural and linguistic nuances that can be otherwise lost in translation. Recent state-of-the-art multilingual large language models (LLMs) demonstrate excellent multilingual abilities in various aspects including understanding CS, but the power of CS in eliciting language-specific knowledge is yet to be discovered. Therefore, we investigate the effectiveness of code-switching on a wide range of multilingual LLMs in terms of *knowledge activation*, or the act of identifying and leveraging knowledge for reasoning. To facilitate the research, we first present EnKoQA, a synthetic English-Korean CS question-answering dataset. We provide a comprehensive analysis on a variety of multilingual LLMs by subdividing activation process into *knowledge identification* and *knowledge leveraging*. Our results demonstrate that compared to English text, CS can faithfully activate knowledge inside LLMs especially on language-specific domains, suggesting the potential of code-switching on low-resource language tasks.

1 Introduction

Code-switching (CS), or the practice of alternating between two or more languages or language varieties within an utterance, is a common phenomenon in multilingual societies. There are multiple motivations for code-switching, to compensate for lack of language proficiency, to emphasize certain emotions or points, or for group identity (Heredia and Altarriba, 2001; Doğruöz et al., 2021).

In particular, code-switching is an effective tool to embedded cultural meanings for bilinguals. Expressing certain concepts in original language can convey subtle cultural and linguistic nuances that can be lost in translation, and knowledge related to certain language are likely to be more memorized in its own language than in foreign languages. As

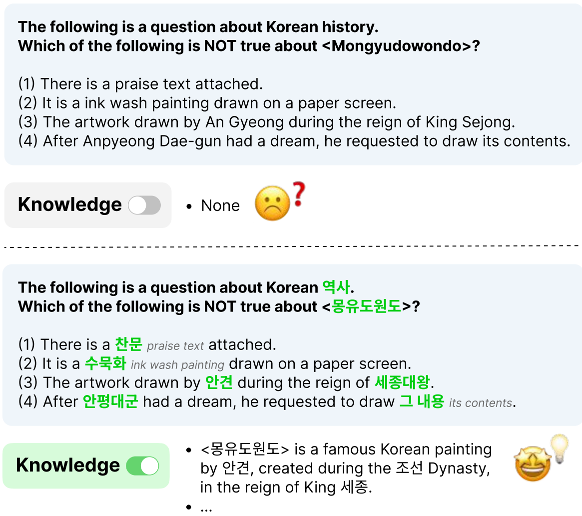

Figure 1: A motivating example of knowledge identification between languages. Compared to a question in English (*top*), a bilingual speaker can "activate" more relevant knowledge with a question in CS (*bottom*).

shown in Figure 1, when a human English-Korean bilingual is given a question that is closely related to Korean culture, a question in English and Korean code-switching is more capable of recalling knowledge about "몽유도원도"[1], because the concept is more familiar in Korean than in English.

Unlike human bilinguals, NLP tasks targeting low-resource languages often rely on machine translation to convert task data from a high-resource language (e.g., English) into the target language. However, crucial semantic nuances may be lost in translation, and machine translation errors are inevitable. To mitigate such risks, we attempt to leverage code-switching as a strategy to minimize nuance loss and reduce translation errors.

There have been continuous, if not abundant, researches on code-switching in the field of computational linguistics (Aguilar et al., 2020; Rizvi et al.,

[1]A landscape painting by An Gyeon in the early Joseon Dynasty requested by Prince Anpyeong, after his dream about Shangri-la. The painting is drawn on silk with ink.

2021). Recently, after the emergence of LLMs with impressive multilingual abilities, a line of work have discovered LLMs' abilities in CS (Huzaifah et al., 2024; Yong et al., 2023; Zhang et al., 2023a). However, the focus of such works are only limited to understanding and generating CS of LLMs, while the effectiveness of CS in tasks that involve low-resource language has not yet been explored.

In light of this, we investigate whether code-switching can effectively activate language-specific knowledge in LLMs. By *knowledge activation*, we refer to the overall process of identifying what knowledge is required, and applying knowledge to answer the question. Therefore, we ask ourselves the following research question: **Can code-switched texts activate language-specific knowledge, or turn on a "knowledge switch" in LLMs?**

To answer the question, we subdivide knowledge activation process into two tasks: (1) In *Knowledge Identification* task, we investigate if querying LLMs in CS and English yield different knowledge from its encoded memory. Specifically, we evaluate the quality of knowledge from different linguistic settings in terms of faithfulness and helpfulness. (2) In *Knowledge Leveraging* task, we observe if LLMs can faithfully ground on knowledge in different settings. We evaluate LLMs' accuracy on question-answering (QA) when given with knowledge.

Meanwhile, a crucial challenge when it comes to code-switching is the data scarcity. There is a limited number of CS datasets, let alone culture-focused data (Doğruöz et al., 2021). Since CS often happens in conversations, data are not easily available and the quality is not ensured. To address the shortage of data, efforts have been made to synthetically generate code-switching corpus based on linguistic theories (Pratapa et al., 2018; Rizvi et al., 2021; Salaam et al., 2022). However, these works rely on syntactic parsers and part-of-speech taggers that support limited languages, and the quality of text are highly dependent on the performances of those tools. Therefore, we first construct EnKoQA, a synthetic English-Korean code-switching dataset to explore the potential of CS in low-resource language task. Following Matrix Language Frame Model (Myers-Scotton, 1997), we synthesize Korean QA datasets (Kim et al., 2024b; Son et al., 2024) that encompass various aspects of Korea into English-Korean code-switched questions.

We conduct experiments with EnKoQA and provide extensive analysis on a wide range of multilingual LLMs. The experimental results reveal that CS is able to faithfully activate language-specific knowledge that are encoded in multilingual LLMs compared to high-resource language and target language translation; this tendency was more prominent on domains that specifically requires knowledge in target language and culture.

The contributions of our work are as follows:

- To the best of our knowledge, this work is the first to analyze the effectiveness of code-switching on knowledge activation to LLMs by introducing two tasks.
- We propose a qualified English-Korean code-switching QA dataset that is synthesized upon two Korean-centric datasets, and conduct extensive experiments on various families of multilingual LLMs.
- Experimental results on extensive LLMs indicate that code-switching has advantages in knowledge activation especially on language-specific domains, suggesting the potential of code-switching text as a tool for conveying cultural nuances in target language tasks.

2 Preliminaries & Related Work

In this section, we provide preliminary knowledge about code-switching, and explore relevant studies from conventional and computational linguistics.

2.1 Code-Switching Theories

Many linguistic theories attempt to explain the grammatical construction of code-switched text, such as Equivalence Constraint theory and Free Morpheme Constraint theory proposed by Poplack (1980). Equivalence Constraint (EC) theory suggests that code-switching occurs at points in a sentence where the structures of both languages are grammatically compatible. Free Morpheme Constraint (FMC) theory suggests that code-switching cannot occur between a bound morpheme and a lexical base. (*e.g.*, "He is look-ando for a book." is a wrong code-switch.)

However, these theories have limitations in that the theory can only be applied to two language with similar or equivalent syntactic structures. EC and FMC theories are not applicable to English-Korean code-switching text, due to the different sentence structure of Korean and English (Park and Yun, 2021). In this regard, we adopt Matrix Language Frame Model to construct our code-switching dataset.

2.2 Matrix Language Frame Model

Matrix Language Frame (MLF) model is a code-switching theory proposed by Myers-Scotton (1997). MLF model posits that in any instance of code-switching, one language provides the morpho-syntactic framework of the sentence. This is known as the *matrix language*. The other language, called the *embedded language*, contributes to additional content, usually in the form of words or phrases, but follows the grammatical rules set by the matrix language. In other words, matrix language dominates the sentence structure, while the embedded language is integrated within that structure. Content morphemes can be in both languages, but functional morphemes come from matrix language. Taking Figure 1 as an example, "그 내용" which translates to "its contents" can be embedded into English sentence, but functional morpheme such as "to" cannot.

2.3 Code-Switching for Language Models

Previous works introduce benchmarks for evaluating code-switching ability of multilingual language models across multiple tasks (Aguilar et al., 2020; Khanuja et al., 2020). More recent works focus on the capability of LLMs in code-switching. Zhang et al. (2023a) discover performance of multilingual LLMs in various code-switching tasks, including sentiment analysis and language identification. Yong et al. (2023) explore prompting multilingual LLMs to generate code-mixed data. Shankar et al. (2024) introduce a prompting technique called in-context mixing for effective in-context learning in LLMs. Although these benchmarks encompass a variety of tasks, the analysis of LLMs' code-switching capabilities in terms of knowledge retrieval and utilization has not yet been investigated.

2.4 Code-Switched Data Synthesis

Data synthesis for code-switching has been approached in various ways. Several studies utilize parsers and neural models to synthesize code-switched text based on EC theory (Pratapa et al., 2018; Rizvi et al., 2021). Similarly, Salaam et al. (2022) extract phrases from source language and reintegrate them into target language. In recent efforts to address data scarcity in low-resource settings, LLMs have been employed to generate synthetic data (Li et al., 2023). However, using LLMs specifically for synthesizing code-switched data remains unexplored.

3 ENKOQA: English-Korean Code-Switching QA Testset

To compare the effectiveness of code-switching with dominant language and translation in target language when performing language-specific tasks, we introduce ENKOQA, a synthetic English-Korean code-switching dataset that is designed based on MLF model. In this section, we first discuss the details of data construction (§ 3.1), and evaluate performances of LLMs on the dataset (§ 3.2, 3.3).

3.1 Dataset Construction

Data Sources. We leverage two multiple-choice Korean-centric question-answering datasets that encompass various aspects of Korean language and culture. CLIcK (Kim et al., 2024b) consists of 1,995 multiple-choice QA pairs, classified in two main categories (Culture, Language) and 11 sub-categories. In this work, we only utilize data of eight sub-categories from Korean Culture category as our work aims to evaluate the effect of CS on activating Korean-specific knowledge. HAE-RAE (Son et al., 2024) is a Korean benchmark dataset originally crafted to capture cultural and contextual nuances inherent to the Korean language. We use 1,027 multiple-choice QA pairs regarding Korean culture. Both datasets are sourced from official Korean exams, textbooks, and text on the internet. We combine two datasets and merge common categories (*i.e.*, Society, Geography, and Law), resulting in 2,372 QA pairs in nine categories: Popular, Economy, Politics, Tradition, General Knowledge, Society, Geography, History, and Law. More details of original datasets are provided in Appendix A.1.

