Unlocking the Non-Native Language Context Limitation: Native Language Prompting Facilitates Knowledge Elicitation

Anonymous ACL submission

Abstract

Multilingual large language models (MLLMs) struggle to answer questions posed in non-dominant languages, even though they have acquired the relevant knowledge from their dominant language corpus. In contrast, human multilinguals can overcome such non-native language context limitations through Positive Native Language Transfer (PNLT). Inspired by the process of PNLT, we analogize the dominant language of MLLMs to the native language of human multilinguals, and propose ***Nat**ive **Lan**guage Prompting* (***NatLan***) to simulate the PNLT observed in human multilinguals. It explicitly creates native language contexts for MLLMs to facilitate the elicitation of the rich native language knowledge during question-answering, unlocking the limitations imposed by non-native language contexts. By employing multi-MLLM collaboration, NatLan reduces the workload on each MLLM in simulating PNLT and refines semantic transfer. On the C-Eval benchmark, NatLan provides up to a **10.1%** average accuracy improvement and up to a **5.0%** increase in the hard-level subset across five MLLMs, surpassing all top-notch related methods. Our code is available at https://github.com/AnonyNLP/NatLan.

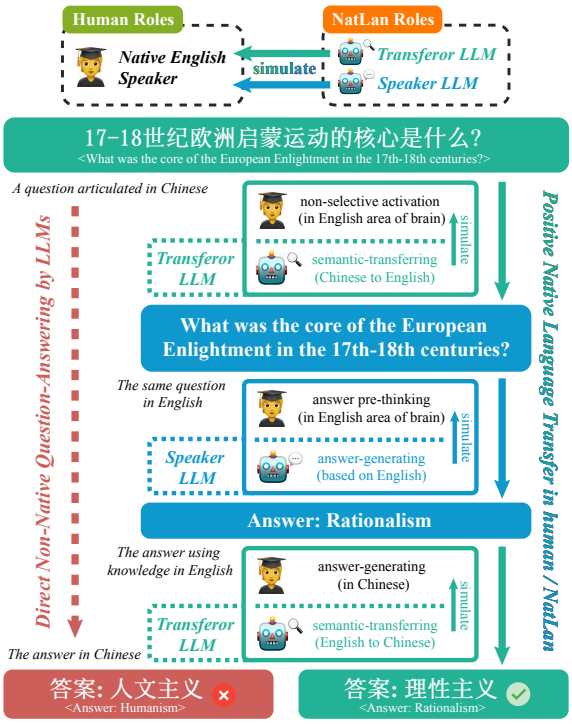

Figure 1: To address the failure of knowledge elicitation when directly answering in non-native language (Left), NatLan simulates the Positive Native Language Transfer of a human multilingual by utilizing two different MLLMs (Right). English meanings appear in < >. The roles involved in the process are displayed at the top.

1 Introduction

Multilingual large language models (MLLMs) (Brown et al., 2020; Achiam et al., 2023) have propelled the advancement of nearly all natural language processing tasks across various languages (Xu et al., 2024; Li et al., 2024). However, it's observed that MLLMs perform poorly on questions articulated in non-dominant languages, as depicted in Figure 1 (Left), failing to answer some questions that they could address when presented in their dominant language (i.e., the language with the highest proportion during training, such as English for Llama (Touvron et al., 2023a), which accounts for over 70% of the tokens in the pretraining corpus). This indicates that MLLMs are unable to effectively apply the rich knowledge acquired from dominant language contexts when operating in non-dominant language contexts. Although this inability can be attributed to the insufficient training arising from the differing volumes (Xue et al., 2021; ImaniGooghari et al., 2023) and quality (Sitaram et al., 2023) of training data across various languages, constructing large-scale, high-quality data across all languages is extremely labor-intensive and not feasible.

In contrast, such issues rarely occur in human multilinguals. Although human multilinguals also possess a most proficient language, typically their

native language, they can still correctly answer a question posed in their less proficient non-native languages, provided they have already acquired relevant knowledge in their native language (Nsengiyumva et al., 2021). In cognitive science, the process of using the rich knowledge acquired in one's native language to benefit addressing questions in a less proficient non-native language is known as the *Positive Native Language Transfer* (*PNLT*) (Gass and Selinker, 1992). As depicted in Figure 1 (Right), for human multilinguals, the native language regions of their brain are non-selectively activated when addressing questions in a non-native language (Zeng et al., 2022), then they can autonomously perform pre-thinking in their native language before responding, thereby flexibly invoking knowledge acquired in their native language.

Given that Ren et al. (2024) have observed significant similarities between MLLMs and the human brain in language processing, we analogize ***the dominant language of MLLMs*** (hereafter referred to as the native language) to ***the native language of human multilinguals***. Phenomena similar to PNLT have also been observed to occur autonomously in MLLMs: they tend to generate intermediate representations (Wendler et al., 2024) and output tokens (Marchisio et al., 2024) in their native language when addressing questions posed in a non-native one. However, since the cognitive capabilities of MLLMs fall considerably short of those of the human brain (Chemero, 2023), relying solely on a single MLLM for autonomous implicit processing cannot replicate the PNLT of human multilinguals.

