EMCEE: Improving Multilingual Capability of LLMs via Bridging Knowledge and Reasoning with Extracted Synthetic Multilingual Context

Hamin Koo **Jaehyung Kim**
Yonsei University
{hamin2065,jaehyungk}@yonsei.ac.kr

Abstract

Large Language Models (LLMs) have achieved impressive progress across a wide range of tasks, yet their heavy reliance on English-centric training data leads to significant performance degradation in non-English languages. While existing multilingual prompting methods emphasize reformulating queries into English or enhancing reasoning capabilities, they often fail to incorporate the language- and culture-specific grounding that is essential for some queries. To address this limitation, we propose EMCEE (**E**xtracting synthetic **M**ultilingual **C**ont**e**xt and m**e**rging), a simple yet effective framework that enhances the multilingual capabilities of LLMs by explicitly extracting and utilizing query-relevant knowledge from the LLM itself. In particular, EMCEE first extracts synthetic context to uncover latent, language-specific knowledge encoded within the LLM, and then dynamically merges this contextual insight with reasoning-oriented outputs through a judgment-based selection mechanism. Extensive experiments on four multilingual benchmarks covering diverse languages and tasks demonstrate that EMCEE consistently outperforms prior approaches, achieving an average relative improvement of 16.4% overall and 31.7% in low-resource languages.[1]

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities across a wide range of applications, from natural language understanding to generation (Brown et al., 2020; Touvron et al., 2023; Team et al., 2024). Among the numerous factors contributing to LLM performance, the quality and quantity of pre-training data stand out as particularly critical (Kaplan et al., 2020; Hoffmann et al., 2022; Gunasekar et al., 2023). However, given that most pre-training data are in English (Ouyang et al., 2022; Touvron et al., 2023), LLMs inevitably exhibit an English-centric bias. As a result, their performance often deteriorates when handling non-English queries (Shi et al., 2022). As the use of LLMs in the real-world rapidly increases, this disparity highlights the urgent need for developing LLMs or methods that can perform effectively across diverse languages, especially those with limited resources.

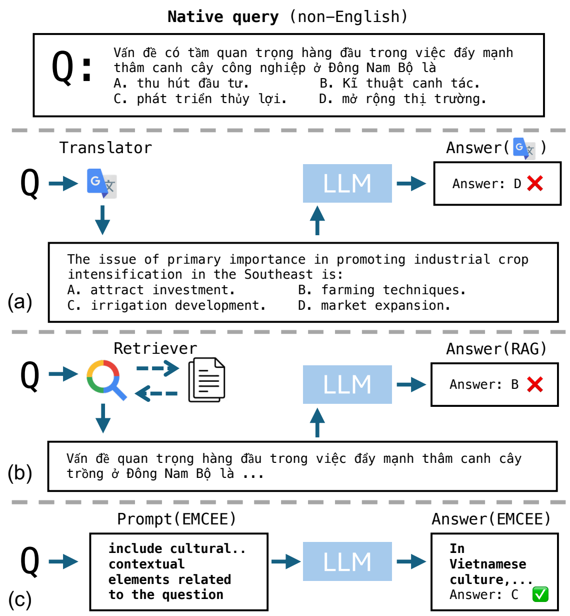


Figure 1: **Different multilingual prompting methods.** Given a Vietnamese query from social science category in M3-Exam (Zhang et al., 2023), **(a)** Translating the query into English using the external translator results in an incorrect answer. **(b)** Even with retrieval-augmented generation (Google Custom Search API), the model remains incorrect. **(c)** However, with EMCEE (Ours) prompt that extracts relevant context from LLM itself, the LLM successfully produces the correct answer.

Addressing multilingual queries with LLMs is particularly challenging because they often involve heterogeneous requirements that demand different strategies. For example, queries in domains such as mathematics and natural science primarily depend on reasoning skills, which are less sensitive to the query language itself. Existing multilingual

[1] Code: https://github.com/hamin2065/EMCEE

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2800–2822

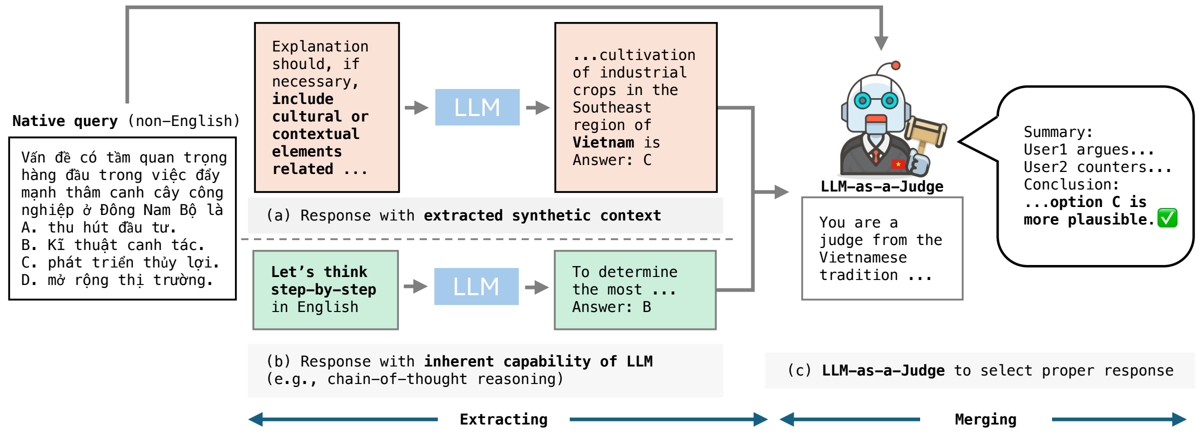


Figure 2: **Overview of EMCEE.** (a) LLM receives a non-English (native) query along with an instruction to *extract relevant synthetic context*, producing a context-enriched response. (b) In parallel, LLM generates a *reasoning-focused response* using only its inherent knowledge, without additional context. (c) An *LLM-as-a-Judge* module then compares the two responses and selects the final answer based on contextual relevance and reasoning adequacy.

prompting approaches have mostly targeted this case by translating non-English queries into English (Liu et al., 2024) or by inducing intermediate reasoning in English (Huang et al., 2023). On the other hand, queries in domains such as language or social science rely heavily on knowledge grounded in specific countries or cultures; in such cases, incorporating appropriate knowledge is more crucial than applying abstract reasoning. However, this remains a significant challenge, as LLMs often struggle to acquire and utilize language- and culture-specific information due to the limited representation of such non-English data in pre-training corpora; yet, how to effectively handle such knowledge-intensive multilingual queries remains under-explored. Furthermore, since both reasoning-oriented and knowledge-intensive queries are often mixed in the real-world, an effective system should be able to answer both cases simultaneously, introducing additional complexity.

Contribution. To fill this gap, we propose a new prompting-based framework called EMCEE (**E**xtracting synthetic **M**ultilingual **C**ont**e**xt and m**e**rging), which enables LLMs to better handle heterogeneous multilingual queries. Specifically, EMCEE consists of two key components: (1) an *explicit extraction step* that generates synthetic context to surface query-relevant knowledge from the LLM, and (2) a *merging step* that adaptively integrates this knowledge with reasoning. The extraction step is motivated by Sun et al. (2023), which demonstrates that explicitly eliciting hidden knowledge into textual form is often more effective than relying on the model to recall it implicitly. Building on this insight, EMCEE enriches the context for answering multilingual queries by incorporating synthetic data, which is especially crucial for knowledge-intensive cases. Unlike retrieval-augmented generation, this approach neither depends on external retrievers nor databases; instead, our framework fully leverages the LLM's own internal knowledge representation to generate synthetic, language-specific context. Simultaneously, EMCEE derives an alternative reasoning-focused response, which tends to be more effective for reasoning-oriented queries. Finally, EMCEE employs an LLM-as-a-Judge (Zheng et al., 2023) to merge the LLM's reasoning across the two candidate responses, enabling the system to determine whether a knowledge-grounded or reasoning-oriented path is more suitable for each query.

We show the effectiveness of EMCEE through extensive evaluations on four multilingual datasets. These datasets cover various languages and tasks such as multiple-choice question answering, text generation, language understanding, and reasoning. For example, compared to a method that uses native query with native instructions, EMCEE achieves an average relative improvement of 16.4% across all datasets, with a notable improvement of 31.7% in low-resource languages. These results suggest that EMCEE effectively captures and leverages cultural knowledge, addressing a key limitation of prior work. In addition, we find that EMCEE is compatible with various state-of-the-art LLMs, including API-based and Open-source ones. Overall, these findings underscore the importance of cultural knowledge extraction for non-English queries and the need for continued research in multilingual prompting that incorporates cultural awareness.

2 Enhancing Multilingual Capability of LLMs with Extracted Synthetic Context

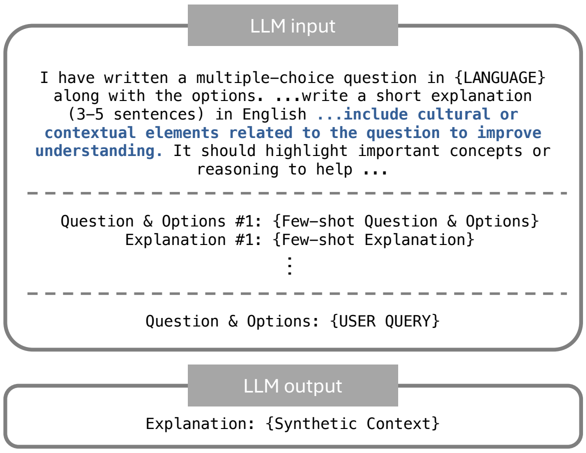
2.1 Preliminary

We consider the problem of generating an accurate response for a non-English input query, referred to as a *native query*. As most LLMs are trained primarily on English data, their performance on non-English queries is often limited (Shi et al., 2022). To mitigate this, *multilingual prompting* aims to reformulate the input prompt so that it better aligns with the model's inherent knowledge. For instance, Liu et al. (2024) translates the native query into English, and Huang et al. (2023) further proposes to construct a mixed input prompt composed of native queries and English instructions that induce chain-of-thought (CoT) reasoning. However, these methods are limited in capturing language- and culture-specific contexts, including social conventions and background knowledge embedded in the native language, that is critical for knowledge-intensive queries.