Automatic Translation. As most LLMs are trained on English-dominant corpora, we regard the multilingual LLM as a bilingual whose matrix language is English but also fairly competent in Korean. To generate code-switched text that follows the MLF model, we need parallel data in Korean and English to extract semantically important words or phrases from Korean text and embed into English text. We first automatically translate all Korean query-choices pairs into English using `gpt-3.5-turbo`. The model is instructed to *translate the following {query} and {choices} to English* with an one-shot demonstration in a desired output format. In this paper, we henceforth refer to query-choices pair as *question*.

Model		Economy	General	Geography	History	Law	Politics	Popular	Society	Tradition	Total
GPT-4o	CS	**91.53**	78.41	**69.04**	74.79	**55.86**	90.48	95.12	**63.7**	85.14	**78.23**
	EN	89.83	75.00	66.19	61.97	52.64	84.52	95.12	60.4	74.32	73.33
	KO	83.05	61.36	56.23	56.62	39.54	80.95	85.37	50.83	67.12	55.31
GPT-3.5	CS	**71.19**	47.73	44.48	32.91	35.40	**70.24**	**80.49**	49.17	57.21	**54.31**
	EN	**71.19**	**48.86**	**45.55**	**36.32**	**36.55**	66.67	63.41	**52.64**	**62.61**	53.76
	KO	54.24	34.66	34.88	27.78	28.28	52.38	56.10	38.94	54.95	36.64
Claude 3.5	CS	93.22	**72.16**	72.95	**73.08**	62.53	**86.90**	95.12	67.66	**84.23**	**78.65**
	EN	89.83	71.59	67.97	61.54	55.63	85.71	92.68	63.20	75.23	73.71
	KO	72.88	42.05	52.67	44.23	40.00	73.81	75.61	53.63	60.36	50.51
Solar	CS	**83.05**	**55.11**	**54.09**	**63.46**	**42.76**	**80.95**	**85.37**	**54.29**	**75.23**	**66.03**
	EN	74.58	46.02	49.47	39.53	**42.76**	77.38	65.85	51.16	62.61	56.60
	KO	72.88	35.80	**54.09**	37.39	38.62	76.19	75.61	48.68	58.11	47.22
Llama3 70B	CS	79.66	51.70	**50.53**	**49.36**	44.14	**80.95**	**75.61**	**57.43**	65.77	**61.68**
	EN	**83.05**	**57.39**	**50.53**	45.94	**45.75**	73.81	73.17	53.30	**66.67**	61.07
	KO	74.58	44.32	47.69	37.39	34.25	69.05	65.85	46.04	55.41	44.98
Llama3 8B	CS	**69.49**	**40.34**	36.30	35.68	**35.63**	**75.00**	**73.17**	45.05	**54.05**	**51.63**
	EN	64.41	39.77	37.72	**37.39**	32.64	67.86	63.41	**45.21**	53.6	49.11
	KO	71.19	29.55	**38.43**	32.69	28.28	67.86	60.98	40.26	45.50	38.15
Gemma2 27B	CS	79.66	46.02	**48.75**	**41.03**	**45.29**	**77.38**	**78.05**	54.79	**65.32**	59.59
	EN	**84.75**	**53.41**	48.40	40.6	41.84	72.62	**78.05**	**54.95**	63.96	**59.84**
	KO	76.27	40.91	46.62	34.40	34.02	75.00	73.17	48.51	56.76	45.11
Gemma2 9B	CS	**79.66**	42.05	44.13	**40.17**	41.15	**73.81**	**80.49**	**53.30**	**65.77**	**57.84**
	EN	76.27	**46.02**	**49.47**	38.46	**42.30**	69.05	73.17	52.15	63.51	56.71
	KO	76.27	38.64	46.26	32.48	31.49	66.67	68.29	44.55	54.50	42.45

Table 1: QA performances of multilingual LLMs on CS, English, and Korean settings. **Bold** indicates higher score between CS and English on each model. **Green** indicates the highest score from each domain.

Generating Candidates in Different Levels. Now that we obtain parallel data in both languages, the next step is to embed Korean content morphemes into English sentence. However, as code-switching mostly happens spontaneously, there does not exist a certain formula for mixing two languages. Moreover, replacing every content word with its Korean equivalent may seem rather artificial. To address this, we simulated a natural code-switching by creating various versions of code-switched texts at different ratios (30%, 50%, 70%, and 90%), then manually selecting a version that represents the best quality and most naturalness. Specifically, given a question in both languages and a specified proportion, `gpt-3.5-turbo` identifies content words from the Korean question and integrates them into the English question according to the specified proportion. To collect contexts of various semantic importance, we employ two prompts that define "content word" differently; one defines content words as nouns or noun phrases, while the other identifies them as semantically important elements within the context. Through this process, we collect total eight code-switched candidates per question. Lastly, a single candidate that faithfully follows MLF and appropriately code-switches is selected. Comprehensive details about dataset construction are provided in Appendix A.

3.2 Experimental Settings

Models. We conduct extensive analysis on two groups of state-of-the-art multilingual LLMs: (1) Proprietary LLMs that are available via APIs, such as GPT-3.5, GPT-4o (OpenAI, 2023), and Claude 3.5 Sonnet (Anthropic, 2024). (2) Open-source LLMs such as Solar (10.7B, Kim et al., 2024a), Llama3 (8B, 70B, Dubey et al., 2024), and Gemma2 (9B, 27B, Gemma Team, 2024). More details about the models are in Appendix B.1.

Implementation Details. To compare performances of LLMs in dominant language, translation in target language, synthesized mixture of dominant and target language, we evaluate on English, translated Korean, and CS questions. For Korean translation baseline, we back-translate English translation text to Korean using Google Translate API. We simply ask the model to *read the following {question} and choose the most appropriate answer*. Full-length prompts are provided in Table 7.

3.3 Results

Overall. As shown in Table 1, the performance on CS significantly outperforms English and Korean across many LLMs for all domains. The gap is more prominent in models with advanced multilingual abilities, such as GPT-4o, Claude 3.5, and Solar.

CS questions excel at language-specific domains. It is also worth noting that the gap between performance of CS and English is substantially large on language-sensitive domains such as History and Tradition. Even for Llama 3 and Gemma 2 models which do not perform well on CS questions, show higher scores on CS for such domains. On the other hand, domains that are relatively general (*e.g.*, Society, General), and domains that require expert-level knowledge (*e.g.*, Politics, Law), show little or no increase. This finding indicates that asking LLMs in CS is much more effective when it comes to activating Korean-centric knowledge.

CS outperforms target language translation. We additionally compare code-switching with translated Korean to observe whether CS has advantages in minimizing translation errors. Overall, Korean translation baseline shows lowest performance among three baselines. This suggests that while translating task in target language yields undesirable results, CS can faithfully encapsulate meanings and linguistic cues that may be lost in translation, highlighting the potential of leveraging CS for performing non-dominant language tasks.

Ratios do not affect performance. To ensure that the ratio of code-switching does not influence models' performances and our dataset is constructed under fair process, we calculate Code-Mixing Index (CMI) scores (Srivastava and Singh, 2021) and report corresponding accuracy in Tradition and History domains. As shown in Table 4, we can see that accuracy scores are quite evenly distributed across all ratios, suggesting that there is no distinct tendency between CMI and accuracy.

4 Can Code-Switched Questions Activate a "Knowledge Switch" in LLMs?

From Section 3.3, we observe that most LLMs are able to answer correctly to questions in CS than in other baselines. To further investigate on the effectiveness of CS in activating language-specific knowledge, we formulate two tasks: *Knowledge Identification* and *Knowledge Leveraging*. We evaluate the tasks in CS and English questions, the two baselines that share the same matrix language.

4.1 Knowledge Identification

Task Description. When a human English-Korean bilingual is given a question about Korean culture, they will first try to identify what specific knowledge is required to answer the question, and then apply the knowledge to find the correct answer. Depending on which language the question is written in, the quantity and quality of the knowledge may vary, as described in Figure 1. Language-specific knowledge is likely to be encoded much abundantly in its own language, so reading the question in CS will allow more effective knowledge activation than in English. In this sense, knowledge identification task evaluates LLMs' ability to identify what knowledge is prerequisite for the question. Specifically, the LLM is asked to write a list of factual knowledge that are necessary for solving the given question in one or two sentences.

Evaluation Criteria. For a qualitative analysis on knowledge identification, we evaluate the quality of a knowledge list based on two criteria: *Faithfulness* evaluates whether the generated knowledge is factually correct and the model does not output hallucination. *Helpfulness* evaluates whether the knowledge is relevant to the question, and helpful for answering the question correctly.

4.2 Knowledge Leveraging

Task Description. We refer to Knowledge Leveraging as applying the identified knowledge into reasoning. In specific, the model should be able to find a correct answer based on the knowledge it has identified from the Knowledge Identification task. Therefore, we provide knowledge identified by each model and instruct the model to find the answer using the knowledge. To encourage the models to properly ground on knowledge, we adopt Chain-of-Thought reasoning (Wei et al., 2023) and prompt the models to generate reasoning steps that lead to the final answer. We conduct experiments on the entire dataset and report accuracy score.

4.3 Experimental Setup

Implementation Details. The top two baselines that showed excellent performance in Section 3.3, CS, and English, are chosen for evaluation. We conduct experiments using the same models as in Section 3.2. For knowledge identification, we instruct the model to write a list of factual knowledge that are required for solving the given question in one or two sentences. For knowledge leveraging, we pass on previously identified knowledge and ask the model to select an answer and explain why. The full-length prompts are provided in Table 8 and 9.