Considering that explicit prompts enhance the consistency of MLLMs with brain cognitive language processing (Ren et al., 2024), we attempt to design specific prompting processes that explicitly guide multiple MLLMs to collaboratively simulate the PNLT of human multilinguals when addressing questions in non-native languages. This aims to replicate a brain-like cognitive process, thereby addressing the issue of MLLMs' inability to effectively utilize the rich native language knowledge.

In this study, we propose ***Nat**ive **Lan**guage Prompting* (***NatLan***), which decomposes PNLT simulation into semantic-transferring and answer-generating, sequentially undertaken by two distinct MLLMs, referred to as the *Transferor LLM* and the *Speaker LLM*. Through the collaboration of two MLLMs, NatLan reduces the workload on each MLLM involved in simulating the PNLT of human multilinguals, and leverages the outstanding capabilities of the Transferor LLM in the non-native target language (hereafter referred to as the target language) to achieve the semantic transfer from the target language to the native language. As depicted in Figure 1 (Right), NatLan simulates PNLT by first using the Transferor LLM to translate questions from the target language into the native language of the Speaker LLM before the Speaker LLM answers. This approach explicitly creates native language contexts for the Speaker LLM to elicit the rich native language knowledge, unlocking the limitations imposed by the non-native language contexts on the effective application of knowledge when answering questions in the target language.

Applied to five MLLMs (Speaker LLMs) (Touvron et al., 2023b; Jiang et al., 2023; Team et al., 2024; Abdin et al., 2024), NatLan achieves up to a **10.1%** average accuracy improvement in the C-Eval benchmark of question-answering (Huang et al., 2023), as well as up to a **5.0%** increase in the hard-level subset, surpassing all top-notch related methods (Schulhoff et al., 2024). Furthermore, we explore how the semantic capabilities of three Transferor LLMs (Bai et al., 2023) impact the effectiveness of NatLan. This study contributes to advancing the understanding of MLLMs from the perspective of explicit PNLT simulation.

2 Related Work

Positive Native Language Transfer in Multilingualism. For human multilinguals, previous work (Wu et al., 2022; Gao et al., 2023) indicated that they tend to subconsciously process texts in the native language when using other languages, with the native language regions of the brain being non-selectively activated (Zeng et al., 2022). This facilitates the effective access of native language knowledge to address questions in non-native languages, without the need for the question to be presented specifically in the native language context. Similarly, English-centric LLMs tend to generate intermediate representations (Wendler et al., 2024) and outputs in English (Marchisio et al., 2024). Ren et al. (2024) noted that explicit prompts contribute to the consistency of LLMs with human brain cognitive language processing. Our proposed NatLan explicitly simulates the Positive Native Language Transfer (PNLT) in prompting processes to facilitate the activation of regions similar to the native language areas in the human brain within MLLMs, thereby achieving brain-like knowledge elicitation.

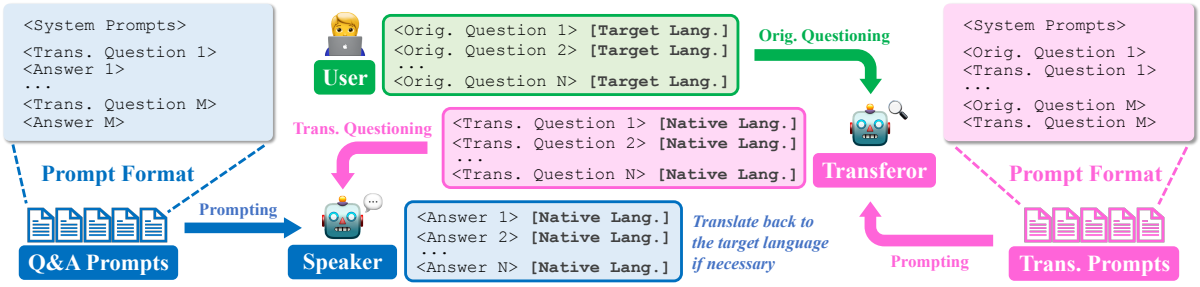

Figure 2: Question-answering workflow of the proposed NatLan. (i) **Semantic-transferring** by the Transferor, which translates the input questions from the target language into the native language of the Speaker. (ii) **Answer-generating** by the Speaker in the native language. In collaboration, the Transferor and the Speaker utilize distinct (five-shot, i.e., $M = 5$) prompts. Details of the prompts can be seen in Appendix A.1.

Translate First Prompting. Translate First Prompting (Schulhoff et al., 2024) aims to leverage the strength of MLLMs in English. Etxaniz et al. (2024) introduced Self-Translation, which requires MLLMs themselves to perform translation tasks, before answering questions. However, this encounters *Language Comprehension Bottlenecks*: if the model has poor capabilities in the target language, it may not complete the translation task accurately, leading to performance instability. Shi et al. (2022) used external Neural Machine Translation (NMT) systems to translate the questions. However, unlike MLLMs (Vilar et al., 2023; Guo et al., 2024; Kang et al., 2024), NMT systems lack rich semantic understanding capabilities, resulting in overly literal translations (Tu et al., 2017; Zhu et al., 2023; Ai et al., 2023). Our proposed NatLan reinterprets the effectiveness of translate-first prompting from the perspective of PNLT in human multilinguals and suggests employing multi-MLLM collaboration to achieve a more effective PNLT simulation.