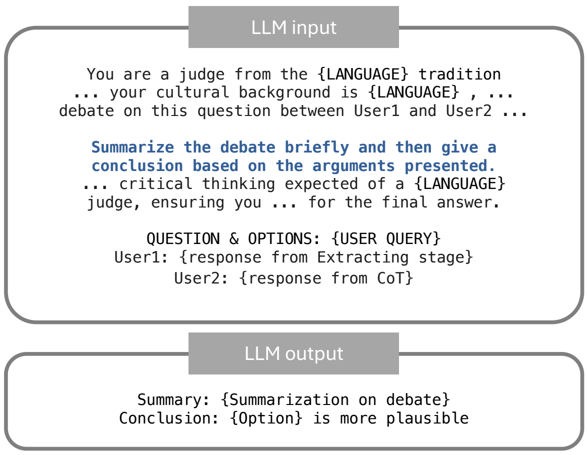
2.2 EMCEE: Extract Synthetic Multilingual Context and Merge with Reasoning

To address these limitations, our key idea is to extract query-relevant information from the LLM itself and use it as an additional context to improve the response quality. As illustrated in Figure 2, EMCEE follows a two-stage framework designed to enhance the model's ability to answer native queries by incorporating relevant contextual knowledge. The first stage, *extraction*, elicits from the LLM the implicit background knowledge required to answer the query. The second stage, *merging*, integrates this knowledge with the model's reasoning to produce a more accurate final response.

Extracting relevant context from LLM. The goal of this stage is to extract query-relevant knowledge from the LLM in the form of *synthetic context*, supplementing information that the model often overlooks in direct query answering. To this end, we prompt the LLM to generate background knowledge necessary for answering the native query. In line with prior work, studies have shown that explicitly eliciting knowledge encoded in model parameters can improve downstream performance across diverse tasks (Sun et al., 2023; Wang et al., 2023a). Following this perspective, our approach exploits the internal knowledge embedded within the LLM's parameters, without relying on external retrieval or databases. Concretely, we design an instruction template that guides the model to extract relevant information, typically within 3–5 sentences (see Figure 3 (a)). The extracted context is then concatenated with the native query to create an augmented input, enabling the LLM to generate responses that are better informed and contextually grounded. To further improve the quality of the extracted context, we employ a few-shot prompting setup with exemplars demonstrating how to extract essential background knowledge. This overall setup enables the model to surface knowledge spanning cultural, historical, and domain-specific dimensions. Additionally, to leverage LLMs' well-established ability to follow English instructions, we conduct this extraction step with English instructions (see Appendix B.4 for more details).



(a) Prompt and LLM response for extracting



(b) Prompt and LLM response for merging

Figure 3: **Input prompt and LLM response example.** These figures illustrates the overview of input prompt and response of LLM during (a) Extracting and (b) Merging processes, respectively.

Merging synthetic context with reasoning. While extracting synthetic context enables the LLM to better handle knowledge-intensive multilingual queries, the degree of reliance on such knowledge varies widely across tasks. For instance, in the M3-Exam dataset (Zhang et al., 2023), questions from social science or language subjects benefit from contextual grounding, whereas those from mathematics or natural-science primarily rely on logical reasoning. To accommodate this diversity within a unified framework, we generate an additional reasoning-focused response and use it with the context-enriched one to obtain the final output.

Specifically, we adopt CoT prompting with English instruction to obtain an additional response focusing on the LLM's inherent reasoning capabilities. We then employ an LLM-as-a-Judge framework (Zheng et al., 2023) to select the most appropriate response, as illustrated in Figure 2. The judge model is instructed to consider the linguistic and contextual background of the query, such as local customs, social norms, or discipline-specific reasoning patterns, thereby selecting the response that best fits the query's context (see Figure 3 (b)). By selectively combining contextual and reasoning-oriented responses, EMCEE merges the LLM's reasoning to evaluate and choose the response best suited to each query, ensuring robust performance across diverse multilingual settings. While it is also possible to prompt the LLM to choose a strategy from the query alone, we observe that our judgment-based merging—where the LLM compares and reasons over the generated responses—yields better performance, likely because it avoids the additional decision-stage errors (see Table 1).

3 Experiments

3.1 Setups

Datasets. We evaluate the effectiveness of EMCEE across four tasks: multiple-choice question answering (MCQA), text generation, language understanding, and reasoning. For each task, we utilize the following benchmarks to ensure a comprehensive, multilingual evaluation: (1) *M3-Exam* (Zhang et al., 2023) is an exam-style dataset collected from diverse countries. We exclude image-based questions and use only text data to assess culturally and linguistically varied test items. In order to conduct a fair evaluation, we remove every question that contains background knowledge and keep only those that can be answered directly from the prompt. (2) *MKQA* (Longpre et al., 2021) is a generation-based evaluation benchmark for open-domain question answering across 26 languages. Following Huang et al. (2023), we select 9 languages. (3) *XNLI* (Conneau et al., 2018) evaluates logical entailment understanding across 14 languages. (4) *XCOPA* (Ponti et al., 2020) evaluates commonsense reasoning in 11 low-resource languages. Across all datasets, English is excluded to prioritize multilingual generalization. To ensure a balanced cross-lingual analysis, we randomly sample 100 examples per language for MKQA, XNLI, and XCOPA, while using the entire filtered M3-Exam. Overall, the combined datasets cover a span of 24 languages. More details about the experimental datasets are provided in Appendix A.1.

Metrics. We use accuracy for M3-Exam, XNLI, and XCOPA since these tasks involve selecting a single correct answer from predefined options. For MKQA, we use a span-based F1 score, which measures the token-level overlap between the predicted answer and the ground truth. In addition, to assess model performance comprehensively, we report our results in three different ways: (1) averaging across all languages (*All*), (2) using only high-resource languages (*High*), and (3) using only low-resource languages (*Low*). Since no exact criteria exist for defining high- and low-resource languages, we use a dataset-specific empirical approach; specifically, we compute the average performance on native-language queries for each dataset and use this as a threshold, *e.g.*, where languages scoring above the threshold are considered high-resource.

Baselines. Following previous works (Huang et al., 2023; Liu et al., 2024), we consider baselines that account for variations in instruction language, chain-of-thought (CoT) reasoning, and translation. We adopt a zero-shot approach for all methods, given that our chosen models are already fine-tuned to follow instructions. (1) NATIVE-BASIC: All instructions and questions are written in the native language without reasoning steps. This setup measures performance entirely within the native language context. (2) ENG-BASIC: This setting is derived from NATIVE-BASIC, except that the instructions are provided in English while the questions remain in the native language. This highlights how partial English usage can affect performance. (3) XLT (Huang et al., 2023): XLT is a prompting method that improves the multilingual capability of LLMs. It instructs the model to translate the question into English and subsequently solve the prob-

Table 1: **Main result.** Test performance of GPT-4o-mini over different multilingual prompting methods on four different tasks. Results are measured using accuracy for M3-Exam, XNLI, and XCOPA. F1-score is used for MKQA. The best and second best scores are highlighted in **bold** and <u>underline</u>, respectively. *All*, *High*, and *Low* categories are determined based on performance of NATIVE-BASIC: the top half of languages are labeled high-resource, and the bottom half low-resource. The numbers in parentheses indicate the number of evaluated languages.

Datasets (→) Methods (↓)	M3-Exam All (8)	M3-Exam High (4)	M3-Exam Low (4)	MKQA All (9)	MKQA High (5)	MKQA Low (4)	XNLI All (14)	XNLI High (7)	XNLI Low (7)	XCOPA All (11)	XCOPA High (6)	XCOPA Low (5)
NATIVE-BASIC	65.2	72.7	57.7	44.1	48.6	38.5	66.2	74.0	58.4	79.3	94.2	61.4
ENG-BASIC	64.5	72.7	56.4	48.1	49.8	46.0	71.4	<u>74.6</u>	68.1	55.3	57.3	52.8
TRANS-GOOGLE	59.9	70.1	49.8	50.3	50.4	50.2	70.5	73.4	67.6	66.0	65.2	67.3
XLT	70.4	76.9	63.8	<u>51.1</u>	50.9	<u>51.5</u>	72.6	74.3	71.0	<u>91.1</u>	95.8	<u>85.4</u>
NATIVE-COT	70.4	76.5	64.4	45.5	49.2	40.9	68.4	70.1	66.6	90.0	<u>96.3</u>	82.4
ENG-COT	74.6	81.8	67.3	49.4	49.5	49.3	<u>73.2</u>	73.7	<u>72.7</u>	90.5	96.0	83.8
RAG (NATIVE)	68.7	80.4	57.0	42.5	45.3	39.0	67.4	69.9	64.9	83.6	92.3	73.2
RAG (ENG)	72.1	**84.0**	63.9	44.7	44.9	44.5	70.4	71.1	69.7	87.9	94.0	80.6
EMCEE (ROUTE)	<u>76.2</u>	83.1	<u>69.2</u>	50.8	<u>51.6</u>	49.8	73.1	73.9	72.3	90.5	96.0	83.8
EMCEE (Ours)	**77.4**	<u>83.3</u>	**71.5**	**52.3**	**52.2**	**52.4**	**74.3**	**74.7**	**73.9**	**92.0**	**96.8**	**86.2**

lem step-by-step in English. (4) TRANS-GOOGLE (Liu et al., 2024): All instructions and questions are translated (via Google Translate API) into English before being processed by the model. This baseline shows how external machine translation impacts overall performance. (5) NATIVE-COT: Both instructions and CoT reasoning are provided in the native language, allowing the model to "think" in that language. (6) ENG-COT: Both instructions and CoT reasoning are carried out in English, assessing how full English prompts with detailed reasoning steps influence results. (7) RAG (NATIVE): This setting extends NATIVE-BASIC by appending retrieved passages (in the native language) obtained via the Google Custom Search API as an external context. (8) RAG (ENG): This variant combines the ENG-BASIC setup with retrieval-augmented generation, where English instructions are used alongside retrieved passages. Specific prompt templates of these settings are presented in Table 17.

Implementation details. We employ GPT-4o-mini (`gpt-4o-mini-2024-07-18`) (OpenAI, 2024). To mitigate the influence of random sampling, we strictly set the temperature to 0.0 for API-based models and use greedy decoding for Llama models. See more details in Appendix A.3.

3.2 Main results

Table 1 presents the results of our experiments on four multilingual benchmarks using GPT-4o-mini. Overall, EMCEE consistently surpasses all non-RAG baselines across datasets, achieving a 16.4% average relative improvement over the basic prompting method, NATIVE-BASIC, and further improving upon non-RAG baselines by 3.2% (ENG-COT) and 3.9% (XLT). Notably, EMCEE shows the largest gains in low-resource languages—a result especially important for multilingual modeling. Compared to NATIVE-BASIC, it improves by 23.7% on M3-Exam, 36.1% on MKQA, 27.7% on XNLI, and 40.4% on XCOPA, averaging a 32.0% gain. EMCEE also exceeds ENG-COT and XLT by 4.0% and 4.4%, respectively. In addition, Figure 4 further confirms that our framework delivers substantial boosts for low-resource languages.