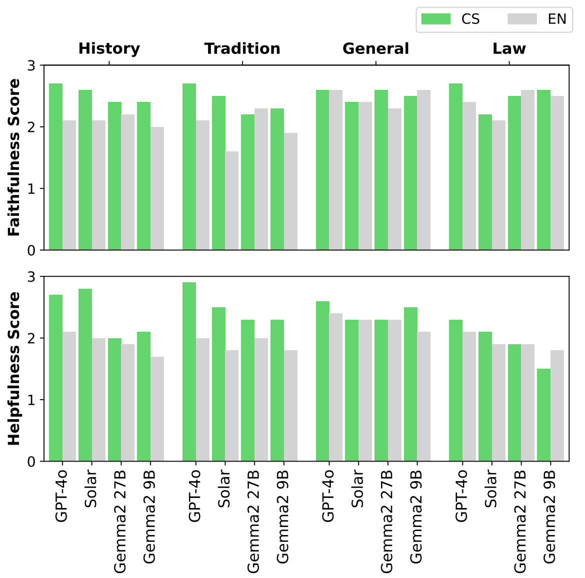

Figure 2: Human evaluation results on faithfulness (*top*) and helpfulness (*bottom*) of knowledge lists identified from **CS** questions and **English** questions.

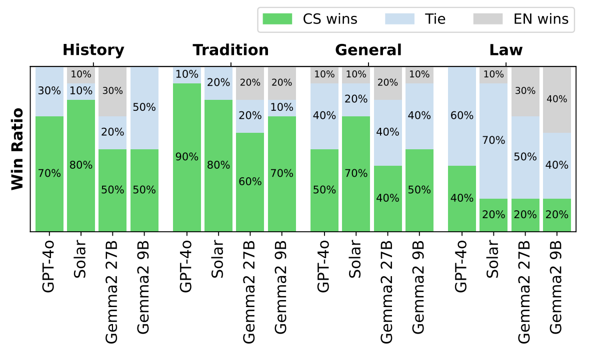

Figure 3: Human evaluation results on pairwise comparisons between knowledge lists identified from **CS** questions and **English** questions.

Evaluating Knowledge Identification. To effectively evaluate knowledge identification from CS and English questions, we refer to Section 3.3 and choose two domains where CS performance is higher (*i.e.*, History, Tradition), and two domains that have minimum difference (*i.e.*, General, Law). Moreover, we select four models with different performances and sizes (*i.e.*, GPT-4o, Solar, Gemma2 27B, Gemma2 9B). Specifically, we sample 10 questions from each domain and model, resulting in 160 samples. Then, we conduct human and LLM-based evaluation on identified knowledge.

Human Evaluation We employ four human evaluators who are fluent in both Korean and English and completed Korean public education, thus qualified to evaluate questions sourced from Korean proficiency tests for foreigners and the Korean College Scholastic Ability Test. For faithfulness and helpfulness, the evaluator is asked to rate a knowledge list on a Likert scale from 1 to 3. In pairwise evaluation, we provide two knowledge lists in a random order and ask the evaluator to select a list that is overall more effective for answering the question. Details on evaluation criteria and evaluator information are provided in Appendix C.1 and C.2.

LLM-based Evaluation As we conduct human evaluation on quite small amount of samples, we additionally conduct LLM-as-a-judge evaluation (Zheng et al., 2023) to amplify our analysis. Specifically, we use GPT-4o as the evaluator, using identical instructions with human evaluators on 40 questions for 9 domains and 8 models, 360 samples in total. Full prompts are provided in Appendix C.1.

5 Analysis on Knowledge Identification

5.1 Human Evaluation

Faithfulness. In the upper row of Figure 2, we observe a significant gap in faithfulness scores between CS and English in both History and Tradition. The discrepancy is more salient in Tradition where cultural nuances is much important, implying that asking questions in CS is much successful in capturing cultural nuances and meanings. In General domain, the scores for CS and English are almost identical (or even better in English for Gemma2 9B), indicating that the difference in knowledge activated by CS questions compared to English questions is minimal when addressing general and common facts. In Law, although knowledge from CS is slightly more faithful than that from English, their absolute scores are lower than those in other domains, suggesting that models fail to identify faithful knowledge that requires domain expertise.

Helpfulness. The lower row of Figure 2 presents evaluation results for helpfulness. It is intuitive that faithful knowledge serves as a valuable source for answering questions, and as a result, the evaluation of helpfulness shows a similar trend to that of faithfulness. In History and Tradition, the gap between CS and English becomes larger in helpfulness, emphasizing the effectiveness of the CS setting in identifying both faithful and helpful knowledge. It is also notable that the scores for helpfulness are particularly high for GPT-4o and Solar, models in which performance in CS surpasses that in English to a large extent (§ 3.3). In contrast, the

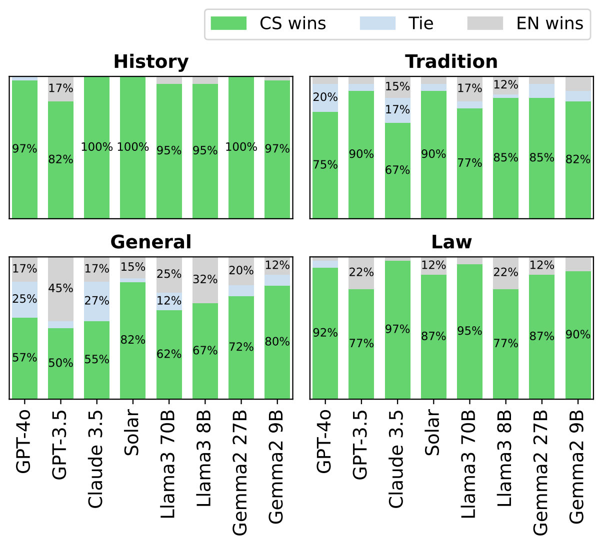

Figure 4: LLM-as-a-judge evaluation results on pairwise comparison between knowledge lists identified from **CS** questions and **English** questions.

Model	History	Tradition	General	Law
GPT-4o	0.41	0.64	0.62	0.62
Solar	0.26	-0.09	0.38	0.02
Gemma2 27B	0.25	0.52	0.17	0.34
Gemma2 9B	0.20	-0.07	0.05	0.24

Table 2: Cohen's kappa (κ) correlation scores between human and LLM-as-a-judge evaluation. Gray indicates poor agreement.

helpfulness scores in the Law domain are considerably lower for both CS and English compared to other domains. Given that the Law domain requires expert-level legal knowledge, the models struggle to grasp the legal context, leading to difficulties in accurately identifying helpful knowledge sources from both CS and English questions.

Pairwise Comparison. In Figure 3, the win ratio for CS is higher in History and Tradition, demonstrating that CS questions can activate more essential knowledge sources for question answering. On the contrary, in domains where CS does not show its effectiveness, the win ratio of CS is comparatively lower (*i.e.*, General) or the ratio of Tie is high (*i.e.*, Law). Especially in the case of Law, the quality of knowledge lists generated from CS questions is evaluated as equivalent to, or even worse than, that generated from English questions.

5.2 LLM-based Evaluation

We observe in Figure 6 and Figure 7 that the score gap between CS and English in both faithfulness and helpfulness are minimal. In fact, CS scores are even or lower for some cases, which are inconsistent with human evaluation results. However, it is still worth noting that LLM-as-a-judge also assigns higher scores for advanced models, and overall scores were lower in History and Tradition.

On the other hand, LLM judgement scores in pairwise evaluation generally agree with the human evaluations. We compute Cohen's Kappa (κ) score in Table 2, and follow interpretations from Landis and Koch (1977).[2] Consistent with human evaluation, the LLM judge votes CS for most cases, and the agreement is stronger with advanced models (*i.e.*, GPT-4o), on culture-intensive domains (*i.e.*, History, Tradition).

While other domains fairly agree with human judgment, Law shows exceptional results. Specifically, the LLM-as-judge evaluation reports a significantly higher win ratio for CS in the Law domain compared to human evaluation. However, considering that tie ratio is substantial in human evaluation as well, we speculate that LLM-as-a-judge gives a win to CS on knowledge that human evaluators regarded comparable quality with English setting.

6 Analysis on Knowledge Leveraging

We present the visualized results of accuracy in both CS and English settings in Figure 5, with detailed scores reported in Table 11. Consistent with the results in Section 3.3, all models demonstrate generally higher accuracy for CS questions compared to English questions, indicating that CS effectively activates knowledge across various domains. To be specific, the performance in the CS setting exceeds that of English in every domain for GPT-4o, Claude 3.5, and Solar. These models not only identify faithful and helpful knowledge (§ 5.1), but also answer questions while accurately grounding on that knowledge; this shows that CS questions robustly activate essential knowledge. On the contrary, Llama3 and Gemma2 families do not seem to benefit from CS questions. Specifically, these models show poor performance in both CS and English settings in several domains, such as Geography and Law. Taking into account that these domains require domain-specific expertise, it is likely that their lack of understanding contributes to low accuracy, let alone CS failing to activate Korea-focused knowledge.

[2] Landis and Koch (1977) interprets 0–0.20 as slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1 as almost perfect agreement.

