Cognitive Priming Prompting Facilitates Knowledge Elicitation in Multilingual Large Language Models

Baixuan Li[1] **(baixuan@seu.edu.cn)** **Yunlong Fan**[1] **(fanyunlong@seu.edu.cn)**

Tianyi Ma[2] **(matiany3@msu.edu)** **Zhiqiang Gao**[1,*] **(zqgao@seu.edu.cn)**

[1]School of Computer Science and Engineering, Southeast University, Nanjing 211189, China
[2]Department of Computer Science and Engineering, Michigan State University
[*]Corresponding Author

Abstract

Multilingual large language models (MLLMs) typically underperform when answering questions in non-native languages compared to their native language. Although existing translate-then-answer prompting methods partially alleviate this issue, their performance remains suboptimal compared to directly answering questions in the native language. Moreover, current studies lack a clear explanation for this gap. In this study, we attribute this issue to incomplete Cognitive Priming, a phenomenon observed from human cognition. And we point out that while existing methods achieve **Language Priming** (**LP**), they overlook **Domain Priming** (**DP**). To address this, we propose **Cog**nitive **Prim**ing Prompting (**CogPrim**), which employs a Role-Enhanced Multi-MLLM Collaboration strategy to ensure both LP and DP, thereby improving knowledge elicitation for non-native QA tasks. Across five language question-answering benchmarks, CogPrim increases accuracy by up to **31.28%**, surpassing all state-of-the-art related methods. This approach contributes to advancing the understanding of human-like cognitive behaviors in MLLMs, fostering better MLLM service to a broader range of multilingual users.

Keywords: large language models; cognitive priming; question answering; few-shot prompting; multilingualism

Introduction

Cognitive priming (Schacter, 1992) is widely observed in human cognition, referring to the enhanced efficiency in processing information that follows prior exposure to specific stimuli, such as particular languages (Marian & Neisser, 2000) or domain-relevant terminology (Neely, 1977). For multilingual individuals with relevant expertise, describing a question in their native language and employing precise domain-specific terms substantially facilitates understanding and answering. For example, an English-speaking chemistry expert is more likely to efficiently access relevant domain knowledge if the question includes the specialized term "completely combusted" rather than the more generic "completely burned" or its non-native equivalent.

Similarly, the phenomenon of cognitive priming has also been observed in multilingual large language models (MLLMs) (Touvron, Lavril, et al., 2023). Recent studies have highlighted that MLLMs achieve higher accuracy when answering questions described in their native language (i.e., the language with the highest proportion during training[1]) compared to those described in a non-native language (Etxaniz, Azkune, Soroa, Lacalle, & Artetxe, 2024). This suggests that

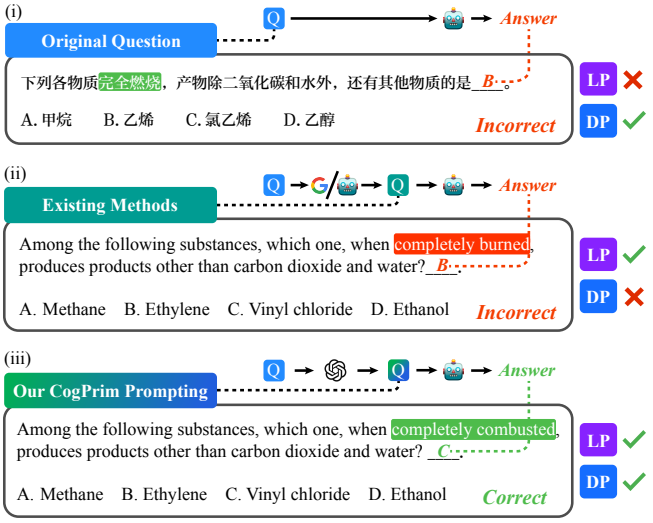

Figure 1: The presence of Language Priming (LP) and Domain Priming (DP) affects knowledge elicitation when MLLMs answer non-native language (Chinese) questions. Identical icons represent the same question/model.

in the question-answering (QA) process, knowledge acquired during training is elicited more effectively when the question is presented in the native language, but less effectively when it is presented in non-native languages[2], as depicted in Figure 1 (i). Existing approaches suggest that translate-then-answer prompting (Schulhoff et al., 2024) can be used to mitigate this issue. These methods involve first translating the question into the MLLM's native language and then prompting the MLLM to answer the translated question.