We also compare against retrieval-augmented generation (RAG) methods (RAG-NATIVE and RAG-ENG) that utilize external context, which is different from how EMCEE uses synthetic context. Although these RAG methods offer minor gains over simpler baselines, their performance remains lower than that of reasoning-oriented approaches such as ENG-COT and XLT, and far behind EMCEE in most cases. For example, on M3-Exam, RAG-ENG reaches 72.1% compared to 74.6% for ENG-COT and 77.4% for EMCEE, with similar patterns across MKQA, XNLI, and XCOPA. This suggests that external retrieval offers limited benefit, while EMCEE's extraction of *synthetic multilingual context* from the model's own knowledge yields richer, query-aligned context. This highlights that EMCEE fully exploits the internal knowledge and reasoning capabilities of the underlying LLM itself, without relying on any external retrieval modules or auxiliary corpora, thereby making it a purely self-contained prompting framework.

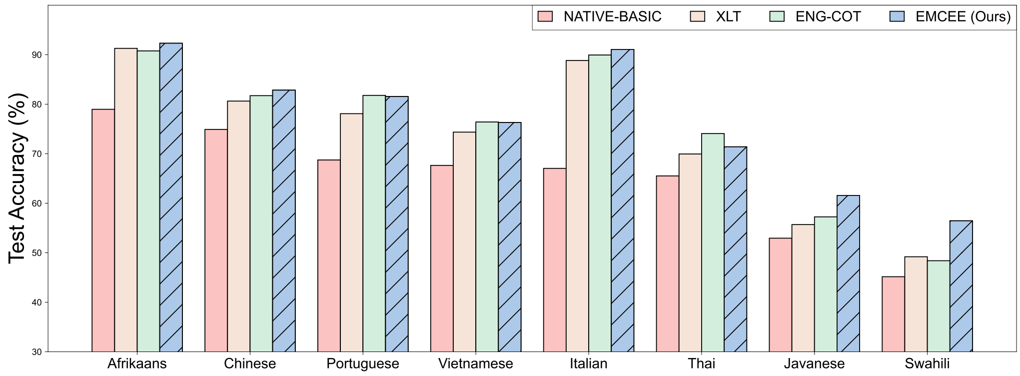


Figure 4: **Overall language-wise improvement.** Test accuracy of GPT-4o-mini over four different multilingual prompting methods on M3-Exam. More results on other datasets and LLMs are presented in Appendix B.

Table 2: **Ablation study of EMCEE.** Test accuracy of GPT-4o-mini on M3-Exam with different configurations of the proposed components in EMCEE. CoT, ExT, MeR are abbreviations of Chain-of-Thought, ExTracting, and MeRging, respectively.

Methods	CoT	ExT	MeR	All / High / Low
NATIVE-BASIC	✗	✗	✗	65.2 / 72.7 / 57.7
ENG-COT	✓	✗	✗	74.6 / 81.8 / 67.3
	✗	✓	✗	74.7 / 82.0 / 67.5
	✓	✗	✓	75.2 / 83.4 / 67.1
EMCEE	✓	✓	✓	77.4 / 83.3 / 71.5

Lastly, we further include a routing variant, EMCEE (Route), which lets the LLM select between extraction and reasoning pathways based on language characteristics of the query alone instead of our merging step. While it achieves strong results overall, it occasionally underperforms the unified EMCEE due to routing errors or suboptimal pathway selection. This implies that merging the LLM's reasoning over both extraction and reasoning strategies is more robust than deciding on a single strategy based solely on the query.

3.3 More analyses

In this section, we conduct additional analyses of EMCEE on the M3-Exam dataset, with GPT-4o-mini as the default LLM unless otherwise specified.

Ablation study. We conduct an ablation study to evaluate the contributions of the three core components of EMCEE: (1) Chain-of-Thought (CoT) reasoning, (2) synthetic knowledge extraction (ExT), and (3) our merging (MeR) mechanism guided by an LLM-as-a-Judge framework. The results are summarized in Table 2. First, by comparing the results on the first and third rows, one can verify that using the extracted knowledge as an additional input context leads to improved performance. Next, in the fourth row, we consider a hybrid setup for the ablation. We generate an additional CoT response using a positive temperature and apply the same EMCEE merging step between this sampled response and the original CoT-generated response. In other words, the response based on extracted information is substituted with another CoT-generated response. As shown in the improvement on low-resource languages (67.1% → 71.5%) between the fourth and fifth rows, the impact of merging two responses from different knowledge sources is evident.

Table 3: **Compatibility across various LLMs.** Average accuracy of three different LLMs on M3-Exam for all languages (All). The full results are presented in Table 14. The best and second best scores are highlighted in **bold** and <u>underline</u>, respectively.

Methods	GPT-4o	Claude-Haiku	Llama-3.1-8B
NATIVE-BASIC	78.1	67.4	49.8
TRANS-GOOGLE	68.4	60.1	<u>53.9</u>
XLT	81.8	74.0	25.8
NATIVE-COT	<u>83.9</u>	74.8	36.8
ENG-COT	82.4	<u>75.4</u>	40.4
EMCEE (Ours)	**85.7**	**75.6**	**56.9**

Compatibility across LLMs. We perform additional experiments to showcase the efficacy of EMCEE across various LLMs. In particular, we evaluate GPT-4o (`gpt-4o-2024-08-06`), Claude-3.5-Haiku (`claude-3-5-haiku-20241022`) (Anthropic, 2024), and Llama-3.1-8B-Instruct (Dubey et al., 2024) on M3-Exam with results summarized in Table 3. Compared to other baselines, EMCEE consistently achieves the highest performance. Notably, while results on Llama-3.1-8B-Instruct show that XLT, NATIVE-COT, and ENG-COT fail to outperform NATIVE-BASIC, our method maintains the highest scores. Overall, in GPT-4o, our method achieves 8.9% relative improvement over NATIVE-

(1) Native query	(2) Translated query
Pagupon Iku Jenenge Kandhang ... A. Pitik B. Macan C. Jaran D. Dara	Pagupon is the name of the cage for ... A. Chicken B. Tiger C. Horse D. Dove

(3) Incorrect responses by ENG-COT: 1. A. Pitik ... "chicken" in Javanese ... referred to as a "kandhang pitik." ... 4. D. Dara ... means "dove" in Javanese ... referred to as a "kandhang dara." ... **the most appropriate answer would be related to domesticated animals. The correct answer is A. Pitik, as it directly refers to a barn for chickens.**

(4) Correct response by Extracting: ... In Javanese, the term "kandhang" refers to a cage or pen for animals. Among the options provided, "Dara" refers to a pigeon or dove in Javanese. **Therefore, the correct answer is D, Dara, as it directly relates to the type of animal that would be housed in a "kandhang" specifically designed for pigeons.** The other options refer to different animals that would have their own respective enclosures.

(5) EMCEE's response by Merging (3) & (4) : User1 argues that the question is about the name of a barn for domesticated animals, specifically identifying "kandhang pitik" for chickens as the most appropriate answer. User2 counters that "pagupon" specifically refers to pigeons, and therefore "kandhang dara" is the correct term for a pigeon shelter. **Conclusion: Given that "pagupon" directly translates to "pigeon" in Javanese, and the question specifically asks for the name of the shelter for that animal, [option] D is more plausible.**

Figure 5: **Qualitative example from M3-Exam.** This figure shows: (1) native Javanese questions and options, (2) the translated one by GPT-4o, (3) incorrect responses produced by ENG-COT, (4) correct responses obtained through our cultural knowledge extraction, and (5) an EMCEE response that integrates and compares the outputs from (3) and (4). More qualitative examples are provided in Appendix C.

Table 4: **Test accuracy on GlobalOpinionQA.** Best and second best scores are highlighted in **bold** and <u>underline</u>, respectively.

Methods	All	High	Low
NATIVE-BASIC	65.3	76.8	53.7
ENG-BASIC	<u>68.6</u>	<u>78.9</u>	<u>58.3</u>
TRANS-GOOGLE	68.5	**79.7**	57.4
XLT	55.2	66.0	44.5
NATIVE-COT	63.2	73.9	52.4
ENG-COT	65.4	74.9	53.9
EMCEE (Ours)	**69.0**	77.6	**60.4**

BASIC on average, while the improvements were 10.8% for Claude and 12.5% for Llama. These findings indicate the effectiveness of our approach across diverse architectures and resource conditions. Full results are provided in Table 14.

Cultural alignment on subjective QA. Here, we further investigate whether EMCEE is capable of steering LLM responses to align with specific groups of users, which has become increasingly important as LLMs are applied to subjective tasks. For this evaluation, we use GlobalOpinionQA (Durmus et al., 2024)—a dataset comprising questions and answers from cross-national surveys designed to capture diverse global opinions. Since the dataset was originally created in English, we employ GPT-4o to translate it into our target languages and filter it to include only countries with more than 40 samples. Our filtered dataset covers data from over 30 countries (see Appendix A.1 for full details), totaling 2,609 samples. We configure the LLM-as-a-Judge by *country* rather than by *language*, since the dataset contains overlapping languages (*e.g.*, Spanish for both Spain and Peru), allowing us to distinguish between culturally distinct regions. As shown in Table 4, EMCEE achieves the highest overall accuracy and particularly strong performance in low-resource countries, highlighting its robustness in culturally diverse contexts. Although TRANS-GOOGLE slightly outperforms EMCEE in high-resource settings—probably because the data were originally in English—these results confirm the effectiveness of EMCEE in steering LLM outputs toward region-specific perspectives by leveraging extracted synthetic knowledge.

Table 5: **Results on Aya-8B.** Accuracy on four M3-Exam languages that are supported by multilingual model Aya-8B. Best scores are highlighted in **bold**.

Methods	zh	it	pt	vi	Avg
NATIVE-BASIC	39	45	52	48	46.0
ENG-BASIC	31	**46**	52	26	38.8
EMCEE	**56**	40	**54**	**49**	**49.8**

Evaluation on Aya-8B. To further verify the generality of our findings, we conduct additional experiments using AYA-8B (Üstün et al., 2024), a model specifically optimized for multilingual understanding. We focus on four languages that are covered by both AYA-8B and the M3-Exam dataset—Chinese (zh), Italian (it), Portuguese (pt), and Vietnamese (vi)—and sample 100 queries for each language. As shown in Table 5, EMCEE improves the average performance of Aya-8B from 46.0 and 38.8 to 49.8. Gains are observed in Chinese, Portuguese, and Vietnamese, while Italian remains challenging, suggesting that EMCEE remains beneficial on average even for multilingual-specialized models.