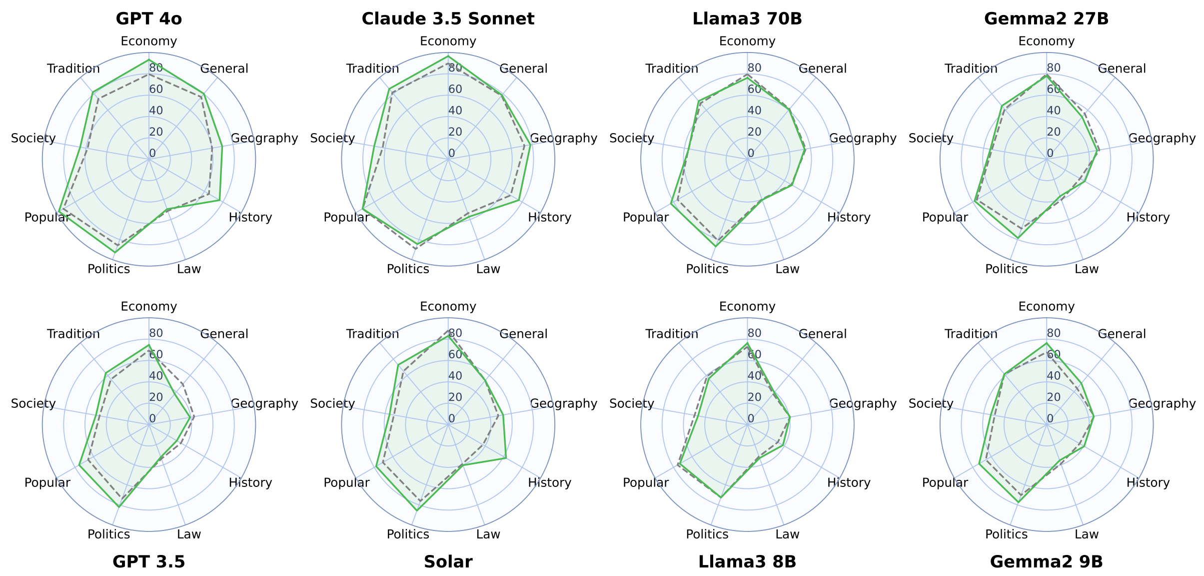

Figure 5: Radar charts of knowledge leveraging performances on all domains across various multilingual LLMs. **Green** line is code-switching and dashed **gray** line is English. We report accuracy for the evaluation metric.

English Questions hallucinate more than CS. Although we informed the models that the answer is in one of the choices, we notice that the majority of incorrect responses were "None of the above". The errors may derive from either hallucinated knowledge or failing to follow instructions faithfully. Therefore, we provide additional analysis on erroneous outputs in Table 12. We report the results in the format of *number of errors that derived from knowledge hallucination / total number of None errors*. Errors that are not from hallucination are caused by poor instruction-following. Overall, we observe that performance on English questions results in more errors compared to CS across all LLMs, and most of them were hallucination errors. This indicates that models hallucinate much frequently when English questions are given, again highlighting the effectiveness of CS over English setting. It is also worth noting that Gemma2 families hallucinate largely on History and General, supporting our finding in Figure 2 and Figure 5 which respectively illustrates poor performance on human evaluation and QA accuracy.

Case Study. We examine a sample case to compare the capability of code-switching and English on knowledge activation. Table 13 shows the knowledge and answer generated by Solar in Tradition. The question asks about "정월대보름", a Korean traditional holiday that celebrates the first full moon of lunar new year. We observe that CS question preserves unique terms such as "정월대보름" and "귀밝이술" in Korean; this helps the model to successfully activate faithful knowledge, consequently leading to the correct answer. However, in the case of English, not only are these cultural nuances lost in English question, but the model misunderstood the question to asking about "Dan-o", another Korean traditional holiday. Solar lacks in knowledge about "정월대보름" in English, or fails to activate encoded knowledge with its English translation. We also provide a case study of CS failing in knowledge activation in Appendix D.4.

7 Conclusion

In this paper, we explore the efficacy of code-switching in activating language-specific knowledge embedded in LLMs. Utilizing two Korean-centric QA datasets, we synthesize EnKoQA, a qualified English-Korean code-switching QA dataset. We formulate two tasks and conduct experiments across various multilingual LLMs. Our analyses demonstrate that code-switching can effectively activate knowledge within LLMs compared to English text, particularly in language-specific domains. Our work suggests the potential of code-switching as an effective strategy for eliciting language-specific knowledge from LLMs. We hope our work can motivate NLP community to explore more potential of code-switching in diverse aspects, and leverage them as an effective tool to train and instruct multilingual LLMs.

Limitations and Future Work

In this work, we focus on code-switching between English and Korean, specifically limiting the scope to Korea-specific knowledge. However, it is important to note that this study serves as a single case focused on the Korean context and leaves room for expanding the scope of code-switching to other cultures and languages. For future research, we aim to investigate whether the knowledge activation effect also occurs in other language settings.

Another limitation of our work is that we conduct human evaluations on only a subset of LLMs, domains, and questions. Evaluating the quality (*i.e.*, faithfulness and helpfulness) of knowledge in code-switched text presents inherent and practical challenges, as it necessitates evaluators to be fluent bilinguals. Consequently, we present only partial results for the knowledge identification task.

Lastly, as we rely on a LLM, specifically `gpt-3.5-turbo`, to synthesize our code-switching dataset, the performance of the LLM can affect the quality of the dataset. To mitigate the risk of erroneous samples and to fully leverage the LLM's capabilities, we engage reliable human annotators to review the samples and verify their quality.

In the future, we aim to investigate more potential of code-switching in diverse aspects, including instruction-tuning of LLMs to users effectively using code-switching for multilingual tasks. As we have demonstrated synthesizing monolingual datasets into code-switching text, we hope our work can inspire NLP community to explore the capability of code-switching in enhancing and utilizing multilingual LLMs.

Ethical Consideration

Our work utilizes large language models for data construction. Recent work has highlighted the risks of LLMs in hallucination (Zhang et al., 2023b). In order to prevent any hallucination or harmful contents, we ensure that human annotators examined each sample carefully and create dataset safely.

References

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

Anthropic. 2024. Claude 3.5 sonnet.

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

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

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size.

Roberto Heredia and Jeanette Altarriba. 2001. Bilingual language mixing: Why do bilinguals code-switch? *Current Directions in Psychological Science - CURR DIRECTIONS PSYCHOL SCI*, 10:164–168.

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

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

Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. 2024a. Solar 10.7b: Scaling large language models with simple yet effective depth upscaling. *Preprint*, arXiv:2312.15166.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024b. CLIcK: A benchmark dataset of cultural and linguistic intelligence in Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3335–3346, Torino, Italia. ELRA and ICCL.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on*

Empirical Methods in Natural Language Processing, pages 10443–10461, Singapore. Association for Computational Linguistics.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. 2024. tinyBenchmarks: evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 34303–34326. PMLR.

C. Myers-Scotton. 1997. *Duelling Languages: Grammatical Structure in Codeswitching*. Clarendon Press.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Lorenzo Pacchiardi, Lucy G. Cheke, and José Hernández-Orallo. 2024. 100 instances is all you need: predicting the success of a new llm on unseen data by testing on a few instances. *Preprint*, arXiv:2409.03563.

Chanjun Park, Hyeonwoo Kim, Dahyun Kim, Seonghwan Cho, Sanghoon Kim, Sukyung Lee, Yungi Kim, and Hwalsuk Lee. 2024. Open ko-llm leaderboard: Evaluating large language models in korean with ko-h5 benchmark. In *The 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*.

Eunsun Park and Hongoak Yun. 2021. The grammatical constraint and grammatical encoding of Korean-English code switching. *The Journal of Mirae English Language and Literature*, 26(1):177–204.

Shana Poplack. 1980. Sometimes i'll start a sentence in spanish y termino en espaÑol: toward a typology of code-switching 1. *Linguistics*, 18:581–618.

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

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

Cesa Salaam, Franck Dernoncourt, Trung Bui, Danda Rawat, and Seunghyun Yoon. 2022. Offensive content detection via synthetic code-switched text. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6617–6624, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Bhavani Shankar, Preethi Jyothi, and Pushpak Bhattacharyya. 2024. In-context mixing (ICM): Code-mixed prompts for multilingual LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4162–4176, Bangkok, Thailand. Association for Computational Linguistics.

Guijin Son, Hanwool Lee, Suwan Kim, Huiseo Kim, Jae cheol Lee, Je Won Yeom, Jihyu Jung, Jung woo Kim, and Songseong Kim. 2024. HAE-RAE bench: Evaluation of Korean knowledge in language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7993–8007, Torino, Italia. ELRA and ICCL.

Vivek Srivastava and Mayank Singh. 2021. Challenges and limitations with the metrics measuring the complexity of code-mixed text. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 6–14, Online. Association for Computational Linguistics.

Rajan Vivek, Kawin Ethayarajh, Diyi Yang, and Douwe Kiela. 2024. Anchor points: Benchmarking models with much fewer examples. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1576–1601, St. Julian's, Malta. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

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

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

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023b. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Dataset Details

A.1 Details of Source Data

We used two datasets in our experiments:

HAE RAE BENCH 1.1 is available at `https://huggingface.co/datasets/HAERAE-HUB/HAE_RAE_BENCH_1.1`.

CLIcK is available at `https://huggingface.co/datasets/EunsuKim/CLIcK`.

A.2 Statistics of EnKoQA

We provide statistics of EnKoQA per domain in Table 3.

Domain	#
Economy	59
General	176
Geography	281
History	468
Law	435
Politics	84
Popular	41
Society	606
Tradition	222
Total	2,372

Table 3: Number of samples in EnKoQA.

A.3 Code-Mixing Index

We report CMI scores for our dataset in Table 4. In specific, we tokenized the sentence using `bert-base-multilingual-cased`, then removed all noisy tokens such as numbers or tags and counted the ratio of $\frac{num\ of\ Korean\ tokens}{num\ of\ all\ tokens}$. We report the distribution of QA accuracy on different CMI scores in Tradition and History, two domains where CS proved its effectiveness. If CMI is close to 0, sentence is mostly written in English, and close to 100 means vice versa. The number of samples at each end (0-10, 90-100) was very small, causing outliers. We can see that accuracy scores are quite evenly distributed across all ratios, suggesting that there is no distinct tendency between CMI and accuracy performance.

It is important to note, however, that code switching metrics such as CMI, while offering a quantitative measure of token-level composition, are inherently limited in capturing the nuanced semantic and syntactic characteristics of code-switched texts. These metrics primarily rely on surface-level token ratios, which can inadvertently assign high scores to linguistically or contextually meaningless sequences. Consequently, they may over-represent the presence of meaningful code-switching patterns while failing to account for the deeper linguistic interplay that defines effective code-switching. For a more comprehensive discussion of these limitations, please refer to Srivastava and Singh, 2021.