However, cognitive priming is multifaceted, with different types of stimuli affecting information processing from various dimensions. This study focuses on two key forms of priming: **(i) Language Priming (LP)** (Marian & Neisser, 2000; Marian & Spivey, 2003; Zeng, Chen, & Guo, 2022), which leverages exposure to a MLLM's most frequently encountered (native) language to more effectively activate learned knowledge, and **(ii) Domain Priming (DP)** (Meyer & Schvaneveldt, 1971; Tulving & Thomson, 1973; Neely,

[1]Such as English for Llama (Touvron, Lavril, et al., 2023), which accounts for over 70% of the tokens in the pretraining corpus.

[2]For the MLLM depicted in Figure 1, English serves as the native language, while Chinese represents the non-native language.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

1977; Liu, Chen, & Xu, 2022), which enhances concept recall through domain-specific terminology closely tied to specialized knowledge. While current translate-then-answer prompting methods (Schulhoff et al., 2024) successfully achieve LP by translating questions from non-native into native languages, they often translate domain-specific terms overly generic and literal, failing to achieve DP adequately. For example, as illustrated in Figure 1(ii), translating the specialized chemical term "completely combusted" merely as "completely burned" dilutes the precise domain cue, impairing subsequent knowledge elicitation.

Specifically, we identify two primary reasons for these shortcomings: **(i) Limited translation model functionality**: paired with an MLLM that handles QA tasks in the native language, Shi et al. (2022) paired an external neural machine translation (NMT) model with an MLLM to implement the translate-then-answer process (*NMT-Translation*). However, constrained by their training paradigms, NMT models cannot adjust their translation style based on domain-specific instructions and can only produce overly generic and literal translations. **(ii) Excessive workload on a single model**: Some methods employ the same MLLM for both translation and QA, implementing *Self-Translation* (Zhang, Li, Hauer, Shi, & Kondrak, 2023; H. Huang et al., 2023). While MLLMs outperform NMT models by adjusting translation styles based on instructions, a single MLLM often struggles to balance translation and QA due to its differential aptitudes. Strong QA ability in the native language is offset by limited proficiency in non-native languages, hindering accurate domain-specific translation. On the other hand, MLLMs focused on multilingual translation often lack sufficient QA capabilities.

To overcome these limitations, we propose **Cog**nitive **Prim**ing Prompting (**CogPrim**), which not only translates non-native questions into an MLLM's native language but also preserves key domain-specific terminology. By simultaneously achieving Language Priming (LP) and Domain Priming (DP), CogPrim better elicits the MLLM's learned knowledge. Our proposed CogPrim employs a **Role-Enhanced Multi-MLLM Collaboration** strategy (Talebirad & Nadiri, 2023; Dong, Jiang, Jin, & Li, 2024) with two specialized roles: a Translator LLM, which translates questions from a non-native to the native language of the Speaker LLM based on domain-specific instructions, and a Speaker LLM, which focuses on answering the translated questions. Employing a specialized MLLM as the translator addresses the inherent functional constraints of NMT-Translation, while distributing workload alleviates the burden on a single MLLM in Self-Translation. As shown in Figure 1(iii), CogPrim accurately translates Chinese into English (ensuring LP) and retains fine-grained domain terminology (ensuring DP), thus more effectively eliciting the MLLM's domain-relevant knowledge.

Our contributions are primarily as follows: (i) We propose a cognition-inspired MLLM prompting strategy (CogPrim) that integrates both language and domain priming, facilitating knowledge elicitation in non-native language QA tasks. Experiments on five MLLMs with English as the native language[3] demonstrate up to a **31.28%** accuracy increase over five non-native QA benchmarks, surpassing all state-of-the-art related methods. (ii) We enhance the accessibility of MLLMs for non-English-native users, enabling performance comparable to native-level English prompts without relying on self-performed translations or external tools. In summary, this work bridges cognitive priming with existing MLLM prompting methods, advancing the community's understanding of human-like cognitive behaviors in MLLMs and offering better service to a broader range of multilingual users.