Table 6: **Results on stronger LLMs.** Accuracy on a subset of M3-Exam (100 samples per language) for GPT-5 and Claude-4-Sonnet. Best and second best scores are highlighted in **bold** and <u>underline</u>, respectively.

	GPT-5			Claude-4-Sonnet		
Methods	All	High	Low	All	High	Low
NATIVE-BASIC	<u>74.3</u>	<u>83.8</u>	**64.8**	79.3	**90.0**	68.5
ENG-COT	68.4	81.3	55.5	<u>81.5</u>	<u>89.5</u>	<u>73.5</u>
XLT	58.4	70.0	46.8	**82.3**	<u>89.5</u>	**75.0**
EMCEE (OURS)	**76.0**	**87.5**	<u>64.5</u>	**82.3**	<u>89.5</u>	**75.0**

Table 7: **Cost analysis.** Average test accuracy, input and output token usages, and incurred cost ($) of GPT-4o-mini on M3-Exam with other cost-intensive baseline and EMCEE. The better scores are highlighted in **bold**.

Methods	Acc ↑	Input / Output Tokens ↓	Cost ↓
3 × ENG-COT & MERGE	76.9	**282k** / 176k	$0.149
EMCEE (OURS)	**78.8**	539k / **99k**	**$0.140**

Generalization to stronger LLMs. We further evaluate EMCEE on GPT-5 and Claude-4-Sonnet using a subset of the M3-Exam dataset (100 per language). As shown in Table 6, EMCEE achieves the best overall accuracy on GPT-5 (76.0), outperforming all baselines and showing strong gains on high-resource languages (87.5). On Claude-4-Sonnet, EMCEE attains 82.3, matching the strongest baseline (XLT) while maintaining robust performance across both high- and low-resource settings. Overall, these results indicate that EMCEE generalizes well to stronger LLMs and consistently provides stable improvements over both multilingual and English-based reasoning baselines.

Cost analysis. As the proposed EMCEE is based on multi-step inference of LLMs, one might be concerned about its computational cost with higher token usage. To demonstrate that the gain from EMCEE is not merely a result of increased computation, we consider other cost-intensive approaches. Specifically, we consider a three-sample variant of the ENG-COT baseline, which enhances reasoning diversity by generating multiple chain-of-thought paths and submitting all of them to the LLM-as-a-Judge module (analogous to the fourth row in Table 2). Although this approach naturally improves accuracy by exploring more reasoning paths, it also substantially increases computational cost. As shown in Table 7, EMCEE attains higher accuracy (78.8%) at a lower total cost ($0.140), even with greater input token usage,

Table 8: **Generalization to reasoning LLMs.** Average test accuracy on M3-Exam using Qwen3-8B with and without think-mode. The best and second best scores are highlighted in **bold** and <u>underline</u>, respectively.

Methods	All	High	Low
w/o Think			
NATIVE-BASIC	37.8	44.6	31.0
ENG-BASIC	45.6	49.2	42.0
EMCEE (OURS)	**67.3**	**80.0**	**54.7**
w/ Think			
NATIVE-BASIC	49.5	67.1	31.8
ENG-BASIC	45.6	49.2	42.0
TRANS-GOOGLE	49.3	65.4	33.1
EMCEE (OURS)	<u>65.0</u>	<u>79.4</u>	<u>50.7</u>

indicating that its extraction and merging steps effectively focus on task-relevant information and minimize redundant reasoning. Since end-to-end latency may vary across hardware and network configurations, we report token consumption as a consistent measure of inference efficiency.

Additional experiment with thinking models. We further evaluate our framework using the recent Qwen3-8B model (Team, 2024), which provides an optional "think-mode". We conduct experiments on the full M3-Exam dataset, comparing multiple baselines and our EMCEE pipeline under both think-mode settings. As shown in Table 8, enabling the think-mode substantially benefits native prompting (from 37.8% to 49.5%), suggesting that internal reasoning processes can be helpful in some scenarios. However, this improvement is not consistent across settings, as ENG-BASIC exhibits identical performance (45.6%) regardless of whether think-mode is enabled, indicating the limited generalizability of think-mode under prompt-based reasoning. In contrast, EMCEE consistently enhances accuracy in both configurations (67.3% without think-mode and 65.0% with think-mode). This shows that the extraction component, rather than implicit reasoning augmentation of think-mode, is the main contributor to EMCEE's effectiveness.

Qualitative examples. As shown in Figure 5, we provide a qualitative example to support further analysis. (1) The original question is written in Javanese, and (2) the translated version indicates that it is asking about the specific name for the enclosure in which "Pagupon" is kept. (3) The response generated by the Eng-CoT method involves interpreting each option by explaining its meaning in English. It then concludes that the question is related to domesticated animals, the incorrect selec-

(1) Native query: ウキウキ・ウェイク・ミー・アップは誰の歌ですか。 (Whose song is 'Wake Me Up Before You Go-Go'?)

(2) Correct Answer: ワム! (Wham!)

(3) Extracted information: The song "ウキウキ・ウェイク・ミー・アップ" (Ukiuki Wake Me Up) is performed by the Japanese artist Koda Kumi...She is a prominent figure in the J-Pop scene ...

Figure 6: **Failure mode on MKQA.** The extraction step injects irrelevant language-specific context into a query whose correct answer is Wham!, illustrating how unnecessary contextualization can mislead the model.

tion of option A. Although one might superficially associate "cage" with "chicken," this interpretation is culturally and contextually incorrect in Javanese, where the term "Pagupon" is mainly used for pigeons or doves. (4) In contrast, our task-relevant knowledge extraction process correctly identifies D as the answer. (5) Finally, the LLM-as-a-Judge confirms that the answer is D instead of A.

Failure mode: irrelevant contextualization. While the extraction step is designed to provide helpful cultural and contextual grounding, it can sometimes introduce irrelevant context when the query concerns a globally identifiable entity rather than culture-specific knowledge. As shown in Figure 6, the query asks, "Whose song is 'Wake Me Up Before You Go-Go'?" While the correct answer is "Wham!", the extracted explanation regarding Japanese cultural knowledge incorrectly identifies the song as performed by the Japanese artist Koda Kumi, introducing irrelevant local cultural information. This mismatch illustrates that the extraction step can misinterpret the query's intent, especially when the query refers to a globally recognized entity rather than a culturally specific one.

4 Related Works

Multilingual prompting of LLMs. Recent advances in multilingual prompting have shown that translation-based strategies can substantially improve LLM performance on non-English tasks. XLT (Huang et al., 2023) proposes cross-lingual thought prompting, which translates non-English inputs into English for step-by-step reasoning, yielding strong multilingual performance. Similarly, Liu et al. (2024) shows that simple translation pipelines can yield competitive outcomes, although they may lose language-specific nuances. Despite these gains, translation-based approaches often fail to capture language-specific nuances, whereas our framework instead directly synthesizes task-relevant contextual signals from the model.

Synthetic data and context extraction using LLMs. LLMs have recently been explored as sources of synthetic supervision, where models generate their own auxiliary data or contextual signals to enhance downstream performance. Recitation (Sun et al., 2023) introduces a self-contained prompting strategy that allows LLMs to generate task-aligned knowledge without an external retrieval. Query2Doc (Wang et al., 2023a) expands queries into pseudo-documents, improving information grounding for open-domain and QA tasks. Other works explore self-generated supervision or adaptive retrieval using model-generated signals (Wang et al., 2023b; Li et al., 2024).

Response merging and self-evaluation with LLMs. Recent research increasingly explores ways to integrate multiple LLM-generated responses through structured synthesis or self-assessment. Multi-agent debate approaches (Du et al., 2024; Khan et al., 2024) show that iterative discussions among models can enhance reasoning quality and reduce factual errors. Likewise, Liang et al. (2023) proposes balancing divergent and convergent thinking, where LLMs first generate diverse hypotheses and then consolidate them into a coherent final answer. Building on this idea, we adopt the LLM-as-a-Judge paradigm to assess and combine extracted synthetic contexts with reasoning-focused outputs, ensuring that the final response remains both logically consistent and semantically faithful to the intended objective.

5 Conclusion

In this paper, we propose EMCEE, a multilingual prompting framework designed to enhance the handling of non-English queries in LLMs. We address the limitations of previous approaches by (1) explicitly extracting synthetic context from LLMs and (2) integrating it with reasoning pathways through an LLM-as-a-Judge mechanism to produce contextually appropriate and logically coherent responses. Empirical results show the effectiveness of EMCEE, particularly in low-resource language settings. These highlight the importance of incorporating query-relevant knowledge for non-English inputs and underscore the need for continued research in multilingual prompting from this view.

Limitations

While our approach yields notable performance improvements, the computational cost associated with multiple LLM inferences remains a concern. Nevertheless, as shown in Table 7, the performance gains are substantial enough to justify the increased cost. Notably, the ENG-COT method performs three rounds of distinct reasoning followed by a merging step, resulting in a total of three inferences—comparable to our framework. However, despite incurring a higher cost, ENG-COT achieves lower accuracy than our method, indicating that simply increasing the number of inference rounds does not necessarily lead to better results.

Also, integrating retrieval-augmented generation (RAG) to provide external cultural or contextual references could mitigate this limitation by supplementing the model with relevant knowledge. Nonetheless, even when compared against RAG-based baselines (RAG (NATIVE) and RAG (ENG)), EMCEE consistently outperforms them on most benchmarks, demonstrating that its internally extracted context is both more targeted and efficient than externally retrieved passages. Furthermore, our method is also effective for improving the multilingual Aya model (see Appendix 3.3), which is explicitly trained to handle low-resource languages, confirming that EMCEE generalizes effectively without additional training or external retrieval resources.

Broader Impact and Ethical Implications

We believe that our work can help make LLM advancements more accessible to non-English users. As LLMs become increasingly integral to digital communication, it is important to ensure fair access for people who speak different languages. By reducing language barriers, our research takes a key step toward making AI technologies less biased and more inclusive.

Acknowledgments

All authors are affiliated with the Department of Artificial Intelligence at Yonsei University. This research was supported in part by Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2020-II201361, Artificial Intelligence Graduate School Program (Yonsei University); No. RS-2025-02215344, Development of AI Technology with Robust and Flexible Resilience Against Risk Factors; No. RS-2025-25442405, Development of a Self-Learning World Model-Based AGI System for Hyperspectral Imaging).