3 Multi-MLLM Collaboration

Due to the varying capabilities of different LLMs, previous work has proposed using multiple LLMs to fulfill distinct roles within a collaborative framework (Talebirad and Nadiri, 2023; Dong et al., 2024). In this study, we decomposed the Positive Native Language Transfer (PNLT) simulation into two more straightforward sub-processes: (i) semantic-transferring and (ii) answer-generating.

Since one single MLLM's capabilities are insufficient for simulating the PNLT of human multilinguals, we designed a Multi-MLLM Collaboration framework and defined distinct roles for different MLLMs, collaborating to simulate the PNLT progressively, with their respective targets and required characteristics outlined as follows:

(i) **Transferor** requires MLLMs that are proficient in the target language and also possess strong capabilities in the native language of the subsequently mentioned Speaker LLMs. It undertakes semantic-transferring: translating questions from the target language into the native language of the Speaker LLMs, and translating the responses of Speaker LLMs back into the target language when required.

(ii) **Speaker** requires MLLMs that excel in their native language (the dominant language during training) and are capable of understanding the target language, though not necessarily to an exceptional degree. It undertakes answer-generating: understanding questions translated by the Transferor and providing answers based on their acquired knowledge.

The Multi-MLLM Collaboration reduces the workload on each MLLM and alleviates the capability bottlenecks by assigning different MLLMs to each specific sub-process within PNLT.

4 Native Language Prompting

Utilizing our constructed Multi-MLLM Collaboration framework, we further proposed ***Nat**ive **Lan**guage Prompting* (***NatLan***) to simulate the PNLT of human multilinguals. The question-answering workflow is illustrated in Figure 2.

As depicted in Figure 2, NatLan initially constructs domain-specific translation prompts (Pink) to provide domain-specific contexts, facilitating the Transferor LLMs' grasp of proper terms specific to the domain. This enables the accurate and coherent semantic transfer of the original questions from the target language to the native language. Subsequently, the proposed NatLan constructs domain-specific Q&A prompts (Blue), which also provide

domain-specific contexts, promoting knowledge recall by the Speaker LLMs for specific domain questions. It is important to note that the Q&A prompts at this stage exhibit the translated question examples, ensuring consistency with the process undertaken by the Speaker LLMs, namely answering the translated questions in their native language.

By employing NatLan, we present questions semantically transferred into the native language to the Speaker LLMs before answering, which mimics the PNLT, facilitating the rich native language knowledge elicitation in the Speaker LLMs.

5 Experiments

To explore the improvements that NatLan brings to knowledge elicitation, we selected question-answering as the evaluation task because it clearly indicates whether the relevant knowledge in the MLLMs has been correctly elicited. Since the native language (dominant language) of nearly all mainstream MLLMs is English, we have selected *English* as the native language in this study. Subsequently, considering that the level of knowledge elicitation requires sufficient language resources for comprehensive, multidisciplinary capability evaluation, we chose another representative language, *Chinese*, as the target language.

Dataset. Based on the target language (Chinese), we selected the C-Eval benchmark of question-answering (Huang et al., 2023) to assess the knowledge elicited from MLLMs. C-Eval comprises 13,948 multiple-choice questions across 52 different disciplines (subsets), providing a comprehensive knowledge evaluation in Chinese contexts.

NatLan Setup. In the proposed NatLan, the Transferor must be capable of translating the content from the target language (Chinese) into the native language (English) as accurately and coherently as possible. Therefore, we selected the Qwen series MLLMs (Bai et al., 2023) as Transferors, for their leading capabilities in Chinese comprehension among all MLLMs. We chose Qwen models with 4B, 7B, and 14B parameters to analyze the effects of Transferors with varying capabilities on NatLan in §5.5 and §5.6. Additionally, we selected a five representative MLLMs with strong English comprehension skills and the capability to understand Chinese to serve as Speakers. These include models from the Phi (Abdin et al., 2024), Gemma (Team et al., 2024), Mistral (Jiang et al., 2023), and Llama (Touvron et al., 2023b) series. For ease of joint deployment with the Transferor LLMs, all these Speaker LLMs possess a moderate parameter scale, ranging from 3.8B to 7B.

Baselines. Two top-notch related methods most relevant to the NatLan were selected as baselines: (i) **Self-Translation** (Etxaniz et al., 2024), which entails a single MLLM sequentially undertaking the semantic-transferring and answer-generating processes, serving both as the Transferor and the Speaker. (ii) **Google-MT** (Shi et al., 2022), which uses Google Neural Machine Translation (NMT) system (API) as the Transferor and MLLMs as the Speaker. It is important to note that the requirement for Speaker LLMs to possess Chinese comprehension abilities is crucial for conducting Self-Translation and direct evaluations on Chinese questions, ensuring fair performance comparisons. More details are available in Appendix A.1.

5.1 Overall Performance Results

We conducted a comparative analysis of performance between the proposed NatLan method and top-notch related methods across the test sets of 52 different disciplines within the C-Eval benchmark.

As shown in Table 1, while Self-Translation can bring certain improvements for some Speaker LLMs, the performance enhancement is not stable. This instability arises because Self-Translation uses Speaker LLMs as their own Transferors, encountering *Language Comprehension Bottlenecks* in the target language. Specifically, it cannot ensure that Speaker LLMs fully comprehend the inherent semantics of the questions in the target language, thus failing to guarantee accurate and coherent semantic transfers. If semantic transfer errors occur during the translation, it can significantly impair the subsequent behavior of Speaker LLMs, potentially causing the performance of Speaker LLMs in their native language to decline below that of directly answering questions in the target language.