Cognitive Priming Meets Multilingual LLMs

In human cognition, cognitive priming describes how one stimulus influences the response to a subsequent stimulus. In early experiments (Meyer & Schvaneveldt, 1971), it was found that when a word (e.g., "nurse") is preceded by a semantically related word (e.g., "doctor"), individuals are able to judge the word more quickly. This effect has been supported by lexical decision tasks (Schvaneveldt & Meyer, 1973), demonstrating that activation can spread to semantically related concepts. Since MLLMs operate by predicting the next token (Brown et al., 2020), the manner in which context is presented can significantly influence their predictions (Leidinger, van Rooij, & Shutova, 2023). Posing questions in specific patterns that align with how MLLMs previously encountered and learned the relevant knowledge makes it easier for them to recall accurate answers. This phenomenon closely mirrors the origins and effects of cognitive priming.

Given this, we aim to explore whether cognitive priming, which facilitates human knowledge elicitation (with a focus on language priming (Marian & Neisser, 2000) and domain priming (Meyer & Schvaneveldt, 1971) in this study), can enhance the performance of MLLMs in QA. Additionally, we focus on non-native language questions, as they pose greater challenges for MLLMs compared to native language questions, which have undergone more extensive training[4]. Additionally, considering real-world scenarios where users may pose questions in languages other than the native language of MLLMs, our study aims to enhance MLLMs' ability to better serve user groups from diverse linguistic backgrounds.

Role-Enhanced Multi-MLLM Collaboration

Since the capabilities of a single MLLM are limited, and different MLLMs exhibit varying strengths, in order to allow each MLLM to fully leverage its unique advantages, previous work has proposed using multiple MLLMs to fulfill distinct roles within a collaborative framework (Talebirad & Nadiri, 2023; Dong et al., 2024). In this study, the translate-then-answer process for non-native QA is inherently divided into two sub-processes: translating and answering. As depicted in

[3]In our extensive investigation, we found that due to the highest quantity and quality of English corpora during the training process, all existing MLLMs have English as their native language.

[4]Native language questions are predominate in the MLLMs' training data, leading to more extensive training on these questions.

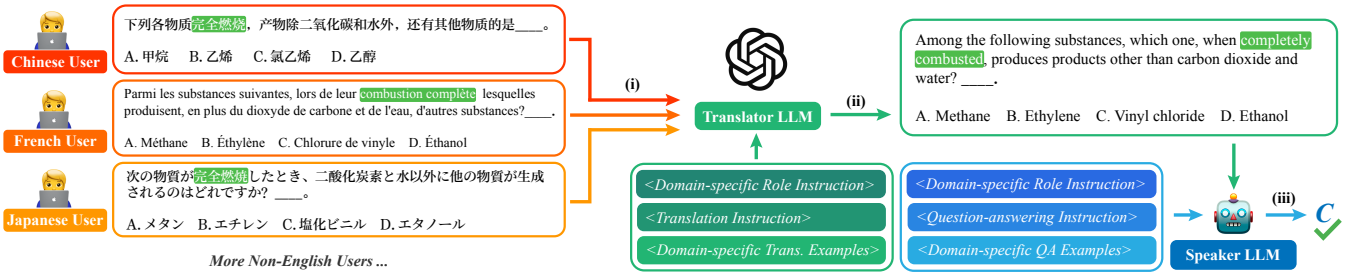

Figure 2: Non-native language question-answering workflow of CogPrim. (i) Non-English users issue queries. (ii) The Translator LLM translates the non-native language questions into the native language (English) of the Speaker LLM. (iii) The Speaker LLM answers the native language question. CogPrim provides 2-shot in-context examples for Trans. and 5-shot for QA.

Figure 2, we designed a Role-Enhanced Multi-MLLM Collaboration framework and defined two distinct roles to handle these sub-processes separately. Their respective targets and required characteristics are outlined as follows:

Translator LLM requires MLLMs to have strong multilingual comprehension and semantic preservation abilities. It needs to translate any received non-native language questions into the native language of the Speaker LLM. **Speaker LLM** requires MLLMs that excel in their native language and are capable of understanding the given non-native language, though not necessarily to an exceptional degree. It needs to rely on its own knowledge to provide answers to the questions translated by the Translator LLM.

Cognitive Priming Prompting

Utilizing our constructed Role-Enhanced Multi-MLLM Collaboration framework, we further proposed **Cog**nitive **Prim**ing Prompting (**CogPrim**) to achieve Language Priming (LP) and Domain Priming (DP) simultaneously during the translate-then-answer process for non-native QA.

Instruction: System Prompt

<For Translator LLM>
Domain-specific Role & Translation Instruction:
You are a professional {non-native language name}-English translator. Translation rules: Proper nouns in English or {non-native language name} need to be translated according to the {discipline name} domain-specific terms, retain the original meaning to the greatest extent, and follow the original format in the translation process.