References

Meta AI. 2024. Introducing llama 3.1: Our most capable models to date.

Anthropic. 2024. Claude 3.5 Haiku.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, and 1 others. 2024. Towards measuring the representation of subjective global opinions in language models. In *Conference on Language Modeling (COLM)*.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, and 1 others. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, and 1 others. 2022. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei.

2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with more persuasive llms leads to more truthful answers. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Jaehyung Kim and Yiming Yang. 2024. Few-shot personalization of llms with mis-aligned responses. *arXiv preprint arXiv:2406.18678*.

Junlong Li, Jinyuan Wang, Zhuosheng Zhang, and Hai Zhao. 2024. Self-prompting large language models for zero-shot open-domain qa. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2023. Encouraging divergent thinking in large language models through multi-agent debate. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2024. Is translation all you need? a study on solving multilingual tasks with large language models. *arXiv preprint arXiv:2403.10258*.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

OpenAI. 2024. GPT-4o mini: Advancing cost-efficient intelligence.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, and 1 others. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2023. Recitation-augmented language models. In *International Conference on Learning Representations (ICLR)*.

CodeGemma Team, Heri Zhao, Jeffrey Hui, Joshua Howland, Nam Nguyen, Siqi Zuo, Andrea Hu, Christopher A Choquette-Choo, Jingyue Shen, Joe Kelley, and 1 others. 2024. Codegemma: Open code models based on gemma. *arXiv preprint arXiv:2406.11409*.

Qwen Team. 2024. Qwen documentation. `https://qwen.readthedocs.io/en/latest/`. Accessed: 2025-10-07.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, and 1 others. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Liang Wang, Nan Yang, and Furu Wei. 2023a. Query2doc: Query expansion with large language models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023b. Self-knowledge guided retrieval augmentation for large language models. In *Findings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*.

A Experimental Details

This section provides comprehensive details on the experimental setup, including dataset descriptions, baseline methods, implementation specifics, and evaluation aspects used in our NLP experiments.

A.1 Datasets

The following tables list example instances from the M3-Exam (Table 9), MKQA (Table 10), XNLI (Table 12), XCOPA (Table 11), and GlobalOpinionQA (Table 13) datasets. The basic prompts used for these datasets are presented in Table 15.

M3-Exam M3-Exam (Zhang et al., 2023) is a multilingual multiple-choice question dataset designed to evaluate LLMs across eight languages: Afrikaans, Chinese, Italian, Javanese, Portuguese, Swahili, Thai, and Vietnamese. It consists of real-world exam questions collected from different countries, ensuring a diverse and authentic assessment of model performance. The dataset includes questions from South Africa for Afrikaans, China for Chinese, Italy for Italian, Indonesia for Javanese, Brazil for Portuguese, Kenya for Swahili, Thailand for Thai, and Vietnam for Vietnamese. Each language's dataset is further categorized into up to four subject areas, reflecting the structure of standardized exams in various educational systems. By incorporating multiple languages and educational contexts, M3-Exam offers a comprehensive benchmark for evaluating multilingual LLMs, particularly in underrepresented languages. Additionally, only text-based questions were used, explicitly excluding any questions requiring image-based inputs. Furthermore, questions that rely heavily on external background knowledge for answering were removed to ensure a fair evaluation of LLM capabilities based solely on textual reasoning. As a result, a total of 5,857 questions were selected for evaluation, providing a focused and controlled benchmark for assessing multilingual model performance, particularly in underrepresented languages.[2]

MKQA MKQA (Longpre et al., 2021) (Multilingual Knowledge Questions and Answers) is a dataset designed for evaluating open-domain question-answering capabilities across multiple languages. It is derived from the Natural Questions (NQ) dataset, which originally contains English questions and corresponding answers. For this study, we focused on nine languages: German, Spanish, French, Japanese, Russian, Thai, Turkish, Vietnamese, and Chinese. We selected 100 questions per language, resulting in a total of 900 samples for evaluation. Additionally, we excluded data with an entity type labeled as 'long-answer' or 'not answerable.' The model's performance was measured using the F1 score, ensuring a robust assessment of multilingual generation capabilities.[3]

XNLI XNLI (Conneau et al., 2018) (Cross-lingual Natural Language Inference) is a multilingual dataset designed to evaluate a model's ability to understand and reason across different languages. It is derived from the MultiNLI dataset, which consists of English sentence pairs labeled for natural language inference (NLI) tasks—entailment, contradiction, or neutral. For this study, we conducted evaluations in 14 languages: French, Spanish, German, Greek, Bulgarian, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi, Swahili, and Urdu. We selected 100 samples per language, resulting in a total of 1400 samples for evaluation. Model performance was measured using accuracy, ensuring a reliable assessment of cross-lingual natural language understanding.[4]

XCOPA XCOPA (Ponti et al., 2020) (Cross-lingual Choice of Plausible Alternatives) is a multilingual dataset designed to evaluate a model's ability to perform causal reasoning across different languages. It is an extension of the original COPA dataset, which consists of sentence pairs where one must determine the more plausible cause or effect. For this study, we conducted evaluations in 11 languages: Estonian, Haitian Creole, Indonesian, Italian, Quechua, Swahili, Tamil, Thai, Turkish, Vietnamese, and Chinese. We selected 100 samples per language, resulting in a total of 1100 samples for evaluation. Model performance was measured using accuracy, providing a standardized assessment of cross-lingual causal reasoning abilities.[5]

GlobalOpinionQA GlobalOpinionQA (Durmus et al., 2024) is a multilingual question-answering dataset designed to evaluate LLMs on opinion-

[2]The M3-Exam dataset was obtained from the official repository: https://github.com/DAMO-NLP-SG/M3Exam.

[3]The MKQA dataset was obtained from the official repository: https://github.com/apple/ml-mkqa.

[4]The XNLI dataset was obtained from Hugging Face: https://huggingface.co/datasets/facebook/xnli.

[5]The XCOPA dataset was obtained from Hugging Face: https://huggingface.co/datasets/cambridgeltl/xcopa.

Table 9: Example data from M3-Exam.

Component	Content
Question	Una scuola ha dieci classi, con una media di 22 alunni per classe. Le classi con 21 alunni sono sei; le classi con 24 alunni sono tre. Quanti alunni ci sono nella decima classe?
Options	A. 20 B. 22 C. 23 D. 25
Ground Truth	B

Table 10: Example data from MKQA.

Component	Content
Question	en moyenne quelle longueur a le wagon de train de marchandises
Answer Type	number_with_unit
Ground Truth	55.0 67.0 pied
aliases	55 67 pied

based questions collected from diverse countries. The dataset consists of survey-based multiple-choice questions covering a range of global topics. For this study, we conducted evaluations using data from 34 countries, including Greece, Sweden, China (non-national sample), Tunisia, Malaysia, Vietnam, Argentina, Russia, Egypt, Indonesia, Jordan, Mexico, Pakistan, Palestinian territories, Turkey, Ukraine, Kenya, France, Germany, Lebanon, Peru, Poland, South Korea, Italy, Spain, Brazil, Chile, Japan, Venezuela, Senegal, the Netherlands, Uganda, the Philippines, and Ethiopia. To ensure a reliable evaluation, we only included languages with at least 40 samples in the dataset. Since the dataset originally included the answer distribution from multiple respondents within the same country, we used the dataset from Kim and Yang (2024), which converts these distributions into a single-answer format by selecting the choice with the highest probability. This standardized format enables a more consistent assessment of model performance across different linguistic and cultural contexts. Model performance was measured using accuracy, providing insights into the model's ability to align with human responses across various regions.

A.2 Measurement details

To distinguish high- and low-resource languages for each dataset, we used performance measured on native queries. Since the classification relies on the model's performance, the high/low-resource split may vary across different models. Table 16 shows the split for M3-Exam, MKQA, XNLI, XCOPA, and GlobalOpinionQA based on GPT-4o-mini.

A.3 Implementation details

EMCEE Details. Table 17 below presents the prompts used for the baselines in our experiments. When selecting few-shot examples, we ensured that they did not overlap with the test dataset. If a train or development dataset was available, we sampled examples from those datasets. In cases where only a test dataset was available, we excluded the few-shot examples from the evaluation set. We used GPT-4o to generate answers for the few-shot examples, providing the ground-truth labels during selection to minimize errors and ensure their reliability. For M3-Exam, which consists of four subject categories, we used four few-shot examples, ensuring representation across different subject categories as much as possible. For other datasets that do not have predefined categories, we used three examples per dataset. In all cases, the examples were randomly selected to avoid selection bias. The few-shot examples are presented in Table 18.

RAG Details. For the retrieval-augmented baselines, we used the Google Custom Search API[6] to retrieve relevant web documents. To ensure fairness with our experimental setup, we only utilized the top-1 ranked document for each query and truncated it to the first 100 words. For RAG (NATIVE), we combined the retrieved document with native-language instructions, whereas RAG (ENG) used equivalent English instructions while keeping the query in the original language. This setup allows us to evaluate the influence of instruction language on retrieval-augmented prompting while maintaining consistent retrieval conditions across both variants.

[6] https://developers.google.com/custom-search/

Table 11: Example data from XCOPA.

Component	Content
Question	effect
Premise	Avevo svuotato le mie tasche.
Choice 1	Estrassi la matrice di un biglietto.
Choice 2	Trovai un'arma.
Label	0 (Choice 1)

Table 12: Example data from XNLI.

Component	Content
Premise	Nun, daran dachte ich nicht einmal, aber ich war so frustriert, dass ich am Ende doch mit ihm redete.
Hypothesis	Wir hatten ein tolles Gespräch.
Label	1 (neutral)

B Additional Analyses

B.1 Full experiment on other datasets

The experimental results from Table 1 in the main paper are presented in graphical form. Specifically, the results for MKQA are shown in Figure 7, XNLI in Figure 8, and XCOPA in Figure 9 to provide a clearer visualization of the findings.

B.2 Effect of language information.

We investigate the role of language information in EMCEE by conducting an ablation study that progressively removes linguistic cues from the Extracting and Merging stages (Table 19). In the full EMCEE setup, both stages use native-language signals in the instruction and the query (Figure 3). We begin from the original EMCEE configuration and first remove instruction-level language signals from the Merging stage, while keeping the Extracting stage unchanged. We then further remove query-level language signals in Merging by replacing the original query with its English-translated version. Next, we remove language information from the Extracting stage as well, while still retaining the query in Merging. Finally, we remove both instruction- and query-level language signals from Merging, yielding a setting in which both stages operate without explicit native-language information. As shown in Table 19, performance remains relatively stable for some high-resource languages. However, the overall results gradually decline as language information is removed, with the largest drop occurring when both instruction- and query-level signals are absent. These findings highlight the importance of explicit language guidance in both stages: instruction-level signals help maintain task alignment, while query-level signals preserve linguistic nuances that are particularly important for cross-lingual generalization, especially in low-resource settings.