Google-MT, by incorporating state-of-the-art Google Neural Machine Translation (NMT) systems as Transferors, ensuring relatively high-quality translations and stable performance improvements. Our proposed NatLan further refines this process by employing additional MLLMs with superior semantic understanding capabilities as Transferors. This addresses the shortcomings of NMT systems, which often produce overly literal translations due to a lack of rich semantic abilities,

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

Table 1: Comparison with top-notch related methods. Performance metrics are measured by the average accuracy. *Lang.* indicates the language of the questions. Red, yellow, and green indicate negative, suboptimal, and optimal enhancement, respectively. The NatLan configurations are the optimal setup reported in §5.5.

thus achieving superior semantic transfer (see §5.3 for details). The proposed NatLan achieves optimal performance across all five Speaker LLMs.

5.2 NatLan Produces More Relative Improvements

To explore in more depth, we conducted a detailed performance analysis of Google-MT and our proposed NatLan method on the validation sets of specific disciplines within the C-Eval benchmark.

We define our analysis process as follows: Considering each discipline individually, we calculate the relative performance improvements brought by NatLan/Google-MT compared to having Speaker LLMs directly answer questions in Chinese (Original). Specifically, this involves computing the relative increase in the number of correct answers provided by NatLan/Google-MT compared to the Original. Subsequently, we apply Min-Max Normalization to the relative improvements achieved by NatLan/Google-MT across various disciplines, resulting in normalized relative improvements.

As shown in Figure 3, NatLan provides more relative improvements than Google-MT in the majority of disciplines. It is important to note that we have excluded disciplines from this analysis where neither method provided more correct answers than the Original. Additionally, since the performance gains from Self-Translation are quite limited and often result in frequent performance declines, this method has not been included in the analysis.

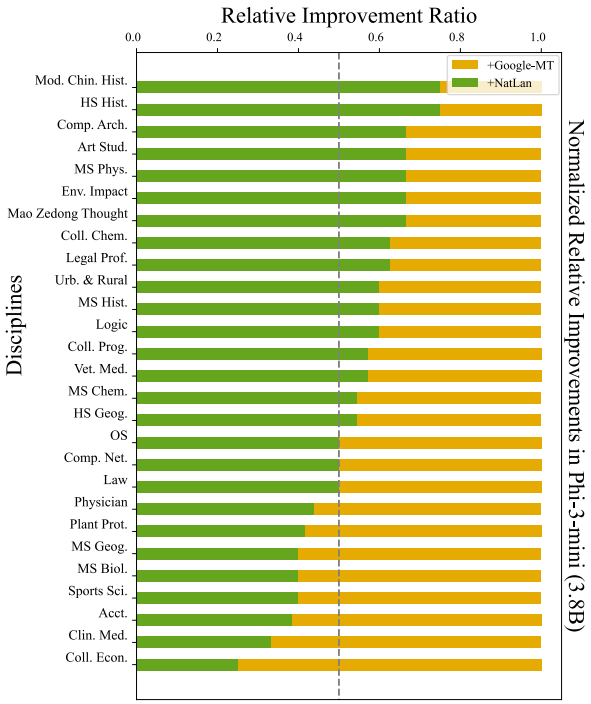

Figure 3: Normalized relative improvements in specific disciplines, with the dashed grey line indicating where their respective relative improvements are equivalent.

5.3 NatLan Refines Semantic Transfer

To substantiate NatLan's superiority in refining semantic transfer, we sampled representative questions from the C-Eval test sets for a comparative analysis. *Original* indicates that the Speaker LLMs respond directly to questions in Chinese.

Enhanced Semantic Coherence. Semantic coherence aims to emphasize the relationships between relevant entities in questions and answers. As shown in the first row of Table 2, NatLan uses "is accessed" to highlight the relationship between the "operand" in the question and the "addressing method" in the answers, reducing the difficulty for Speaker LLMs in recalling the relevant knowledge.

Enhanced Semantic Accuracy. As shown in the second row of Table 2, NatLan uses "Kingdom" instead of "Country", which more accurately captures the folkloric connotation of the term. Additionally, it uses "literacy" instead of "quality",

Original Question	Google-MT Trans. Question	NatLan Trans. Question	Answers
单地址指令中为了完成两个数的算术运算，除地址码指明一个操作数外，另一个采用____方式。 A. 立即寻址 B. 隐含寻址 C. 间接寻址 D. 基址寻址	In order to complete the arithmetic operation of two numbers in a single-address instruction, in addition to the address code indicating one operand, the other one uses ____ method. A. Immediate addressing B. Implicit addressing C. Indirect addressing D. Base addressing	In a single-address instruction to perform arithmetic operations on two numbers, apart from the operand specified by the address code, the other one is accessed using the ____ method. A. Immediate addressing B. Implicit addressing C. Indirect addressing D. Base addressing	Original: C +Google-MT: C +NatLan : B True Label : B
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 +NatLan : C True Label: C

Table 2: Chinese-to-English translation cases in C-Eval test sets. More cases are available in Appendix A.2.

enhancing the semantic precision. Moreover, in option C, it conveys the main reason as "Isolation" rather than merely "Impeded", enabling Speakers to understand the answer more accurately.