<For Speaker LLM>
Domain-specific Role & QA Instruction:
You are a professional {discipline name} expert, and you are currently answering a multiple-choice question about {discipline name}, you need to provide only one option as the answer based on the question, and you only need to return one single capital character as the answer.

As shown in Figure 2, we construct domain-specific instructions and inject domain-specific context through few-shot examples to achieve two effects: **Individual Enhancement:** enabling each MLLM in the framework to recall relevant knowledge via domain-specific role prompting; **Joint Enhancement:** ensuring the Translator LLM retains and conveys domain-specific terms during translation, presenting them in the Speaker LLM's native language for better understanding. This process leverages the Translator LLM's multilingual alignment capabilities to explicitly pass domain-specific information to the Speaker LLM, triggering both LP and DP, and further eliciting its relevant knowledge.

Experiments and Discussion

To investigate whether our proposed CogPrim enhances cognitive priming for MLLMs to facilitate knowledge elicitation, we use multidisciplinary QA as the evaluation task, measuring knowledge elicitation through MLLM response accuracy. Given that nearly all mainstream MLLMs use English as their native language, we select *English (en)* as the native language in this study. To ensure linguistic diversity, we include five representative non-native languages: *Arabic (ar), Chinese (zh), French (fr), German (de), and Japanese (ja).*

Dataset. We use the Multilingual Massive Multitask Language Understanding (MMMLU) benchmark[5], which includes expert-translated versions of MMLU (Hendrycks et al., 2021) in the aforementioned five non-native languages. Each version contains 14,079 multiple-choice questions across 57 disciplines, with identical questions and answers across languages. Additionally, we include the C-Eval Chinese benchmark (Y. Huang et al., 2023) for further ablation and case studies. Unlike MMMLU, which focuses on knowledge from the English-speaking community, C-Eval is centered on Chinese-speaking community knowledge and includes 13,948 questions across 52 disciplines.

CogPrim Setup. In the proposed CogPrim, we selected GPT-4o-mini as a universal translator[6] to accomplish translations from the five non-native languages to the native language, for its comprehensive multilingual understanding ca-

[5] https://huggingface.co/datasets/openai/MMMLU
[6] We can also select smaller models with careful consideration.

Table 1: Performance on MMMLU, measured by QA accuracy (%). Red, yellow, and green indicate negative, suboptimal, and optimal enhancement, respectively. CogPrim employs the universal translator, GPT-4o-mini. *+Human (Gold)* represents the performance of the Speaker LLMs when answering on the human-constructed English version MMLU.

Model	ar	zh	fr	de	ja
Phi-3-mini (3.8B)	33.66	43.04	57.49	55.06	41.38
+Self-Translation	40.58	54.14	62.68	62.23	53.75
+Google-MT	62.99	63.59	65.39	64.32	64.59
+CogPrim (Ours)	64.18	64.23	65.55	65.58	65.00
+Human (Gold)	———		68.10		———
Phi-3-small (7B)	39.24	55.66	67.12	65.29	53.25
+Self-Translation	53.04	64.10	66.22	68.10	63.76
+Google-MT	68.35	69.56	71.46	69.14	70.32
+CogPrim (Ours)	70.52	70.30	72.16	71.92	71.46
+Human (Gold)	———		74.67		———
Gemma-1.1 (7B)	39.72	47.70	50.35	49.36	45.11
+Self-Translation	46.28	49.15	52.46	52.25	48.36
+Google-MT	54.65	55.44	56.99	56.32	55.70
+CogPrim (Ours)	56.10	56.03	56.84	56.72	56.55
+Human (Gold)	———		58.12		———
Mistral-0.3 (7B)	32.25	41.23	49.45	47.96	38.75
+Self-Translation	39.99	46.04	52.32	52.26	46.13
+Google-MT	54.49	55.46	57.14	55.59	55.85
+CogPrim (Ours)	56.08	56.17	57.34	56.93	56.59
+Human (Gold)	———		58.70		———
Llama-2 (7B)	11.88	18.81	18.53	24.45	16.51
+Self-Translation	10.78	15.16	20.05	17.42	12.69
+Google-MT	32.55	31.80	33.22	32.74	29.92
+CogPrim (Ours)	32.54	32.34	33.18	32.79	31.40
+Human (Gold)	———		34.97		———

pabilities. Additionally, to analyze the effects of the Translator LLM with varying capabilities on CogPrim, we chose the Qwen series MLLMs (Bai et al., 2023) as Chinese-to-English translators, for their leading Chinese comprehension capabilities. Furthermore, we selected five representative MLLMs with the strong English QA capability to serve as Speaker LLMs, including models from the Phi (Abdin et al., 2024), Gemma (Team et al., 2024), Mistral (Jiang et al., 2023), and Llama (Touvron, Martin, et al., 2023) series.