Table 19: Accuracy on M3-Exam when removing language signals from instruction and query in the Extracting (ExT) and Merging (MeR) stages.

Methods	ExT	MeR	All / High / Low
EMCEE	both	both	77.4 / 83.3 / 71.5
	both	query	76.3 / 83.3 / 69.3
	query	query	76.4 / 84.1 / 68.8
	no	query	75.5 / 83.5 / 67.5
	no	no	73.1 / 82.9 / 63.4

B.3 Adaptive setting with M3-Exam

We conduct additional experiments comparing EMCEE pipeline with a adaptive contextual reasoning approach. Specifically, while EMCEE performs explicit extraction and merging of cultural and contextual knowledge before reasoning, the adaptive variant prompts the model to decide dynamically whether to generate such context. We sample 100 queries for each of the eight languages in the M3-Exam dataset (800 queries in total) and use the

Table 13: Example data from GlobalOpinionQA.

Component	**Content**
Question	Glauben Sie, dass die Regierung Saudi-Arabiens die persönlichen Freiheiten ihrer Bürger respektiert, oder glauben Sie das nicht?
Options	(A) Ja, respektiert persönliche Freiheiten (B) Nein, respektiert persönliche Freiheiten nicht
Ground Truth	B

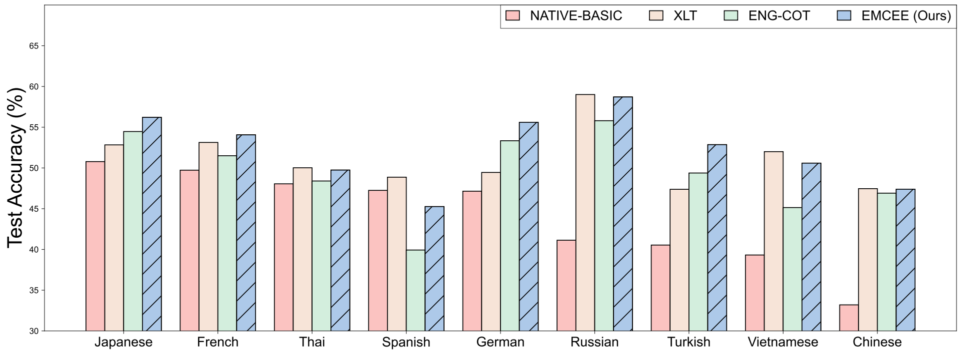


Figure 7: Overall language-wise improvement for MKQA.

`gpt-4o-mini` model for evaluation. The adaptive variant employs the following system prompt, enabling the model to determine on its own whether to produce a contextual note:

> Whenever you decide that a cultural/contextual note is required, output exactly:
> `Explanation: <your explanation here>`
> Otherwise, output exactly:
> `Explanation: None`
> Do not include any other words, headings, or punctuation.

We compare the following two configurations: 1) ADAPTIVE: Standard Chain-of-Thought prompting augmented with the above adaptive instruction. 2) EMCEE: Our original pipeline comprising extraction, reasoning, and merging stages. As summarized in Table 20, the adaptive approach achieves an average accuracy of 69.9%, whereas EMCEE attains 76.1%. These results indicate that explicitly extracting query-relevant knowledge for every query and subsequently determining its relevance yields stronger and more consistent improvements than relying on the model to decide on-the-fly.

Table 20: Comparison between adaptive contextual reasoning (ADAPTIVE) and EMCEE. The best results in each column are shown in **bold**.

Methods	**All**	**High**	**Low**
ADAPTIVE	69.9	79.5	60.3
EMCEE	**76.1**	**84.5**	**67.8**

Table 14: Test accuracy of three different LLMs on M3-Exam. The best and second best scores are highlighted in **bold** and underline, respectively.

Models (→)	**GPT-4o**			**Claude-Haiku**			**Llama-3.1-8B**		
Methods (↓)	All	High	Low	All	High	Low	All	High	Low
NATIVE-BASIC	78.1	84.5	71.7	67.4	76.4	58.3	49.8	60.9	38.7
TRANS-GOOGLE	68.4	78.3	58.5	60.1	67.5	52.7	53.9	64.0	43.8
XLT	81.8	89.2	74.3	74.0	84.4	63.6	25.8	30.7	20.8
NATIVE-COT	83.9	90.1	77.7	74.8	84.8	64.8	36.8	46.6	27.1
ENG-COT	82.9	88.1	76.7	75.4	83.9	**66.8**	40.4	47.2	33.6
EMCEE (Ours)	**85.7**	**91.5**	**79.8**	**75.6**	**85.8**	65.4	**56.9**	**69.3**	**44.5**

Table 15: Basic prompts used for different benchmarks.

Benchmark	Basic Prompt
M3-Exam	The following is a multiple-choice question. {Question} {Options} You should provide the final answer at the end in the format: ‘Answer: [option]’
MKQA	Answer the question in one or a few words in English: {Question} You should provide the final answer at the end in the format: ‘Answer: ’
XNLI	{Premise} Based on the previous passage, is it true that {Hypothesis}? Answer should be in the format of "Answer: [Yes or No or Maybe]".
XCOPA	Here is a premise: {Premise}. What is the {Question}? Help me pick the more plausible option: -choice1: {Choice1}, -choice2: {Choice2} You should only choose one option for your answer. You should answer the question in the format of ‘Answer: [1 or 2]’

Table 16: High- and low-resource subsets for each dataset. For GlobalOpinionQA, we report countries.

Dataset	High-resource languages	Low-resource languages
M3-Exam	Afrikaans, Chinese, Portuguese, Vietnamese	Italian, Javanese, Swahili, Thai
MKQA	German, Spanish, French, Japanese, Thai	Russian, Turkish, Vietnamese, Chinese
XNLI	French, German, Greek, Bulgarian, Russian, Thai, Hindi	Spanish, Turkish, Arabic, Vietnamese, Chinese, Swahili, Urdu
XCOPA	Estonian, Indonesian, Italian, Thai, Turkish, Vietnamese	Haitian Creole, Quechua, Swahili, Tamil, Chinese
GlobalOpinionQA	Greece, Sweden, China (Non-national), Ukraine, Kenya, France, Germany, Lebanon, Poland, S.Korea, Italy, Spain, Brazil, Chile, Japan, Senegal, Netherlands	Tunisia, Malaysia, Vietnam, Argentina, Russia, Egypt, Indonesia, Jordan, Mexico, Pakistan, Palest. ter, Turkey, Peru, Venezuela, Uganda, Philippines, Ethiopia

B.4 Extraction in native language

To examine the impact of instruction language on our method, we conduct an additional ablation study comparing native language and English language instructions for the extraction stage. While our main experiments adopt English instructions, reflecting the fact that LLMs are predominantly trained and optimized on English data (AI, 2024; Liu et al., 2024), we also test the effect of using native language instructions for synthetic knowledge extraction. Previous work (Huang et al., 2023) has shown that introducing English elements into multilingual prompts, for example for disfluency correction or inference tasks, often enhances model performance, which aligns with our design choice.

We evaluate 100 queries per language across eight languages in the M3-Exam (800 queries in total). In one condition, the model is prompted to extract knowledge using a native language instruction (NATIVE-EXTRACT). As shown in Table 21, the native language extraction achieves an average accuracy of 70.8%, while the English instruction variant slightly outperforms it with 72.5%. These results suggest that English based instructions remain advantageous, likely due to the models' pretraining bias toward English, although native language prompting remains a promising direction for future research.

Table 17: Prompts used for different baselines.

Settings	Content
Original Question	Pagupon Iku Jenenge Kandhang .... A. Pitik B. Macan C. Jaran D. Dara
Translated Question	Pagupon Is The Name Of The Cage .... A. Chicken B. Tiger C. Horse D. Dara
NATIVE-BASIC	Ing ngisor iki minangka pitakon pilihan ganda. {Original Question} Sampeyan kudu nyedhiyakake jawaban pungkasan ing pungkasan nganggo format: 'Wangsulan: [pilihan]'.
ENG-BASIC	The following is a multiple-choice question. {Original Question} You should provide the final answer at the end in the format: 'Answer: [option]'.
TRANS-GOOGLE	The following is a multiple-choice question. {Translated Question} You should provide the final answer at the end in the format: 'Answer: [option]'.
XLT	I want you to act as an question answering expert for Javanese. Request: {Original Question} You should retell the request in English. You should do step-by-step answer to obtain an option. You should step-by-step answer the request. You should tell me the answer in this format 'Answer : [option]'.
NATIVE-COT	Ing ngisor iki minangka pitakon pilihan ganda. {Original Question} Ayo mikir kanthi bertahap nganggo basa Jawa, lan weneh jawaban pungkasan ing pungkasan kanthi format: 'Jawaban: [pilihan]'
ENG-COT	The following is a multiple-choice question. {Original Question} Let's think step-by-step in English, and provide the final answer at the end in the format: 'Answer: [option]'

Table 21: Comparison between native language and English instructions for the extraction stage. The best results in each column are shown in **bold**.

Methods	All	High	Low
NATIVE-EXTRACT	70.8	78.8	62.8
ENG-EXTRACT (OURS)	**72.5**	**79.5**	**65.6**

B.5 Robustness across multiple runs

To evaluate robustness with respect to prompt construction and few-shot exemplar selection, we repeated all methods over three independent random seeds, where each run used a different randomly sampled set of few-shot examples. Table 22 reports the mean and standard deviation across the three runs. Overall, the variance across seeds is consistently small for all methods. Native-Basic achieves 68.0 ± 2.4, Eng-CoT achieves 72.9 ± 1.6, XLT achieves 70.5 ± 0.7, and EMCEE achieves 75.2 ± 2.1. Importantly, the performance gains of EMCEE remain larger than the standard deviations of the baselines, and its improvement over Eng-CoT and XLT is preserved across all runs. These results indicate that the effectiveness of EMCEE is not driven by a particular random seed or a specific selection of few-shot exemplars, but reflects a stable improvement across different prompt instantiations.

Table 22: Robustness across three independent runs. The best results in each column are shown in **bold**.