Overall, NatLan leverages the rich semantic capabilities of Transferor LLMs to deliver translations that surpass those of NMT systems, which refines the semantic transfer from the target language to the native language, significantly reducing comprehension failures in Speaker LLMs.

5.4 NatLan Rectifies Knowledge Activation

In our question-answering task setup, since the Speaker LLMs only need to generate the answer options, the last hidden state for predicting the first token reflects the internal knowledge activation pattern used for answer generation, avoiding extraneous influences introduced when generating tokens in different languages. Therefore, we extract it for more in-depth analysis in knowledge activation.

As shown in Figure 4, areas of substantial overlap indicate better alignment of knowledge between the target language (Chinese) and the native language (English). Conversely, the divergences represent different knowledge activations in the Speaker LLMs. When addressing the same questions, significant differences in activation patterns are exhibited when answering directly in Chinese (Original) versus answering based explicitly on knowledge learned in English through NatLan.

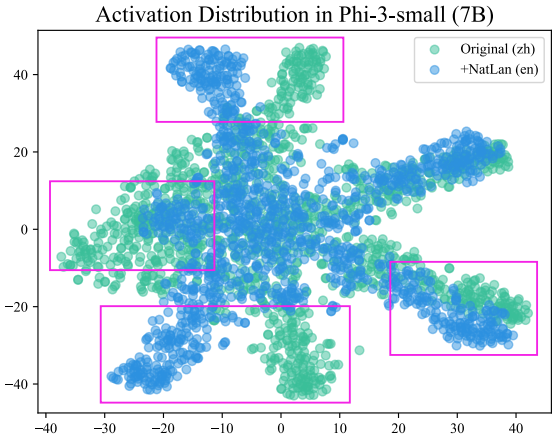

Figure 4: The distribution of activation patterns in Speaker LLMs on the C-Eval validation set, visualized through dimensionality reduction using t-SNE (Van der Maaten and Hinton, 2008). The areas with significant activation differences are highlighted in the pink box.

Considering the potential correlation between knowledge activation differences and the correctness of responses, we sampled the activations of questions and observed intriguing phenomena, as shown in Figure 6 (Top). (i) When the knowledge to answer a question is correctly activated by directly using Chinese for prompting (Original), the resulting knowledge activation shows minimal differences compared to the English knowledge activation guided by NatLan (Yellow). This confirms

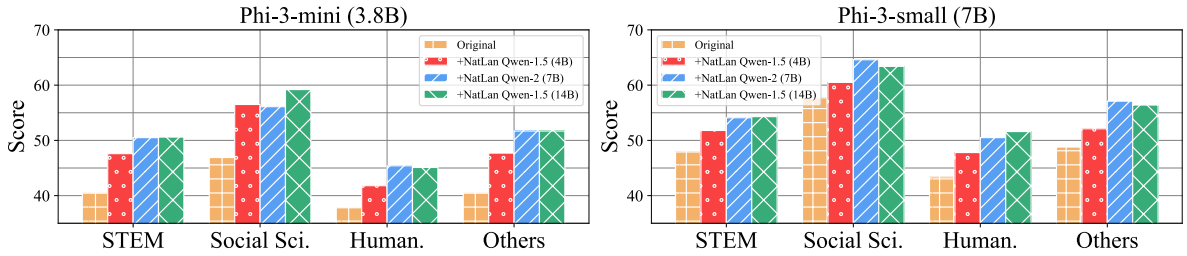

Figure 5: Performance comparison of NatLan using different Transferor LLMs in the C-Eval test sets, divided into four distinct subdomains, with Phi-3-mini (3.8B) (Left) and Phi-3-small (7B) (Right) as the Speaker LLMs.

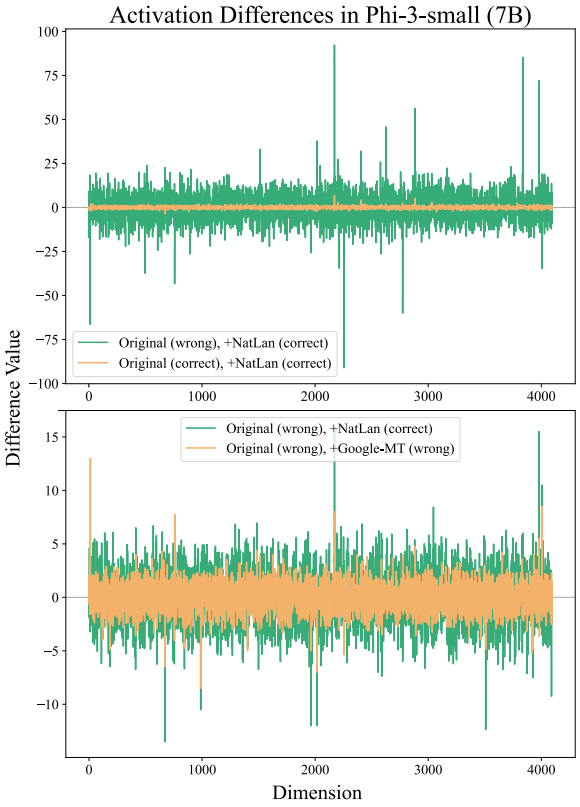

Figure 6: Activation differences between different methods for the same questions. Contents in parentheses indicate the correctness of the Speaker LLMs' responses.

that NatLan effectively simulates PNLT in Speaker LLMs, producing highly similar knowledge activations, i.e. correct answers can be generated independent of different language contexts. (ii) When PNLT cannot occur autonomously and implicitly, and direct prompting in Chinese cannot correctly activate the relevant knowledge, NatLan can explicitly guide the Speaker LLMs to adjust the activation pattern onto the correct track, resulting in significant activation differences (Green).