Baselines. Two top-notch related methods most relevant to the CogPrim were selected as baselines: **(i) Self-Translation** (Zhang et al., 2023; H. Huang et al., 2023), which entails a single MLLM sequentially undertaking the translating and answering processes, serving both as the Translator LLM and the Speaker LLM. **(ii) Google-MT** (Shi et al., 2022), which uses Google Neural Machine Translation system[7] (API) as the translator and MLLMs as the Speakers. It is important to note that the requirement for Speaker LLMs to possess the five non-native languages comprehension abilities is crucial for conducting Self-Translation and direct evaluations on the non-native language, ensuring fair performance comparisons.

[7]Translation-only NMT model (limited functionality), unlike the general-purpose instruction-following MLLMs.

Table 2: Performance on C-Eval, measured by QA accuracy (%). CogPrim employes the Qwen series MLLMs as the Chinese-to-English translators. The reported results represent the optimal performance achieved by Speaker LLMs when using different versions of Qwen as the translator. *Hard* represents the QA accuracy based on assessments conducted on the more challenging subjects, such as "college physics".

Model	zh	zh (Hard)	Model	zh	zh (Hard)
Phi-3-mini (3.8B)	41.2	36.3	Phi-3-small (7B)	49.0	41.6
+Self-Translation	43.8	37.7	+Self-Translation	52.0	42.1
+Google-MT	50.9	40.4	+Google-MT	55.7	42.7
+CogPrim (Ours)	51.3	41.3	+CogPrim (Ours)	55.9	44.7
Gemma-1.1 (7B)	44.4	36.3	Mistral-0.3 (7B)	42.8	32.6
+Self-Translation	41.9	33.9	+Self-Translation	34.8	30.9
+Google-MT	46.7	38.2	+Google-MT	48.0	33.3
+CogPrim (Ours)	47.7	38.6	+CogPrim (Ours)	48.4	35.3
Llama-2 (7B)	21.3	14.7			
+Self-Translation	9.6	10.3			
+Google-MT	25.4	15.1			
+CogPrim (Ours)	27.6	18.6			

Overall Performance Results

As shown in Table 1, we demonstrated the performance of the proposed CogPrim, along with top-notch related methods, on the MMMLU benchmark. Overall, CogPrim achieved accuracy surpassing all top-notch related methods. Furthermore, it closely approached the gold standard performance achieved when answering questions directly on the human expert-constructed English version. Similarly, as shown in Table 2, CogPrim also achieved the best performance.[8]

More specifically, the performance of different methods exhibited a particular incremental pattern, i.e. *Self-Translation* < *Google-MT* < *CogPrim*. We confirmed in the next section that such pattern are positively correlated with the degree of incremental cognitive priming achieved by each method. Moreover, virtually all translate-then-answer methods outperformed direct answers[9] in the non-native language, reflecting the impact of Language Priming (LP).

However, it is noteworthy that on a few MLLMs, particularly Llama-2 (7B), due to the limitations of its own instruction-following capabilities, the performance of Self-Translation was even worse than answering directly in the non-native language. This degradation is caused by the error propagation during the translation phase, further emphasizing the necessity of incorporating additional models better suited for multilingual translation to achieve more stable improvements from the translate-then-answer process.

[8]C-Eval is only available in Chinese and does not have a human expert-constructed English (or other languages) version.

[9]For each MLLM, the first row in Tables 1 and 2 shows its performance on directly answering non-native language questions.

Table 3: Chinese-to-English non-native QA cases in C-Eval. Errors in red and correct/consistent in green.