A.4 Quality Control Guideline

We provide a guideline we used to filter the candidates and select the final candidate.

- Is the question written in English-Korean code-switching, where matrix language is English and semantically important Korean words are embedded into English sentence?
- Do choices also follow the code-switched pattern of query?
- Does the syntactic structure of the sentence follow that of English?
- Are semantically important nouns and noun phrases from Korean sentence, and are they embedded into English sentence?
- Are functional words and grammatical morphemes kept in English?

A.5 Annotation Details

For dataset construction, two Korean native annotators with expert knowledge in Korean culture and equivalently fluent in English manually examine the candidates and select the most naturally code-switched question, then cross-checked each other's assigned share of dataset. If a selected candidate appeared to be incorrect or suboptimal, the annotators engaged in thorough discussions until they reached an agreement on the most appropriate candidate.

Regarding inter-annotator agreement (IAA), although we did not compute a formal IAA score, significant effort was devoted to ensuring high annotation quality through extensive discussion and collaboration among annotators. In specific, the annotation process involved annotators who are fluent in both English and Korean are assigned each

CMI	Tradition				History			
	Solar	Gemma 2 9B	Gemma 2 27B	GPT-4o	Solar	Gemma2 9B	Gemma2 27B	GPT-4o
0–10	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
10–20	100.0	100.0	100.0	100.0	50.00	50.00	100.0	50.00
20–30	56.25	65.62	53.12	71.88	50.00	31.58	36.84	60.53
30–40	72.34	59.57	55.32	76.6	66.67	48.72	45.3	77.78
40–50	80.39	62.75	70.59	84.31	69.54	42.38	42.38	78.15
50–60	83.72	74.42	74.42	93.02	62.35	40.00	35.29	74.12
60–70	85.19	66.67	66.67	92.59	50.00	28.85	46.15	63.46
70–80	66.67	58.33	66.67	91.67	57.89	15.79	21.05	57.89
80–90	66.67	83.33	83.33	100.0	50.00	50.00	25.00	100.0
90–100	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00

Table 4: Distribution of QA accuracy on different CMI scores in Tradition and History. If CMI is close to 0, sentence is mostly written in English, and close to 100 means vice versa. The number of samples at each end (0-10, 90-100) was very small, causing outliers.

portion of the dataset to select a candidate for code-switched question. Following this initial annotation, the annotators cross-checked each other's work to identify any discrepancies. If a selected candidate appeared to be incorrect or suboptimal, the annotators engaged in thorough discussions until they reached an agreement on the most appropriate candidate. This iterative and collaborative process was integral to constructing a high-quality dataset.

A.6 Dataset Size and Quality

Discussion on Dataset Size While we acknowledge the relatively limited size of EnKoQA dataset, we emphasize that quality often matters more than quantity as many studies (Pacchiardi et al., 2024; Maia Polo et al., 2024; Vivek et al., 2024) have demonstrated. Please note that we prioritized creating a high-quality dataset with rigorous manual validation and linguistic alignment, ensuring that the dataset serves as a reliable resource for code-switching research. Additionally, while the size of Korean datasets is often limited given that Korean is a low-resource language, EnKoQA dataset is comparatively larger than the sizes of other Korean datasets. For instance, datasets in the Open Ko-LLM leaderboard (Park et al., 2024), such as Ko-ARC (1.1k), Ko-TruthfulQA (0.8k), and Ko-CommonGen (0.8k), are all smaller in scale than EnKoQA's 2,372 question-answer pairs. This highlights our effort to provide a relatively extensive resource within the constraints of dataset availability for minor languages.

Specifically, our quality control process includes human annotators thoroughly reviewing all LLM-generated samples to assess the quality and naturalness. When any errors or unnatural code-switching patterns were identified, annotators corrected them to ensure that the final dataset adheres to high standards of our quality control. In that sense, GPT-3.5-turbo served as an assistive tool for providing initial candidates, rather than generating final outputs. Therefore, we assert that any potential shortcomings of the translation tool were effectively mitigated through this meticulous human review and correction process.

Translating with GPT-3.5 We have conducted experiments on both GPT-3.5 and GPT-4o for translation and code-switching generation tasks. Interestingly, we observed that after manual examination and correction process, the results from both models were comparable in terms of quality and naturalness. This is due to our rigorous human-in-the-loop workflow that ensures any errors or unnatural expressions are taken care of, regardless of the initial model used. Given this finding, we used GPT-3.5 for its cost efficiency while maintaining high-quality standards through meticulous human examination and refinement. By prioritizing manual validation, we ensured that the final dataset reflects linguistic accuracy and naturalness, independent of the model used for preliminary generation.

A.7 Data Sample

We also provide a sample of original Korean, translated English, and synthesized CS example question in Table 5. Note that unique terms or semantically important words are properly embedded in Korean.

B Experimental Details

B.1 Computational Resources and API Cost.

Llama3 and Gemma2 models. We used Huggingface model cards and run them on two NVIDIA A100 GPUs. Specifically, we used `meta-llama/meta-llama-3-8b-instruct`, `meta-llama/meta-llama-3-70b-instruct`, `google/gemma-2-9b-it`, `google/gemma-2-27b-it`.

GPT-3.5 and GPT-4o. We used up-to-date versions of `gpt-3-5-turbo` and `gpt-4o` APIs. The cost for `gpt-3-5-turbo` was $15 for EnKoQA generation and $6 for experiment inference, while the cost for `gpt-4o` was $23 for experiment inference.

Claude 3.5. We used `claude-3-5-sonnet` API from Anthropic AI[3]. The cost for `claude-3-5-sonnet` was $21 for experiment inference.

Solar We used `solar-mini` API from Upstage[4].

B.2 Prompts

We provide the following prompts used in our experiments. Table 6 contains the prompt used for generating code-switched text candidates across different levels of linguistic complexity. For QA inference tasks, we used the prompt presented in Table 7. The prompt for identifying relevant knowledge in a given context is provided in Table 8, while Table 9 shows the prompt used for leveraging this identified knowledge in downstream tasks.

B.3 Open-ended QA

Out dataset, EnKoQA is multiple-choice QA dataset, following its original source datasets. We additionally explore the potential of code-switching on open-ended QA as well.

Results are shown in Table 10. Using same questions in our dataset, we instruct the model to respond in short answer and compute exact match score. It is noticeable that the performances are very low compared to multiple-choice QA results. We attribute this to the free-form response of open-ended tasks, causing more errors and hallucinations. It is observable that the models barely answer correctly in History and Popular.

C Evaluation Details

C.1 Evaluation Criteria

We provide evaluation guideline for human evaluation.

Faithfulness. Faithfulness evaluates the factual correctness of the knowledge.

- Knowledge list is very faithful. Every knowledge is factually correct.
- Knowledge list is somewhat faithful. Some, not every, knowledge is factually correct.
- Knowledge list is not faithful at all. Every knowledge is hallucinated.

Helpfulness. Helpfulness evaluates how useful the knowledge is for answering the question.

- Knowledge list is very helpful. Every knowledge is relevant to the question, and used for finding the answer.
- Knowledge list is somewhat helpful. Some, not every, knowledge is useful for finding the answer.
- Knowledge list is not helpful at all. All knowledge are irrelevant with the question.

Pair-wise comparison. We comprehensively evaluate the quality of knowledge generated from CS and English questions in terms of both faithfulness and helpfulness. If both are identical, evaluators can choose *Tie*.

In case of LLM-as-a-judge evaluation, same criteria and instructions are given as prompts.

C.2 Human Evaluator Qualifications

For knowledge identification evaluation, collecting qualified bilingual evaluators was not easy due to the inherent challenge in code-switching research of necessitating fluent bilinguals as evaluators. Our dataset is composed of questions from Korean proficiency tests for foreigners and the Korean College Scholastic Ability Test. Thus, it is designed

[3] https://www.anthropic.com/
[4] https://www.upstage.ai/

Lang	QUESTION	CHOICES
KO	다음 글의 (가)에 대한 (나)의 상대적 특성으로 옳은 것은? (단, (가), (나)는 각각 겨울과 여름 중 하나임.) 우리나라는 더위와 추위에 대비하여 대청마루와 온돌 같은 전통 가옥 시설이 발달하였다. 대청마루는 바람을 잘 통하게 하여 (가) 을 시원하게 지낼 수 있도록 설치되었다. 온돌은 아궁이의 열을 방으로 전달하여 (나) 을 따뜻하게 지낼 수 있도록 설치되었다. 대청마루 는 중부와 남부 지역에 발달한 한편, 온돌은 대부분의 지역에 발달하였다.	(1) 평균 상대 습도가 높다. (2) 정오의 태양 고도가 높다. (3) 한파의 발생 일수가 많다. (4) 대류성 강수가 자주 발생한다. (5) 열대 저기압의 통과 횟수가 많다.
EN	What is the correct relative characteristic of (나) in relation to (가) in the following passage? (Note that (가) and (나) refer to either winter or summer.) In Korea, traditional house facilities such as daecheongmaru and ondol have developed to cope with heat and cold. Daecheongmaru is designed to allow good ventilation to keep (가) cool. Ondol transfers heat from the kitchen stove to the room to keep (나) warm. While daecheongmaru is developed in the central and southern regions, ondol is developed in most areas.	(1) The average relative humidity is high. (2) The midday sun's altitude is high. (3) There are many days of occurrence of cold waves. (4) Heavy rainfall often occurs in Daeryuseong. (5) There are many occurrences of passage of tropical cyclones.
CS	What is the correct relative characteristic of (나) in relation to (가) in the following passage? (Note that (가) and (나) refer to either winter or summer.) In 한국, 전통 가옥 시설 such as 대청마루 and 온돌 have developed to cope with heat and cold. 대청마루 is designed to allow good ventilation to keep (가) cool. 온돌 transfers heat from the kitchen stove to the room to keep (나) warm. While 대청마루 is developed in the 중부 and 남부 지역, 온돌 is developed in most areas.	(1) The average 상대 습도 is high. (2) The 정오의 태양 고도 is high. (3) There are many days of occurrence of 한파. (4) 대류성 강수 often occurs. (5) There are many occurrences of passage of 열대 저기압.