It is important to note that, compared to Google-MT, NatLan provides a more significant corrective effect on knowledge activation, as shown in Figure 6 (Bottom). Google-MT is insufficient to correct the knowledge activation in Speaker LLMs to the necessary extent, causing the model to still fail (Yellow). In contrast, NatLan's corrections are more substantial and appropriately directed, enabling the Speaker to produce correct responses (Green).

5.5 Impact of Transferor's Semantic Capabilities on NatLan

Furthermore, we conducted a detailed analysis to evaluate how the semantic capabilities of the Transferor LLMs in the target language affect the overall effectiveness of the proposed NatLan method.

Model	Lang.	Avg.	Avg. (Hard)
Transferor LLMs			
Qwen-1.5 (4B)	zh	60.1	42.3
Qwen-2 (7B)	zh	78.9	56.7
Qwen-1.5 (14B)	zh	74.9	58.9
Speaker LLMs			
Phi-3-mini (3.8B)	zh	41.2	36.3
+NatLan Qwen-1.5 (4B)	en	48.1	37.9
+NatLan Qwen-2 (7B)	en	50.8	39.9
+NatLan Qwen-1.5 (14B)	en	51.3	41.3
Phi-3-small (7B)	zh	49.0	41.6
+NatLan Qwen-1.5 (4B)	en	52.7	41.9
+NatLan Qwen-2 (7B)	en	56.0	43.5
+NatLan Qwen-1.5 (14B)	en	55.9	44.7
Gemma-1.1 (7B)	zh	44.4	36.3
+NatLan Qwen-1.5 (4B)	en	45.0	38.2
+NatLan Qwen-2 (7B)	en	47.7	38.6
+NatLan Qwen-1.5 (14B)	en	47.6	38.0
Mistral-0.3 (7B)	zh	42.8	32.6
+NatLan Qwen-1.5 (4B)	en	45.6	33.6
+NatLan Qwen-2 (7B)	en	48.4	35.3
+NatLan Qwen-1.5 (14B)	en	47.8	35.5
Llama-2 (7B)	zh	21.3	14.7
+NatLan Qwen-1.5 (4B)	en	25.6	18.7
+NatLan Qwen-2 (7B)	en	25.2	17.3
+NatLan Qwen-1.5 (14B)	en	27.6	18.6

Table 3: Performance comparison of NatLan using different Transferor LLMs on the C-Eval test sets.

For this purpose, Qwen series models, which exhibit strong average semantic capabilities in the target language (Chinese) and possess varying levels

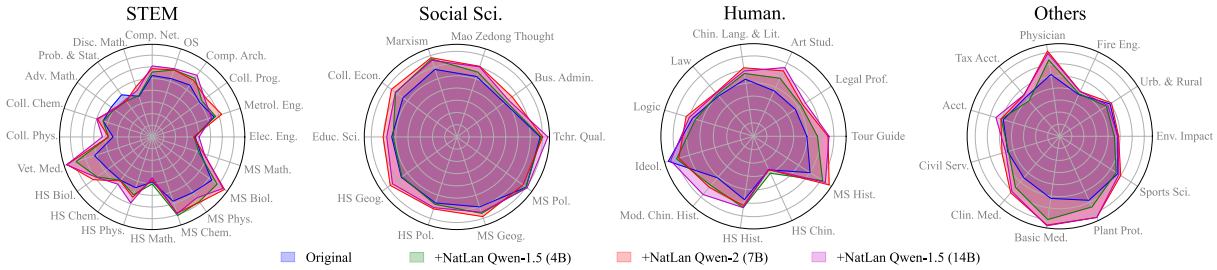

Figure 7: Performance comparison of NatLan using different Transferor LLMs in the C-Eval test sets, divided into 52 distinct disciplines, with Phi-3-small (7B) as the Speaker LLMs. More details are available in Appendix A.4.

of semantic proficiency, were deployed as Transferor LLMs. As shown in Table 3, Qwen-2 (7B) and Qwen-1.5 (14B) exhibit comparable semantic capabilities, each with their own strengths, while Qwen-1.5 (4B) has relatively weaker semantic capabilities in comparison. Furthermore, when they serve as Transferor LLMs, the relative strengths and weaknesses of their semantic capabilities are generally reflected in the varying degrees of knowledge elicitation from the Speaker LLMs.

Specifically, NatLan Qwen-2 (7B) and NatLan Qwen-1.5 (14B) generally provide comparable performance improvements. The former tends to perform better in terms of average accuracy across most models, while the latter excels in average accuracy at the hard level, aligning with their respective strengths. This confirms the pivotal role of the semantic capabilities of Transferor LLMs in the effectiveness of the proposed NatLan method.

5.6 Analysis of NatLan with Different Transferors in Various Domains

More comprehensively, we conducted a fine-grained analysis of the impact of Transferor LLMs on NatLan across four subdomains and even at the level of individual disciplines within C-Eval.