Original Question	Google-MT Trans. Question	CogPrim Trans. Question	Answers
云南民俗中有“女儿国”和“君子国”，这“两绝”的形成与下列哪种因素有关____。 A. 生活水平低 B. 文化素质差 C. 交通闭塞 D. 开发历史短	There are “Daughter Country” and “Gentleman Country” in Yunnan folklore. Which of the following factors is related to the formation of these “two uniques”____. A. Low living standards B. Poor cultural quality C. Impeded transportation D. Short development history	The formation of “the Kingdom of Women” and “the Kingdom of Gentlemen” in Yunnan folklore is related to____. A. Low living standards B. Poor cultural literacy C. Isolation due to poor transportation D. Short development history	Original: B +Google-MT: D +CogPrim : C True Label: C

CogPrim Retains More Domain-Specific Terms

To evaluate the superiority of CogPrim in cognitive priming, we developed a pairwise comparison method. Assuming all methods achieve comparable LP levels, we focus on Domain Priming (DP). Using GPT-4o-mini, we quantify DP by comparing how well each method retains domain-specific terms in translated questions. GPT-4o-mini selects the better translation between two options, identifying the superior method.

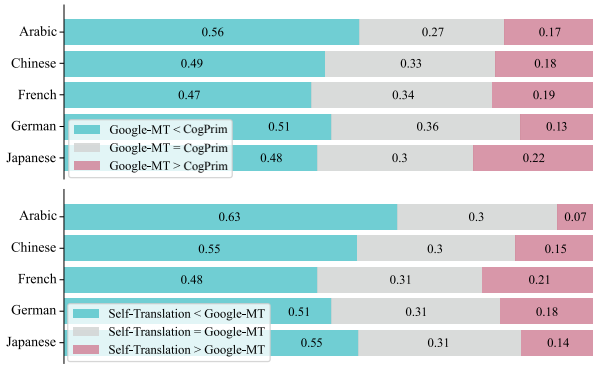

Figure 3: Pairwise compared domain-specific terms advantage ratios on MMMLU. Gray bars indicate nearly equivalent term retention between the two methods.

As depicted in Figure 3, the pairwise comparison results clearly demonstrate the same incremental pattern reported in the previous section, i.e., *Self-Translation* (Phi-3-small (7B)) $<$ *Google-MT* $<$ *CogPrim*. This confirms the effectiveness of CogPrim's improvements in enhancing DP. With comparable LP, the degree of cognitive priming can be reflected by the degree of DP, further validating that our proposed CogPrim provides more effective cognitive priming. This superiority is explicitly demonstrated through non-native QA accuracy.

Two-Aspect Domain Priming Enhancement

To determine how CogPrim retains domain-specific terms and enhances DP in the translate-then-answer process, we summarized the findings into the following two principal aspects:

(i) Accurate descriptions of domain-specific entities. Precisely translating descriptions specific to the domain of the entities, avoiding overly generic or literal translations. As shown in Table 3, Google-MT produces a literal translation such as “Daughter Country” without considering the folkloric context of the Chinese expression, whereas CogPrim accurately uses “the Kingdom of Women”. Similarly, CogPrim uses “literacy” instead of “quality”, etc. **(ii) Explicit descriptions of relationships between domain-specific entities.** Although implicit relationships between domain-specific entities can be inferred with effort, this increases cognitive load. Explicitly describing domain-specific relations between entities avoids ambiguity. As shown in Table 3, CogPrim explicitly clarifies the causal relationship between “poor transportation” and “isolation”, illustrating the geographical context of the question.

Cognitive Priming Triggers Knowledge Activation

We additionally analyzed how enhanced DP in translation impacts knowledge elicitation during the answering process at a more fine-grained activation level. As the Speaker LLMs only need to generate the answer options, the last hidden state for predicting the answer token reflects the internal knowledge activation pattern. Therefore, we extracted it for further analysis. Additionally, on MMMLU, we used the knowledge activation generated by Speaker LLMs when answering on the human-constructed English version MMLU as the gold knowledge activation standard[10] for evaluation.

We measured the average Euclidean distance between activation vectors from various methods and those from the human gold standard. As shown in Table 4, CogPrim achieved the closest approximation to the human gold standard in knowledge activation. Notably, the activation distance evaluation followed the same trend as mentioned earlier: *Self-Translation* $<$ *Google-MT* $<$ *CogPrim*, confirming that enhanced DP, or more effective cognitive priming during translation, triggers knowledge activation in Speaker LLMs.

Furthermore, Figure 5 shows that Google-MT, the best-performing baseline, triggered some knowledge activation but at a level weaker than CogPrim, making it insufficient to correct answers. In contrast, CogPrim, with enhanced DP and more effective cognitive priming, triggered stronger activation, successfully correcting previously incorrect answers.