Methods	Mean ± Std
Native-Basic	68.0 ± 2.4
Eng-CoT	72.9 ± 1.6
XLT	70.5 ± 0.7
EMCEE (OURS)	$\mathbf{75.2 \pm 2.1}$

B.6 Merging step with different combinations

One possible concern is that the gain of EMCEE may stem simply from the merging procedure rather than from the complementary roles of its two components. To examine this possibility, we evaluated several alternative combinations under the same merge-based framework on the full M3-Exam dataset with GPT-4o-mini. As shown in Table 23, EMCEE, which combines EXTRACT and

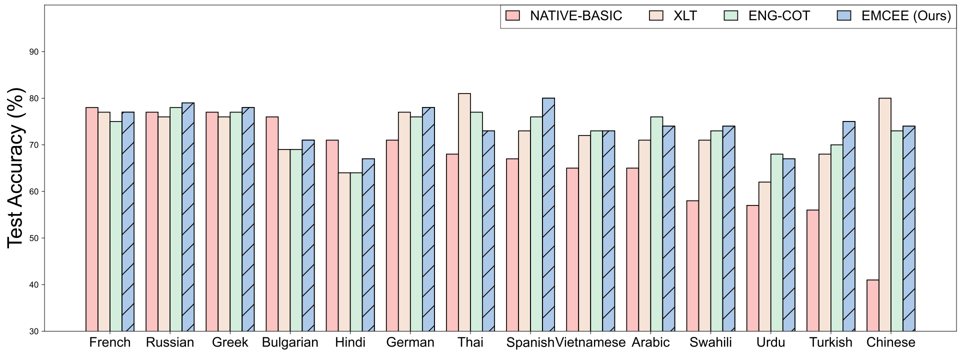


Figure 8: Overall language-wise improvement for XNLI.

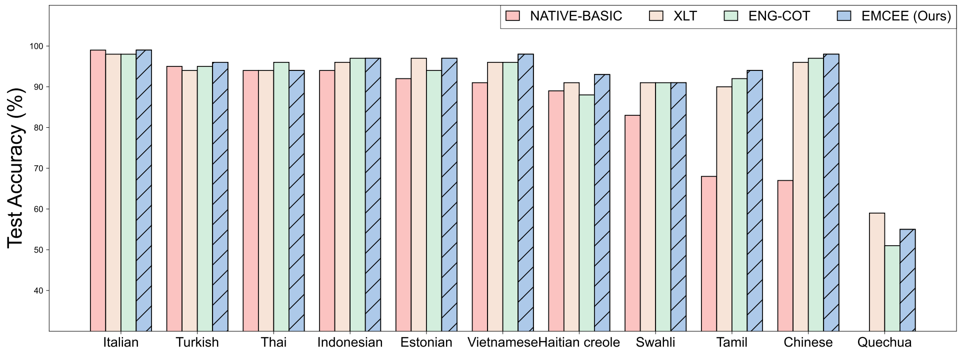


Figure 9: Overall language-wise improvement for XCOPA.

ENG-COT, achieves the best overall performance among all tested combinations. Specifically, alternative combinations such as NATIVE-BASIC + NATIVE-COT (73.1), ENG-BASIC + ENG-COT (74.4), NATIVE-BASIC + ENG-COT (74.6), EXTRACT + RAG (74.5), and ENG-COT + RAG (73.2) all remain below EXTRACT + ENG-COT (EMCEE), which reaches 77.4 average accuracy. This result suggests that the benefit does not arise from aggregation alone. Rather, the specific combination of what is being merged is crucial. We hypothesize that the strength of EMCEE comes from the complementary nature of its two paths. The EXTRACT step provides multilingual, culturally grounded synthetic context by explicitly eliciting background knowledge that the model may fail to recall when directly answering non-English questions. In contrast, the ENG-COT step offers a stable reasoning path by leveraging the model's generally stronger reasoning ability in English. The empirical results indicate that combining these two components is substantially more effective than other merge-based alternatives.

Table 23: Merge-based combinations on M3-Exam. The best results in each column are shown in **bold**.

Methods	All	High	Low
NATIVE-BASIC + NATIVE-COT	73.1	81.5	64.7
ENG-BASIC + ENG-COT	74.4	83.4	65.4
NATIVE-BASIC + ENG-COT	74.6	**83.5**	65.7
EXTRACT + RAG	74.5	83.0	65.9
ENG-COT + RAG	73.2	81.9	64.5
EXTRACT + ENG-COT (EMCEE)	**77.4**	83.3	**71.5**

B.7 Extended RAG baselines

To better understand whether EMCEE simply benefits from replacing retrieval with generation, we compare it against stronger retrieval-based baselines with multiple configurations. In addition to a basic single-passage setting, we evaluate full-passage retrieval, multi-passage retrieval, and a dense reranking pipeline. Table 24 summarizes the results over 100 queries for each M3-Exam language. Among the RAG variants, the strongest baseline is the setting with five retrieved passages followed by reranking, which achieves 71.3 average accuracy. However, this result still falls substantially short of EMCEE, which achieves 76.0 on the same evaluation subset. The gap is particularly notable in the low-resource setting, where the reranked RAG baseline reaches 62.0 while EMCEE achieves 69.3. These results suggest that the advantage of EMCEE is not simply due to using more context or a more elaborate inference pipeline. Instead, explicitly eliciting the model's own multilingual and culturally grounded knowledge can be more effective than relying solely on externally retrieved passages, especially in multilingual scenarios where retrieval may be sparse, noisy, or culturally misaligned.

Table 24: Comparison of extended RAG baselines. Best results are shown in **bold**.

Methods	All	High	Low
RAG			
One passage (100 words)	70.6	79.8	61.5
One passage (full text)	70.4	80.0	60.8
Five passages (100 words each)	69.0	78.0	60.0
Five passages + reranking	71.3	80.5	62.0
EMCEE (OURS)	**76.0**	**82.8**	**69.3**

B.8 Comparison with a recitation-based baseline

We further consider a recitation-based synthetic-context baseline inspired by prior work on self-generated contextual augmentation. In this setting, we run the EXTRACT stage twice and merge the two generated outputs, yielding a baseline that uses only synthetic context generation without the reasoning path of ENG-COT. We denote this variant as EXTRACT + EXTRACT → MERGE (RECITE). As shown in Table 25, this RECITE-style baseline achieves 72.4 average accuracy on the full M3-Exam dataset. Although it performs reasonably well, it remains below both ENG-COT + ENG-COT → MERGE (74.6) and ENG-COT + EXTRACT → MERGE (EMCEE) (77.4). This finding indicates that simply generating more synthetic text is not sufficient. Rather, the complementary combination of a knowledge-oriented path (EXTRACT) and a reasoning-oriented path (ENG-COT) is essential for achieving the full benefit of EMCEE.

Table 25: Comparison with a recitation-based synthetic-context baselines. Best results are shown in **bold**.

Methods	All	High	Low
ENG-COT + ENG-COT → MERGE	74.6	81.8	67.3
EXTRACT + EXTRACT → MERGE (RECITE)	72.4	80.4	64.5
ENG-COT + EXTRACT → MERGE (EMCEE)	**77.4**	**83.3**	**71.5**

C Qualitative Examples

C.1 Case study: when knowledge extraction can be detrimental

In this section, we provide additional qualitative examples illustrating cases where the extraction step can negatively affect performance. We consider two representative situations in which knowledge extraction may be counterproductive. The first occurs when the generated explanation introduces context that is misaligned with the intent of the query, as illustrated by the failure case in the main text (Figure 6). The second occurs when additional cultural or contextual grounding is unnecessary for solving the problem (Figure 10). In both cases, the reasoning-oriented Chain-of-Thought (CoT) response selected by the EMCEE framework tends to provide a more reliable and efficient solution.

In Figure 10, the question simply asks for the result of a basic arithmetic operation ("What is the value of 63.91 $\times$ 10?"). Here, the extracted explanation redundantly elaborates on the concept of multiplying by ten, which adds unnecessary verbosity without contributing to correctness. Since contextual knowledge is irrelevant to mathematical reasoning, this case demonstrates how extraction may distract from the concise reasoning path provided by CoT. Together, these examples highlight why the merging step in EMCEE, which selects between extracted and reasoning-based responses is crucial. When the extracted context is misleading or irrelevant, the model can rely on CoT reasoning to produce a more accurate and faithful answer.

C.2 More examples

Additional qualitative examples with EMCEE are shown in Figures 11 and 12. All the examples are from M3-Exam with GPT-4o-mini.

Table 26: Human evaluation of EMCEE

Metrics	Mean ± Std
Cultural Fidelity	3.4 ± 0.84
Factual Accuracy	4.2 ± 0.42
Stereotype/Bias	1.9 ± 1.10
Misinformation Risk	1.7 ± 0.82

D Human Evaluation of Synthetic Context Quality

To further assess the quality of EMCEE-generated synthetic contexts, we conducted a focused human evaluation on the Korean subset of GlobalOpinionQA. We recruited five native Korean speakers

(1) Native query and Options: Wat is die waarde van 63,91 $\times$ 10? (What is the value of 63.91 $\times$ 10?)
(A) 6,391 (B) 63,910 (C) 6391 (D) 639,1

(2) Correct Answer: (D)

(3) Extracted information: To find the value of 63.91 $\times$ 10, you need to multiply 63.91 by 10. When you multiply a number by 10, you simply shift the decimal point one place to the right.

Figure 10: Qualitative example from M3-Exam.

(1) Native query: Ni sentensi ipi sahihi?
A. Hewala! Nitakusaidia. B. Kefule! Naomba maji. C. Oyee! Tumeshindwa. D. Hamadi! Amekuja.

(2) Translated query: Which sentence is correct?
A. Hewala! I will help you. B. Lock it! I'm asking for water. C. Oye! We have failed. D. Hamadi! He has come.

(3) Incorrect responses by ENG-COT: To determine which sentence is correct, we need to analyze each option based on standard Swahili grammar and usage.
A. "Hewala! Nitakusaidia." - "Hewala" is an informal greeting or expression, and "Nitakusaidia" means "I will help you." This sentence is grammatically correct.
B. "Kefule! Naomba maji." - "Kefule" is not a standard Swahili word or expression. "Naomba maji" means "I am asking for water," which is correct, but the first part makes the whole sentence questionable.
C. "Oyee! Tumeshindwa." - "Oyee" is an informal expression, and "Tumeshindwa" means "We have failed." This sentence is also grammatically correct.
D. "Hamadi! Amekuja." - "Hamadi" is a name, and "Amekuja" means "He/She has come." This sentence is correct as well.
Now, we have three sentences (A, C, and D) that are grammatically correct. However, the question asks for the most appropriate or commonly used expression.
Considering the context of greetings and expressions, "Oyee!" is often used in celebratory or informal contexts, while "Hewala!" is also informal but less common.
Thus, the most appropriate and commonly used expression in this context is option C.