As shown in Figure 5, NatLan consistently achieved stable performance improvements across four subdomains. Moreover, the trends in performance improvements across four subdomains, which correlate with shifts in the semantic capabilities of Transferor LLMs, align closely with the analyses presented in §5.5. Additionally, it can be observed that the degree of improvements brought by NatLan is closely linked to the upper limits of performance of Speaker LLMs in their native language. This implies that when the semantic transfer challenges attributed to Transferors are alleviated, the primary determinant of NatLan's performance increasingly becomes the intrinsic knowledge level of Speaker LLMs in their native language.

Additionally, it should be noted that the degree of performance improvement NatLan delivers varies across more fine-grained disciplines. As shown in Figure 7, in the majority of disciplines, such as Veterinary Medicine (Vet. Med.) and Basic Medicine (Basic Med.), NatLan achieves substantial improvements. We believe that in such disciplines, Speaker LLMs have access to more relevant knowledge in their native language training data compared to the target language. However, in a few rare cases, such as Probability and Statistics (Prob. & Stat.) and Ideological and Moral Cultivation (Ideol.), using NatLan leads to a slight decline in performance. We believe such results are consistent with intuition, as in these disciplines, challenges arise from the complexity of translation, which can lead to semantic transfer errors, or from knowledge that is intimately associated with Chinese. These factors contribute to the diminished performance of Speaker LLMs in their native language (English).

6 Conclusion

It has been observed that MLLMs fail to answer some questions articulated in non-dominant languages, which they could address when presented in their dominant language. To mitigate this, we propose ***NatLan*** to simulate PNLT in the cognitive processes of human multilinguals. It reinterprets the effectiveness of the existing translate-first prompting methods from the perspective of PNLT in human multilinguals and suggests employing multi-MLLM collaboration to alleviate the *Language Comprehension Bottlenecks* and refine semantic transfer, thereby more effectively eliciting relevant knowledge for question-answering. The proposed NatLan achieves up to a **10.1%** average accuracy improvement in the C-Eval benchmark, as well as up to a **5.0%** increase in the hard-level subset, surpassing all top-notch related methods.

Limitations

The Speaker LLMs selected for this study all use English as their dominant language (native language). Although we aimed to assess MLLMs with various native languages, the vast majority of existing MLLMs primarily utilize English as their native language. Even if some MLLMs demonstrate stronger capabilities in other languages, they still cannot significantly outperform the performance under English prompting. Therefore, we encourage future research to explore MLLMs with different native languages other than English, or investigate whether the phenomenon of PNLT can be transferred to other non-native languages through alternative methods. Such explorations could have a profound impact on the development of applications for low-resource languages.

Furthermore, although NatLan significantly enhances the performance of MLLMs, the potential improvements attributable to NatLan are inherently limited by the capabilities of the Transferor LLMs and particularly the Speaker LLMs, where the primary bottlenecks tend to occur. Moreover, as observed in the analysis from §5.6, for a minority of disciplines, NatLan fails to enhance performance. In addition to translation errors produced by Transferor LLMs, another significant factor is that some knowledge is closely tied to specific languages, such as in the Ideology and Moral Cultivation discipline. Employing the native language to address these types of issues may not yield benefits and could instead prevent the successful recall of relevant knowledge. Therefore, we encourage future work to explore the scope of knowledge covered by various languages in MLLMs, aiming to achieve an adaptive and dynamic language switching during question-answering, specifically switching to the language that best encompasses the required knowledge for optimal knowledge elicitation.

Ethical Considerations

LLMs are prone to generating incorrect and potentially biased information. This issue becomes especially significant when LLMs are tasked with responding to sensitive questions. While NatLan enhances the performance of LLMs, it does not eliminate the issue of producing biased or incorrect statements. In light of some potential issues, this study advocates for usage under research purposes. Cautious deployment is advisable when integrating such systems into user-facing applications.

All the datasets and models used in this study are publicly available with permissible licenses. C-Eval benchmark has CC-BY-NC-SA-4.0 License [1], Phi-3-* models have MIT License [2], Qwen-1.5-* models have Tongyi-Qianwen-Research License [3], Qwen-2-* and Mistral-0.3-* models have Apache-2.0 License [4], Llama-2-* models have Llama 2 Community License [5] and Gemma-1.1-* models have Gemma Terms of Use [6].

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

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

Yiming Ai, Zhiwei He, Kai Yu, and Rui Wang. 2023. Tecs: A dataset and benchmark for tense consistency of machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1930–1941.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Anthony Chemero. 2023. Llms differ from human cognition because they are not embodied. *Nature Human Behaviour*, 7(11):1828–1829.

Yihong Dong, Xue Jiang, Zhi Jin, and Ge Li. 2024. Self-collaboration code generation via chatgpt. *ACM Trans. Softw. Eng. Methodol.*

[1] https://spdx.org/licenses/CC-BY-NC-SA-4.0

[2] https://choosealicense.com/licenses/mit

[3] https://huggingface.co/Qwen/Qwen1.5-14B-Chat/blob/main/LICENSE

[4] https://choosealicense.com/licenses/apache-2.0

[5] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf/blob/main/LICENSE.txt

[6] https://ai.google.dev/gemma/terms

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lacalle, and Mikel Artetxe. 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)*, pages 550–564.