[10]This also corresponds to the optimal performance that Speaker LLMs can achieve on this benchmark.

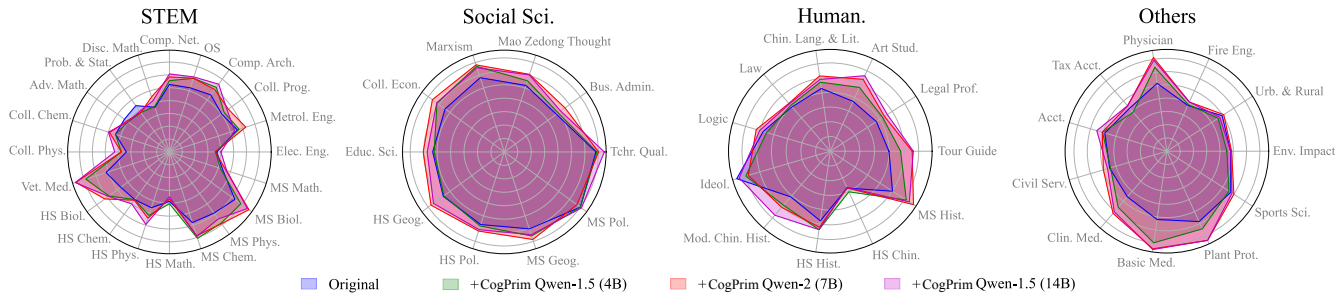

Figure 4: Performance of CogPrim across 52 disciplines on C-Eval, with Phi-3-small (7B) as the Speaker LLM.

Table 4: Average Euclidean distances between knowledge activation (extracted from Phi-3-small (7B)) using various methods and the human gold standard, with smaller distances indicating closer approximation to the human gold standard.

Method	**ar**	**zh**	**fr**	**de**	**ja**
Original	111.40	69.54	21.50	36.72	87.43
+Self-Translation	35.31	20.86	15.60	15.76	22.31
+Google-MT	13.24	12.84	9.06	10.07	13.88
+CogPrim (Ours)	9.84	11.24	7.46	8.38	11.28

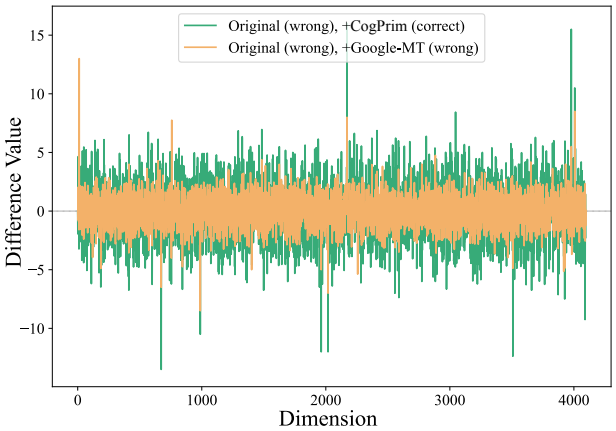

Figure 5: Activation differences between different methods for the same questions. Contents in parentheses indicate the correctness of the (Phi-3-small (7B)) Speaker's responses.

Impact of Domain-Language Characteristics

We further explored the characteristics that exist between specific domains and languages and their impact on CogPrim. As shown in Figure 4, while translating questions from a non-native language (Chinese) to the native language (English) using CogPrim significantly improves accuracy across most disciplines, exceptions exist. In some disciplines, translation does not lead to performance gains. These disciplines can be categorized into the following two types:

(i) Language-Insensitive Disciplines: Disciplines like Probability and Statistics (Prob. & Stat.), which mostly rely on understanding mathematical formulas to answer questions. These mathematical formulas are consistent across languages. In such cases, adding a translation process can introduce potential errors, such as the loss of content in mathematical formulas, impacting the correct understanding of the questions. **(ii) Language-Knowledge Bound Disciplines:** Disciplines like Ideological and Moral Cultivation (Ideol.) are closely tied to specific languages due to cultural and national differences. Training materials in different languages may convey distinct knowledge on similar topics, and translating such questions into English can cause confusion and conflict, hindering accurate knowledge elicitation.

It should be noted that while there is a possibility that applying CogPrim in the aforementioned two types of disciplines may not offer benefits, this is not always the case. The actual occurrence still largely depends on the capabilities of the Translator LLMs and Speaker LLMs involved.