Table 5: An example of Korean, English, and CS from dataset.

at a level that would not be challenging for evaluators whom were born and raised in Korea, received a Korean public education, and graduated prestigious universities. We managed to collect four Korean graduate school students as our evaluators, all of whom are native Korean with sufficient understanding of Korean culture. Also, they possess qualified English exam scores, indicating that they have no problem in understanding Korean-English code-switched texts. To mitigate the shortage of labor force, we designed the evaluation criteria objectively, allowing for an assessment that is not subjective and has clear correct answers. Specifically, we evaluate knowledge identification based on two criteria: faithfulness and helpfulness. Faithfulness evaluates the factualness of the knowledge, so the evaluators are required to use their background knowledge as well as searching from faithful sources where gold knowledge exists. To evaluate helpfulness, evaluators are given a gold answer to the question and determine whether the knowledge is helpful for finding the answer, using their logical reasoning.

D Observations

In this section, we provide additional results and comprehensive observations throughout our work.

D.1 Knowledge Identification Results

We observed that the majority of models benefitted from CS questions. Table 1 shows that scores in CS are higher on all models in Politics, and in case of Law, only three models (GPT-3.5, Llama3 70B, and Gemma2 9B) out of eight models performed worse. We can see in Average score, all models except Gemma2 27B performed better on CS.

D.2 Knowledge Leveraging Results

We provide accuracy results of Knowledge Leveraging in Table 11. Figure 5 is a visualization of this table.

D.3 Error Analysis

We provide full results of error counts in Table 12. Note that as models get smaller and show poor performance in Korean, the number of errors increase. (See Gemma2 families.)

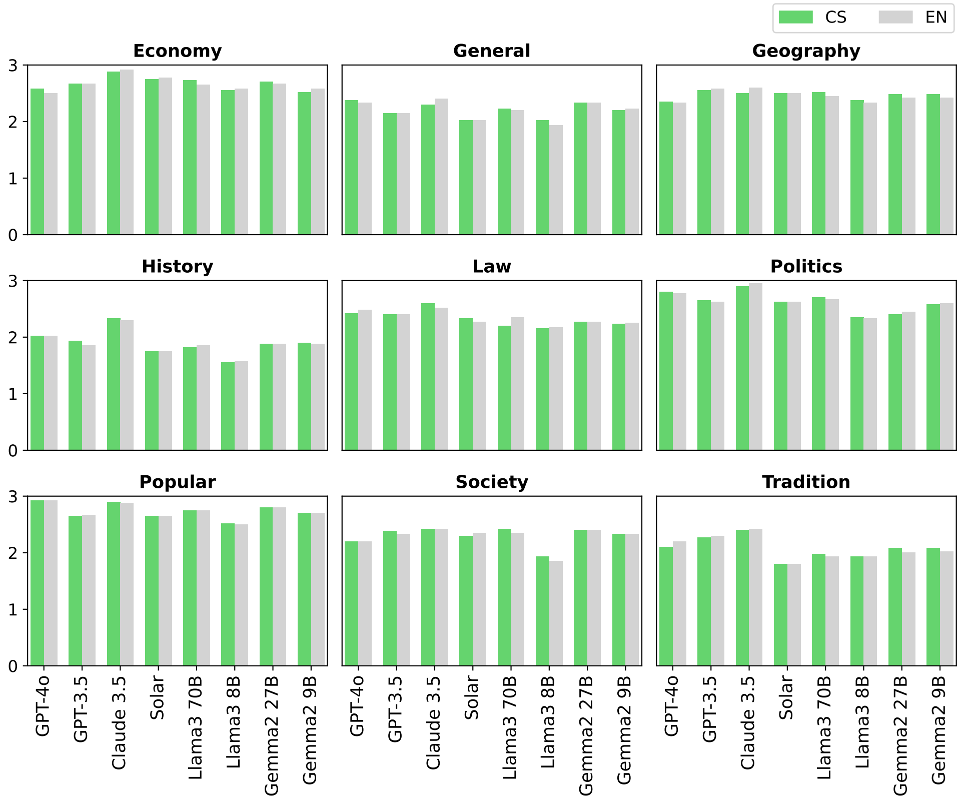

Figure 6: LLM-as-a-judge evaluation results on faithfulness between knowledge lists identified from CS and English questions.

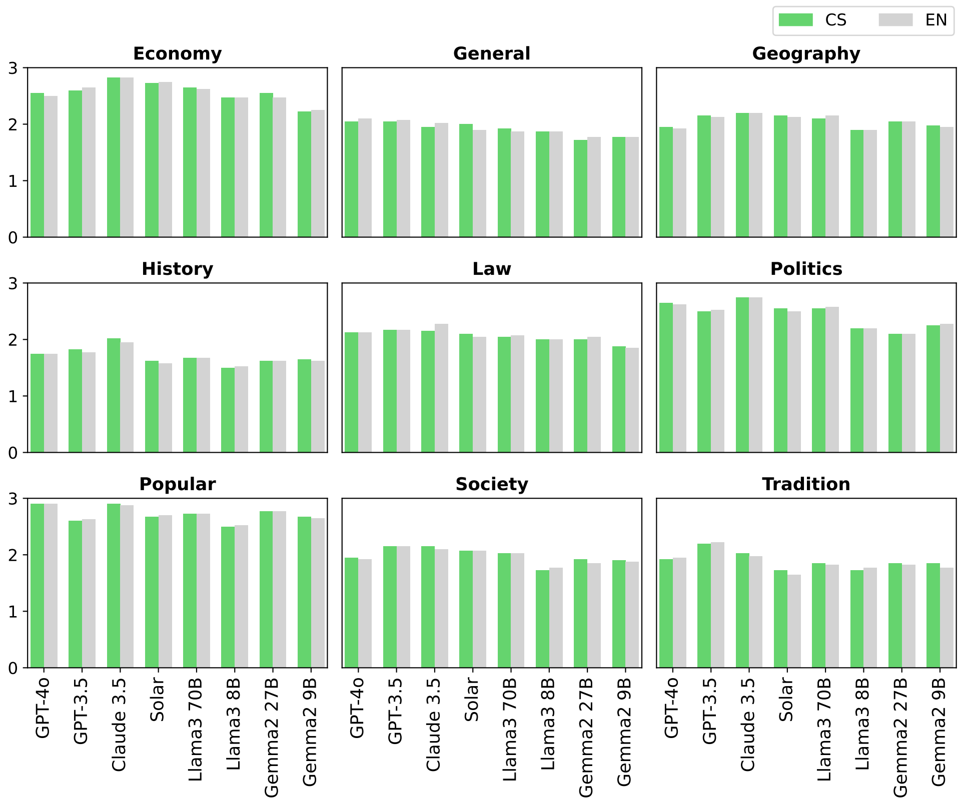

Figure 7: LLM-as-a-judge evaluation results on helpfulness between knowledge lists identified from CS and English questions.

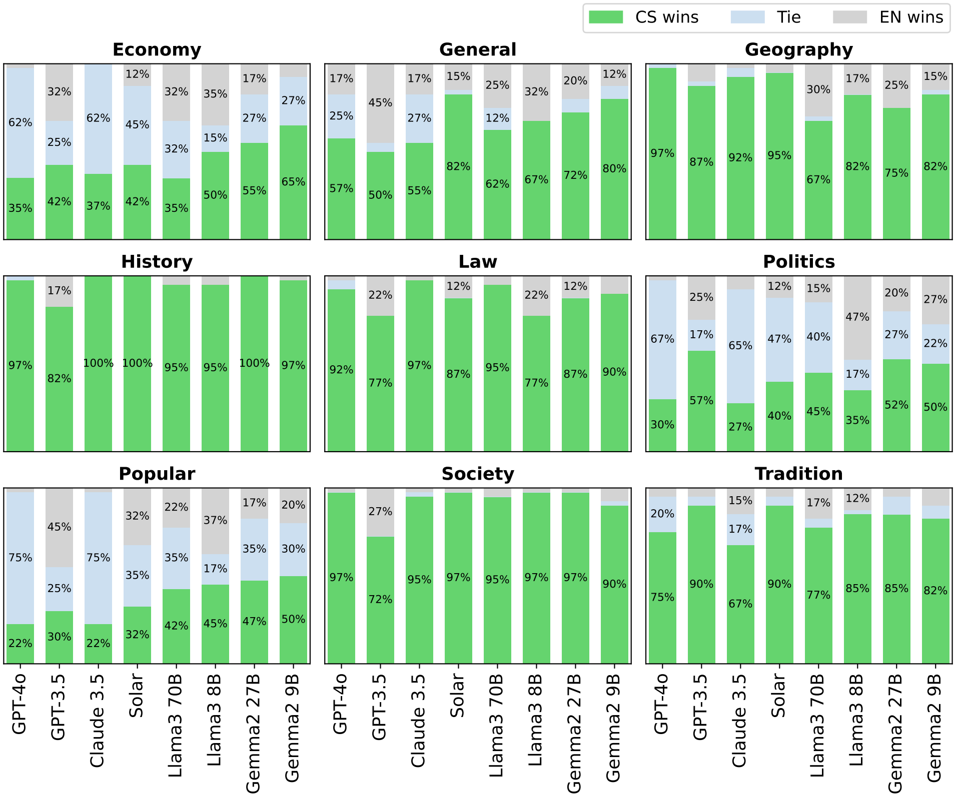

Figure 8: LLM-as-a-judge evaluation results on pairwise comparison between knowledge lists identified from CS and English questions.