(4) Correct response by Extracting: To determine the correct sentence, we need to consider the context and usage of the words. In Swahili, "Hewala!" is an expression of excitement or encouragement, making option A a valid exclamation. However, "Kefule!" in option B is not a commonly recognized expression in Swahili, which makes it less likely to be correct. Option C, "Oyee! Tumeshindwa," uses "Oyee!" as an exclamation, but it conveys a negative context of defeat, which may not be appropriate depending on the intended meaning. Option D, "Hamadi! Amekuja," uses "Hamadi!" as a name or greeting, which is also acceptable. However, the most universally recognized and appropriate exclamation in a positive context is option A: "Hewala! Nitakusaidia," which expresses a willingness to help. Therefore, option A is the most suitable choice.

(5) Example of EMCEE response comparing (3) & (4) : Summary: User1 argues that options A, C, and D are grammatically correct, but suggests that option C is the most appropriate due to its common usage in celebratory contexts. User2 counters that while options A and D are valid, option A is the most suitable because it expresses a positive willingness to help, making it the best choice overall.
Conclusion: Given the arguments, option A is more plausible as it conveys a positive sentiment and is grammatically correct.

Figure 11: Qualitative example from M3-Exam.

who were also fluent in English, and asked them to evaluate the synthetic contexts generated by EMCEE for 51 items. For each item, annotators were shown the question, the answer options, and the generated synthetic context, all in English, and rated the context on a 5-point Likert scale.

We considered four evaluation dimensions. First, *Cultural Fidelity* measures whether the generated context aligns with Korean cultural and social perspectives. Second, *Factual Accuracy* measures whether the background knowledge is correct and trustworthy. We also included two safety-related dimensions: *Stereotype/Bias*, which captures whether the generated context contains culturally inappropriate generalizations or stereotypes, and *Misinformation Risk*, which measures whether the context may mislead readers through inaccurate or unsupported claims.

Table 26 summarizes the results. EMCEE achieves solid performance on both core quality dimensions, with a Cultural Fidelity score of 3.4 ± 0.84 and a Factual Accuracy score of 4.2 ± 0.42. In addition, the safety-related scores remain low, with Stereotype/Bias at 1.9 ± 1.10 and Misinformation Risk at 1.7 ± 0.82, indicating that the generated contexts rarely introduce problematic generalizations or misleading content.

Overall, these findings provide supporting evidence that EMCEE's synthetic contexts are not only culturally grounded and factually reliable, but also relatively safe in terms of bias and misinformation. Although this human study is limited to one language subset and a modest number of annotated examples, it offers an additional validation that the extracted context is generally appropriate for downstream multilingual reasoning.

E Usage of AI Assistants

In preparing this work, we made limited use of AI-based writing assistants to improve sentence clarity, refine grammar, and enhance readability. These tools were employed strictly for editorial purposes and did not contribute to the conception of the research questions, development of methodology, execution of experiments, or interpretation of results. All technical content, analyses, and conclusions presented in this paper were conceived and authored solely by the researchers.

Table 18: Few-shot examples used in the M3-Exam Afrikaans experiment.

Component	**Content**
Question	Identifiseer die hoofsin in die onderstaande sin. Gesonde mense drink elke dag in die winter water.
Options	(A) Mense drink water. (B) Mense drink elke dag water. (C) Mense drink in die winter water. (D) Mense drink.
Category	Language
Explanation	In the given sentence, G̈esonde mense drink elke dag in die winter water,ẅe are looking for the main clause, which is the complete thought or statement that provides the essential information without additional modifiers. The main subject is m̈ense(̈people) and the main verb is d̈rink(̈drink). Option (A) M̈ense drink waterc̈aptures this essential information, as it includes the subject M̈enseänd the verb d̈rink,älong with the object ẅater,f̈orming the core of the sentence without the additional details about when or how often the action occurs. Options B, C, and D include modifiers or less complete versions of the main statement.
Question	IOpwekking van elektrisiteit deur vallende water, word ... genoem.
Options	(A) kernkrag (B) hidro-elektrisiteit (C) transformasie (D) aardlekkasie
Category	Natural Science
Explanation	The generation of electricity by falling water is termed ḧydro-electricity,ẅhich is option (B). This process involves converting the kinetic energy of falling or flowing water into electrical energy using turbines and generators. Unlike nuclear power (option A) which relies on nuclear reactions, or transformation (option C) which refers to changing one form of energy to another, hydro-electricity specifically pertains to water. Aardlekkasie (option D) refers to ground leakage, which is unrelated to the generation of electricity through water. Understanding the specific terminologies associated with different energy sources is crucial in selecting the correct answer.
Question	Om te verhoed dat die gruweldade van WWII weer plaasvind, het die 'Algemene Vergadering van die VN' .... in Desember 1948 aanvaar.
Options	(A) die Wet op Afsonderlike Geriewe (B) die paswette (C) die Universale Verklaring van Menseregte (D) die Suid-Afrkaanse Handves van Regte
Category	Social Science
Explanation	The correct answer is (C) the 'Universal Declaration of Human Rights.' After the atrocities of World War II, the United Nations sought to ensure such events would not happen again by establishing a set of fundamental human rights that all nations should uphold. In December 1948, the General Assembly of the United Nations adopted the Universal Declaration of Human Rights, which outlines basic rights and freedoms to which all humans are entitled. Options (A), (B), and (D) are incorrect as they pertain to specific laws or documents related to South Africa and are unrelated to the global context addressed by the UN in 1948.
Question	Watter van die volgende getalle is 'n ⟨b⟩priemgetal⟨/b⟩?
Options	(A) 1 (B) 2 (C) 0 (D) 4
Category	Math
Explanation	A prime number is defined as a natural number greater than 1 that has no positive divisors other than 1 and itself. In this context, understanding prime numbers is crucial. Option A (1) is not a prime number because it only has one divisor, which is itself. Option C (0) is not a prime number because it is not greater than 1. Option D (4) is not a prime number because it has divisors other than 1 and itself, namely 2. Therefore, the only option that fits the definition of a prime number is B (2), as it is greater than 1 and can only be divided by 1 and 2.

(1) Native query: Essa foi, provavelmente, a última Copa do Mundo de Futebol Feminino em que o público teve a chance de ver o trio de veteranas Marta, Cristiane e Formiga atuando juntas em campo. Formiga, que já chegou aos 41 anos e agora disputou sua sétima Copa, não deve voltar aos gramados para o mundial de 2023. No atual elenco da Seleção Feminina, por exemplo, são poucas as jogadoras que têm menos de 30 anos — sendo que, no masculino, essa faixa etária não é vista com tanta frequência entre os jogadores convocados para a Seleção. Disponível em: https://capricho.abril.com.br. Acesso em: 3 set. 2019 (adaptado).
A longa permanência de atletas do futebol feminino na Seleção brasileira, numa faixa etária considerada avançada para o esporte, ocorre devido à
(A) pouca competitividade nos campeonatos. (B) força do vínculo entre as atletas e a Seleção. (C) capacidade de adaptação e resistência muscular. (D) falta de incentivo para a modalidade.

(2) Translated query: This was probably the last Women's World Cup in which the public had the chance to see the trio of veterans Marta, Cristiane and Formiga working together on the field. Formiga, who has reached the age of 41 and has now played in her seventh World Cup, is not expected to return to the pitch for the 2023 World Cup. In the current squad of the Women's National Team, for example, there are few players who are under 30 years old — and, in men, this age group is not seen as frequently among players called up for the national team. Available at: https://capricho.abril.com.br. Accessed on: 3 Sep. 2019 (adapted).
The long stay of female football athletes in the Brazilian team, in an age group considered advanced for the sport, occurs due to the
(A) little competitiveness in championships. (B) strength of the bond between the athletes and the National Team. (C) adaptation capacity and muscular resistance. (D) lack of incentive for the modality.

(3) Incorrect responses by ENG-COT: To analyze the question, we need to consider the context provided in the passage regarding the long careers of female athletes in the Brazilian national football team, particularly focusing on the older age of players like Formiga.
1. Option (A) - pouca competitividade nos campeonatos: This suggests that the lack of competitive opportunities might lead to players staying longer in the national team. However, this does not directly explain why older players are still selected; it rather implies a stagnation in player development.
2. Option (B) - força do vínculo entre as atletas e a Seleção: This option points to a strong emotional or professional connection between the players and the national team. This could be a valid reason for older players continuing to play, as they may feel a strong commitment to represent their country.
3. Option (C) - capacidade de adaptação e resistência muscular: This option suggests that older players have developed the ability to adapt and maintain their physical condition. While this could be true, it does not directly address the reason for their prolonged presence in the national team.
4. Option (D) - falta de incentivo para a modalidade: This implies that there is insufficient support for younger players to emerge, which could lead to older players remaining in the team longer. This could be a contributing factor, but it does not specifically highlight the emotional or professional ties that might keep veteran players in the squad.
Considering these points, option (B) seems to be the most relevant as it directly addresses the emotional and professional commitment of the players to the national team, which could explain why they continue to play at an older age.
Final answer: Answer: B

(4) Correct response by Extracting: The passage highlights the presence of veteran players in the Brazilian women's national football team, particularly noting that many athletes are older than 30, which is considered an advanced age for competitive sports. This situation can be attributed to the lack of opportunities and resources in women's football, which often leads to fewer younger players being developed and promoted to the national team. As a result, the existing veteran players remain in the team longer due to the limited influx of younger talent. Therefore, the most appropriate answer is (D) falta de incentivo para a modalidade, as it reflects the systemic issues in women's football that contribute to the prolonged presence of older athletes in the national team.

(5) Example of EMCEE response comparing (3) & (4) : Summary: The debate centers around the reasons for the long careers of older female athletes in the Brazilian national football team, particularly focusing on the presence of veteran players like Formiga. User1 argues that the strong emotional and professional bond between the players and the national team (Option B) is the primary reason for their continued participation. In contrast, user2 contends that the lack of opportunities and resources in women's football (Option D) leads to fewer younger players being developed, resulting in older players remaining in the team longer.
Conclusion: While both arguments present valid points, user2's perspective on the systemic issues in women's football (Option D) provides a broader context for understanding the situation. The lack of incentive and support for the development of younger talent is a significant factor that explains the prolonged presence of older athletes in the national team.

Figure 12: Qualitative example from M3-Exam.