Conclusion and Limitations

This study reveals that MLLMs underperform on questions in non-native languages compared to their native ones. We interpret this phenomenon through the lens of Cognitive Priming in human cognition, particularly Language Priming (LP) and Domain Priming (DP). Furthermore, we propose a novel CogPrim method to enhance cognitive priming in MLLMs when answering non-native questions, facilitating knowledge activation while also advancing the community's understanding of the integration of MLLMs with human cognition. Moreover, it makes a positive contribution to enabling MLLMs to better serve a broader range of multilingual users.

However, our work is constrained by current MLLMs' reliance on English as their native language; even MLLMs with improved capabilities in other languages (e.g., Qwen) still achieve higher accuracy with English contexts. Future research should develop MLLMs natively trained in various languages or explore advanced language-transfer techniques. Moreover, the effectiveness of CogPrim depends on the quality of both Translator and Speaker LLMs. In some domains, where knowledge is language-insensitive or tightly bound to a specific language, switching to English does not improve, and may even hinder, knowledge elicitation. We encourage future efforts to enable dynamic language switching guided by the knowledge domain, ultimately promoting more flexible and effective multilingual question-answering.

References

Abdin, M., Jacobs, S. A., Awan, A. A., Aneja, J., Awadallah, A., Awadalla, H., ... others (2024). Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., ... others (2023). Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... others (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Dong, Y., Jiang, X., Jin, Z., & Li, G. (2024). Self-collaboration code generation via chatgpt. *ACM Trans. Softw. Eng. Methodol.*. doi: 10.1145/3672459

Etxaniz, J., Azkune, G., Soroa, A., Lacalle, O., & Artetxe, M. (2024). Do multilingual language models think better in english? In *Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 2: Short papers)* (pp. 550–564).

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. In *International conference on learning representations.*

Huang, H., Tang, T., Zhang, D., Zhao, W. X., Song, T., Xia, Y., & Wei, F. (2023). Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the association for computational linguistics: Emnlp 2023* (pp. 12365–12394).

Huang, Y., Bai, Y., Zhu, Z., Zhang, J., Zhang, J., Su, T., ... others (2023). C-eval: a multi-level multi-discipline chinese evaluation suite for foundation models. In *Proceedings of the 37th international conference on neural information processing systems* (pp. 62991–63010).

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., ... others (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Leidinger, A., van Rooij, R., & Shutova, E. (2023). The language of prompting: What linguistic properties make a prompt successful? In *Findings of the association for computational linguistics: Emnlp 2023* (pp. 9210–9232).

Liu, J., Chen, Y., & Xu, J. (2022). Saliency as evidence: Event detection with trigger saliency attribution. In *Proceedings of the 60th annual meeting of the association for computational linguistics* (pp. 4573–4585).

Marian, V., & Neisser, U. (2000). Language-dependent recall of autobiographical memories. *Journal of Experimental Psychology: General*, *129*(3), 361.

Marian, V., & Spivey, M. (2003). Bilingual and monolingual processing of competing lexical items. *Applied Psycholinguistics*, *24*(2), 173–193.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of experimental psychology*, *90*(2), 227.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of experimental psychology: general*, *106*(3), 226.

Schacter, D. L. (1992). Priming and multiple memory systems: Perceptual mechanisms of implicit memory. *Journal of cognitive neuroscience*, *4*(3), 244–256.

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., ... others (2024). The prompt report: A systematic survey of prompting techniques. *arXiv preprint arXiv:2406.06608*.

Schvaneveldt, R. W., & Meyer, D. E. (1973). Retrieval and comparison processes in semantic memory. *Attention and performance IV*, 395–409.

Shi, F., Suzgun, M., Freitag, M., Wang, X., Srivats, S., Vosoughi, S., ... others (2022). Language models are multilingual chain-of-thought reasoners. In *The eleventh international conference on learning representations.*

Talebirad, Y., & Nadiri, A. (2023). Multi-agent collaboration: Harnessing the power of intelligent llm agents. *arXiv preprint arXiv:2306.03314*.

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., ... others (2024). Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., ... others (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... others (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological review*, *80*(5), 352.

Zeng, T., Chen, C., & Guo, J. (2022). First language translation involvement in second language word processing. *Frontiers in Psychology*, *13*, 986450.

Zhang, X., Li, S., Hauer, B., Shi, N., & Kondrak, G. (2023). Don't trust chatgpt when your question is not in english: A study of multilingual abilities and types of llms. In *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 7915–7927).