D.4 Case Study

In this section, we provide a case study of Gemma2 9B on Law domain. In Table 14, hallucinations are observed in the knowledge generated from CS question. According to the Civil Act of the Republic of Korea, individuals under the age of 14 can only enter into binding contracts with the consent of their legal guardians. Additionally, individuals between the ages of 14 and 19 are not deprived of contractual effect; rather, they are granted the right to cancel such agreements at their discretion. Moreover, the knowledge generated in English incorrectly applies the U.S. standard, which defines minors as those under 18 years of age, instead of the Korean standard, which applies to individuals under 19 years of age. This finding suggests that English question is not helpful for identifying necessary and language-specific knowledge.

Prompt for generating CS candidates

You are a bilingual who can speak both English and Korean fluently. I will give you a Korean and English (<Korean>, <English>) pair. They are semantically the same. Your job is to write Korean-English code-switching text with certain switching level by mixing Korean and English text. Example 1 is an example of code-switched text in different levels of code-switching. Use Example 1 as reference to understand the level of code-switching. Read the instructions carefully and solve the Task.

Instructions:
- Maintain English word order, that is, Subject-Verb-Object.
- Find semantically important given nouns and noun phrases from the text, and change `{level}` percent of them to Korean.
- Keep functional words in English.
- Keep the indicators such as (가), (나), ㄱ, ㄴ, 갑, 을 in Korean.

[Example 1]
<Korean>
제주도는 점성이 작고 유동성이 큰 마그마가 여러 차례 분출하여 형성된 방패 모양의 화산섬이다. 하지만 한라산의 정상부는 종 모양의 화산으로 이루어져 있으며, 산허리에는 오름으로 불리는 기생화산이 많이 형성되어 있다.

<English>
Jeju Island is a shield-shaped volcanic island formed by multiple eruptions of small-sized and highly fluid magma. However, the top of Hallasan Mountain consists of a cone-shaped volcano, and many parasitic volcanoes called Oreum are formed on the hillsides.

<Code-switch with 30 percent of Korean>
Jeju Island is a shield-shaped 화산섬 formed by multiple eruptions of small-sized and highly fluid magma. However, the top of 한라산 consists of a cone-shaped volcano, and many 기생화산 called 오름 are formed on the hillsides.

<Code-switch with 50 percent of Korean>
Jeju Island is a 방패 모양의 화산섬 formed by multiple eruptions of small-sized and highly fluid 마그마. However, the top of 한라산 consists of a cone-shaped 화산, and many 기생화산 called 오름 are formed on the hillsides.

<Code-switch with 70 percent of Korean>
제주도 is a 방패 모양의 화산섬 formed by multiple eruptions of 크기가 작고 유동성이 큰 마그마. However, the top of 한라산 Mountain consists of a 종 모양의 화산, and many 기생화산 called 오름 are formed on the 산허리.
<Code-switch with 90 percent of Korean>
제주도 is a shield-shaped 화산섬 formed by multiple 분출 of small-sized and 유동성이 큰 마그마. However, the 정상부 of 한라산 consists of a cone-shaped 화산, and many 기생화산 called 오름 are formed on the 산허리.

[Task]
<Korean>
`{question}`

<English>
`{translation}`

<Code-Switch>

Table 6: Prompt for generating code-switched text candidates in diffferent levels.

Prompt for QA (CS)
You will be given a question and choices about Korea. The text are written in English-Korean code-switching, where matrix language is English and semantically important Korean words are embedded into English sentence. Your job is to answer the question. Read the [QUESTION] and choose the most appropriate answer from [CHOICES]. Only write your answer number in parentheses, like (1). Do not repeat the question or choice. Use Example 1 as a reference to answer Example 2. <Example 1> [QUESTION] Which city is the 수도 of 한국? [CHOICES] (1) 뉴욕 (New York) (2) 서울 (Seoul) (3) 파리 (Paris) (4) 도쿄 (Tokyo) [ANSWER] (2) <Example 2> [QUESTION] {question} [ANSWER]

Table 7: Prompt for QA inference.

Prompt for Knowledge Identification
You are a bilingual who is fluent in both Korean and English, and is knowledgeable about South Korea. You will be given a multiple choice question about South Korea. The text are written in English-Korean code-switching, where matrix language is English and semantically important Korean words are embedded into English sentence. Your job is to follow the instructions and write a list of knowledge that is necessary to know for solving the question correctly. Instructions: - Write a list of factual knowledge that are required for solving the question. Try to write each knowledge in one or two sentences. You can write in whichever language you can explain better, either Korean or English. Start this task with [KNOWLEDGE] tag. - Only write knowledge that you definitely know. Do not write incorrect information. - Do not repeat input text in your response. Do not generate new question. Stick to input text that is given to you. I will give you an example for reference. «Example 1» [QUESTION] Read the following question and choose the most appropriate answer. Who is the person who greatly defeated the soldiers of the 당나라 in the 안시성 싸움? [CHOICES] (1) 양만춘 (2) 서희 (3) 김유신 (4) 강감찬 (5) 윤관 [KNOWLEDGE] 1. 안시성 싸움 (Siege of Ansi): 안시성 싸움 (645 AD) was a famous military conflict between 고구려 and the 당 Dynasty. 고구려, under the leadership of 양만춘 (Yang Man-chun), successfully defended the 안시성 against the powerful 당 forces led by Emperor 태종. 2. 양만춘 (Yang Man-chun): He was the general who commanded the defense of 안시성, playing a key role in defeating the 당나라 army. 3. 서희 (Seo Hee): A 고려 diplomat famous for negotiating with the 거란 to avoid invasion, but not involved in the 안시성 싸움. 4. 김유신 (Kim Yu-shin): A general from the 신라 Kingdom, instrumental in the unification of the 한반도, but not involved in this specific battle. 5. 강감찬 (Gang Gam-chan): A 고려 military commander known for his victory over the 거란 in the 귀주대첩, unrelated to 안시성. 6. 윤관 (Yun Gwan): A 고려 general famous for his campaigns against the Jurchen, unrelated to the 한반도. Now solve this. «Example 2» [QUESTION] {question} [CHOICES] {choices}

Table 8: Prompt for Knowledge Identification task.

Prompt for Knowledge Leveraging
You are a bilingual who is fluent in both Korean and English, and is knowledgeable about South Korea. You will be given a multiple choice question and a list of knowledge that are relevant to the question. The text are written in English-Korean code-switching, where matrix language is English and semantically important Korean words are embedded into English sentence. Your job is to follow the instructions and select one choice from [CHOICES]. Instructions: - Using given [KNOWLEDGE], explain concisely what and why you think is the answer. You can write in whichever language you can explain better, either Korean or English. Start this task with [EXPLANATION] tag. - Choose your final choice from [CHOICES]. The answer is one of the [CHOICES], so do not say 'none of the above'. You must write a index number in parentheses, like (1). Start this task with [ANSWER] tag. - Do not repeat input text in your response. Do not generate new question. Stick to input text that is given to you. I will give you an example for reference. «Example 1» [QUESTION] Read the following question and choose the most appropriate answer. Who is the person who greatly defeated the soldiers of the 당나라 in the 안시성 싸움? [CHOICES] (1) 양만춘 (2) 서희 (3) 김유신 (4) 강감찬 (5) 윤관 [KNOWLEDGE] 1. 안시성 싸움 (Siege of Ansi): 안시성 싸움 (645 AD) was a famous military conflict between 고구려 and the 당 Dynasty. 고구려, under the leadership of 양만춘 (Yang Man-chun), successfully defended the 안시성 against the powerful 당 forces led by Emperor 태종. 2. 양만춘 (Yang Man-chun): He was the general who commanded the defense of 안시성, playing a key role in defeating the 당나라 army. 3. 서희 (Seo Hee): A 고려 diplomat famous for negotiating with the 거란 to avoid invasion, but not involved in the 안시성 싸움. 4. 김유신 (Kim Yu-shin): A general from the 신라 Kingdom, instrumental in the unification of the 한반도, but not involved in this specific battle. 5. 강감찬 (Gang Gam-chan): A 고려 military commander known for his victory over the 거란 in the 귀주대첩, unrelated to 안시성. 6. 윤관 (Yun Gwan): A 고려 general famous for his campaigns against the Jurchen, unrelated to the 한반도. [EXPLANATION] The question specifically asks about the 안시성 싸움 (Siege of Ansi) and who defeated the 당나라 soldiers in that battle. Based on historical facts, the leader who played a key role in defending 안시성 and defeating the 당나라 army was 양만춘 (Yang Man-chun). [ANSWER] (1) Now solve this. «Example 2» [QUESTION] {question} [CHOICES] {choices} [KNOWLEDGE] {knowledge}

Table 9: Prompt for Knowledge Leveraging task.

Model		Economy	Geography	History	Law	Politics	Popular	Society	Tradition
GPT-4o	CS	85.00	20.00	40.00	30.00	30.00	05.00	50.00	35.00
	EN	80.00	00.00	05.00	05.00	10.00	00.00	05.00	00.00
	KOR	85.00	65.00	65.00	40.00	75.00	45.00	85.00	95.00
GPT-3.5	CS	70.00	00.00	00.00	20.00	10.00	05.00	10.00	10.00
	EN	75.00	00.00	00.00	10.00	15.00	0.00	05.00	00.00
	KOR	65.00	45.00	05.00	30.00	60.00	20.00	65.00	60.00
Llama3-70B	CS	20.00	00.00	00.00	10.00	15.00	10.00	10.00	00.00
	EN	30.00	05.00	00.00	10.00	20.00	05.00	10.00	00.00
	KOR	60.00	50.00	00.00	40.00	70.00	35.00	55.00	60.00
Llama3-8B	CS	20.00	00.00	00.00	05.00	25.00	00.00	05.00	00.00
	EN	15.00	00.00	00.00	05.00	15.00	00.00	00.00	00.00
	KOR	25.00	30.00	05.00	05.00	50.00	05.00	10.00	20.00

Table 10: QA performances on open-end QA.

Model		Economy	General	Geography	History	Law	Politics	Popular	Society	Tradition	Average
GPT-4o	CS	**93.22**	80.11	**69.75**	76.50	49.66	92.86	97.56	**65.51**	**81.98**	**78.57**
	EN	79.66	76.14	60.14	64.96	**51.49**	85.71	92.68	58.42	73.87	71.45
GPT-3.5	CS	**74.58**	37.50	39.15	30.13	32.41	**82.14**	**75.61**	**50.50**	**63.06**	**53.90**
	EN	69.49	**49.43**	**43.06**	**34.62**	**34.02**	73.81	65.85	47.03	55.41	52.52
Claude 3.5	CS	96.61	**78.41**	78.29	76.50	57.24	84.52	**92.68**	70.46	86.04	80.08
	EN	89.83	77.84	72.60	67.52	53.79	**89.29**	**92.68**	62.38	81.53	76.38
Solar	CS	83.05	**53.98**	**52.31**	**62.61**	**40.46**	**85.71**	**78.05**	**55.78**	**72.97**	**64.99**
	EN	**88.14**	**53.98**	47.69	37.61	39.08	76.19	70.73	51.65	65.32	58.93
Llama3 70B	CS	76.27	60.80	54.45	**48.29**	**40.46**	**86.90**	**82.93**	55.94	**71.17**	**64.13**
	EN	**79.66**	**61.36**	**55.52**	47.65	39.77	80.95	75.61	**56.11**	68.47	62.79
Llama3 8B	CS	**74.58**	37.50	39.15	30.13	32.41	**82.14**	**75.61**	**50.50**	**63.06**	**53.90**
	EN	69.49	**49.43**	**43.06**	**34.62**	**34.02**	73.81	65.85	47.03	55.41	52.52
Gemma2 27B	CS	77.97	51.70	48.04	**41.24**	36.78	**78.57**	**78.05**	**53.63**	**65.32**	**59.03**
	EN	**79.66**	**55.68**	**50.18**	36.11	**40.46**	69.05	75.61	52.15	61.26	57.80
Gemma2 9B	CS	**76.27**	**50.57**	44.84	**40.60**	35.63	**77.38**	**73.17**	**53.14**	**61.71**	**57.03**
	EN	67.80	44.32	**45.20**	35.68	**39.08**	70.24	65.85	49.50	**61.71**	53.26

Table 11: Knowledge leveraging performances of multilingual LLMs on CS and English settings. **Bold** indicates higher score between CS and English on each model. **Green** indicates the highest score from each domain.

Model		Economy	General	Geography	History	Law	Politics	Popular	Society	Tradition	Total
GPT-4o	CS	0/0	0/0	0/0	0/0	0/2	0/0	0/0	0/5	0/1	0/8
	EN	0/8	0/1	1/11	0/15	2/15	1/3	0/0	1/40	0/6	5/99
GPT-3.5	CS	0/0	**1/1**	1/1	**2/2**	0/0	0/0	0/0	**1/1**	0/0	**5/5**
	EN	0/0	**3/3**	**1/1**	**2/2**	**1/1**	0/0	0/0	**6/6**	0/0	**13/13**
Claude 3.5 Sonnet	CS	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
	EN	0/0	0/0	0/0	0/0	0/1	0/0	0/0	0/4	0/0	0/5
Solar	CS	0/0	**5/5**	**1/1**	2/2	**11/11**	0/0	0/0	**5/5**	**1/1**	**20/20**
	EN	0/0	**9/9**	**8/8**	**11/11**	**10/10**	0/0	**1/1**	**4/4**	**2/2**	**35/35**
Llama3 70B	CS	**2/2**	**6/6**	**6/6**	**11/11**	**3/3**	0/0	**1/1**	**3/3**	**2/2**	**34/34**
	EN	**2/2**	**7/7**	**10/10**	**24/24**	**12/12**	**2/2**	0/0	**4/4**	**2/2**	**63/63**
Llama3 8B	CS	**1/1**	**5/5**	**3/3**	**9/9**	**2/2**	**1/1**	**1/1**	0/0	**1/1**	**23/23**
	EN	0/0	**4/4**	**6/6**	**8/8**	**5/5**	**2/2**	0/0	**2/2**	**2/2**	**29/29**
Gemma2 27B	CS	**4/4**	**18/18**	**7/7**	26/28	22/38	**2/2**	2/3	6/9	10/14	97/123
	EN	**3/3**	**28/28**	**5/5**	37/38	7/15	**7/7**	**1/1**	13/20	**5/5**	106/122
Gemma2 9B	CS	**3/3**	**7/7**	**12/12**	**25/25**	18/19	**1/1**	**2/2**	**7/7**	**2/2**	77/78
	EN	**9/9**	**30/30**	**13/13**	**35/35**	**11/11**	**5/5**	0/0	23/28	**9/9**	135/140

Table 12: Counts of None errors. Each cell indicates # of None errors / # of errors due to knowledge hallucination. **Bold** indicates that all errors are caused by hallucination.

Example of Solar from the Tradition domain
[QUESTION] The following is a question about 한국의 문화 and 전통. Foreigners want to experience 한국의 전통 풍습 by celebrating the 정월 대보름. Which of the following is not part of the 체험내용? (The following is a question about Korean culture and traditions. Foreigners want to experience Korean traditional customs by celebrating the first full moon of the lunar year. Which of the following is not part of the experience?) (1) Eat 오곡밥 made from five types of grains. (Eat five-grain rice made from five types of grains.) (2) Drink 귀밝이술, wishing for better hearing. (Drink ear-brightening potion, wishing for better hearing.) (3) Eat food made with 쑥 and 익모초 to overcome the heat. (Eat food made with mugwort and dried hydrangea to overcome the heat.) (4) Munch on 견과류 like 땅콩, 호두, and 밤, hoping for no skin disease. (Munch on nuts like peanuts, walnuts, and chestnuts, hoping for no skin disease.) **[Knowledge from CS]** 1. 정월 대보름 (Jeongwol Daeboreum): It is a traditional Korean holiday that takes place on the first full moon of the lunar new year. It is a time for families to gather, participate in various rituals, and enjoy traditional foods. 2. 오곡밥 (Ogokbap): A traditional rice cake made with five grains (rice, barley, soybeans, sesame seeds, and black beans) that is eaten during 정월 대보름 for good health and longevity. 3. 귀밝이술 (Gwi-bal-i-sul): A traditional Korean drink made from makgeolli (rice wine) that is consumed on 정월 대보름 to improve hearing and communication skills. 4. 쑥과 익모초 (Perilla and Mugwort): These herbs are used in traditional Korean medicine and are believed to have healing properties. While they may be used in some traditional Korean dishes, they are not specifically associated with 정월 대보름. 5. 견과류 (Nuts): Eating nuts like 땅콩 (peanuts), 호두 (walnuts), and 밤 (chestnuts) is a common practice during 정월 대보름 as they are believed to bring good luck and prevent skin diseases. [ANSWER] (3) **[Knowledge from EN]** 1. The first full moon of the lunar year is celebrated in Korea as a traditional holiday called "Dano." 2. One of the customs during Dano is eating "ongchimsik," a rice dish made with five grains. 3. Another custom is drinking "suriseong," a herbal drink believed to improve hearing. 4. Eating "yeot," a dish made with mugwort and dried hydrangea, is also a traditional practice during Dano to help overcome the heat. 5. Nuts like peanuts, walnuts, and chestnuts are often consumed during Dano as a symbol of good health and fortune. 6. Skin disease is not directly related to the traditional customs of Dano. [ANSWER] (4) ...

Table 13: An example of knowledge activation by CS and English questions from the Tradition domain on Solar. The text in parentheses shows English translation.

Example of Gemma2 9B from the Law domain

[QUESTION]
What is the correct legal 판단 on the following case?
17세인 A, B, and C는 entered into a 계약 with 판매업자 D to purchase an expensive 노트북. A did not obtain 부모의 동의, while both B and C presented 부모의 동의서, but C forged the consent form. At the time of the contract, D did not know that A was a 미성년자. *Parents of A, B, and C are all 법정 대리인.
(What is the correct legal judgment on the following case?
17-year-olds A, B, and C entered into a contract with seller D to purchase an expensive laptop. A did not obtain parental consent, while both B and C presented parental consent forms, but C forged the consent form. At the time of the contract, D did not know that A was a minor. * Parents of A, B, and C are all legal representatives.)

(1) B's 부모 can cancel the 계약. (B's parents can cancel the contract.)
(2) The 부모 of A and C can both cancel the 계약. (The parents of A and C can both cancel the contract.)
(3) B cannot cancel the 계약 unlike A. (B cannot cancel the contract unlike A.)
(4) D cannot withdraw the 의사 표시 to conclude a 계약 with A. (D cannot withdraw the expression of intention to conclude a contract with A.)
(5) D can demand a prompt response from A regarding the 취소 of the 계약. (D can demand a prompt response from A regarding the cancellation of the contract.)

[Knowledge from CS]
1. **미성년자 계약의 효력**: 대한민국 민법상 14세 미만의 자는 계약력이 없으며, 14세 이상 19세 미만의 자는 부모의 동의가 없는 계약은 효력이 없다.
2. **부모의 동의**: 미성년자의 계약은 부모의 동의가 있어야 효력이 있다. 동의서가 위조된 경우, 계약은 무효이다.
3. **계약의 취소**: 계약의 당사자는 계약을 취소할 수 있다.
4. **의사표시의 취소**: 의사표시는 당사자의 의사에 따라 취소될 수 있다.

[ANSWER] (Correct answer is (3).)
(2)

[Knowledge from EN]
1. In most jurisdictions, minors (typically under 18 years old) lack the legal capacity to enter into binding contracts.
2. Contracts entered into by minors are generally voidable at the option of the minor.
3. Parental consent can sometimes affect the validity of a contract entered into by a minor, depending on the specific laws of the jurisdiction.
4. Fraudulent misrepresentation, such as forging a parental consent form, can invalidate a contract.

[ANSWER]
(2)

Table 14: An example of knowledge activation by CS and English questions from the Law domain on Gemma2 9B. The text in parentheses shows English translation.