Fei Gao, Lin Hua, Paulo Armada-da Silva, Juan Zhang, Defeng Li, Zhiyi Chen, Chengwen Wang, Meng Du, and Zhen Yuan. 2023. Shared and distinct neural correlates of first and second language morphological processing in bilingual brain. *npj Science of Learning*, 8(1):33.

Susan M Gass and Larry Selinker. 1992. *Language transfer in language learning*, volume 5. John Benjamins Publishing.

Ping Guo, Yubing Ren, Yue Hu, Yunpeng Li, Jiarui Zhang, Xingsheng Zhang, and He-Yan Huang. 2024. Teaching large language models to translate on low-resource languages with textbook prompting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15685–15697.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Jiayi Lei, Yao Fu, et al. 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*, pages 62991–63010.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André FT Martins, François Yvon, et al. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Haoqiang Kang, Terra Blevins, and Luke Zettlemoyer. 2024. Translate to disambiguate: Zero-shot multilingual word sense disambiguation with pretrained language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1562–1575.

Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2024. Improving in-context learning of multilingual generative language models with cross-lingual alignment. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8051–8069.

Kelly Marchisio, Wei-Yin Ko, Alexandre Bérard, Théo Dehaze, and Sebastian Ruder. 2024. Understanding and mitigating language confusion in llms. *arXiv preprint arXiv:2406.20052*.

Dominique Savio Nsengiyumva, Celestino Oriikiriza, and Sarah Nakijoba. 2021. Cross-linguistic transfer and language proficiency in the multilingual education system of burundi: What has the existing literature so far discovered?. *Indonesian Journal of English Language Teaching and Applied Linguistics*, 5(2):387–399.

Yuqi Ren, Renren Jin, Tongxuan Zhang, and Deyi Xiong. 2024. Do large language models mirror cognitive language processing? *arXiv preprint arXiv:2402.18023*.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, et al. 2024. The prompt report: A systematic survey of prompting techniques. *arXiv preprint arXiv:2406.06608*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Sunayana Sitaram, Monojit Choudhury, Barun Patra, Vishrav Chaudhary, Kabir Ahuja, and Kalika Bali. 2023. Everything you need to know about multilingual llms: Towards fair, performant and reliable models for languages of the world. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, pages 21–26.

Yashar Talebirad and Amirhossein Nadiri. 2023. Multi-agent collaboration: Harnessing the power of intelligent llm agents. *arXiv preprint arXiv:2306.03314*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti

Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Zhaopeng Tu, Yang Liu, Lifeng Shang, Xiaohua Liu, and Hang Li. 2017. Neural machine translation with reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting palm for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. *arXiv preprint arXiv:2402.10588*.

Yan Jing Wu, Koji Miwa, and Haoyun Zhang. 2022. Cognitive factors in bilingual language processing. *Frontiers in Psychology*, 13.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Yuqi Ye, and Hanwen Gu. 2024. A survey on multilingual large language models: Corpora, alignment, and bias. *arXiv preprint arXiv:2404.00929*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Tao Zeng, Chen Chen, and Jiashu Guo. 2022. First language translation involvement in second language word processing. *Frontiers in Psychology*, 13:986450.

Lichao Zhu, Maria Zimina, Maud Bénard, Behnoosh Namdar, Nicolas Ballier, Guillaume Wisniewski, and Jean-Baptiste Yunès. 2023. Investigating techniques for a deeper understanding of neural machine translation (nmt) systems through data filtering and fine-tuning strategies. In *Proceedings of the Eighth Conference on Machine Translation*, pages 275–281.

A Appendix

A.1 Implementation Details

In this study, to minimize randomness introduced during the sampling process, we standardized the decoding method across all MLLMs to greedy decoding, which includes both Transferor and Speaker LLMs. Furthermore, all MLLMs involved in the experiments are open-source models of the Instruct/Chat version: Phi-3-mini (3.8B) [7], Phi-3-small (7B) [8], Gemma-1.1 (7B) [9], Mistral-0.3 (7B) [10], Llama-2 (7B) [11], Qwen-1.5 (4B) [12], Qwen-2 (7B) [13], and Qwen-1.5 (14B) [14].

At the same time, as we deployed Transferor LLMs within NatLan that required designing translation prompts, we used GPT-4o [15] to translate the dev sets of various disciplines in the C-Eval benchmark from Chinese to English. This ensures the quality of the translations in the prompts, with each discipline's dev set containing five examples, allowing us to construct five-shot translation prompts for each discipline. We also created five-shot Q&A prompts using the C-Eval dev sets. In practical applications, we provide the MLLMs with prompts corresponding to the discipline currently being tested, thus maximizing the elicitation of their domain-specific knowledge.

Since the Transferor LLMs and Speaker LLMs used in the proposed NatLan method are required to undertake distinct processes, the former are required to translate questions from the target language to the native language, while the latter are required to provide answers based on the translated questions in the native language. Therefore, they use different sets of prompts. First, we report the details of the translation prompts used in our experiments as follows:

```
<System Prompts>
You are a professional Chinese-English translator. Translation rules: Proper nouns in English or Chinese need to be retained without translation, retain the original meaning to the greatest extent, and follow the original format in the translation process.

<Original Question Prompts>
Now help me translate the following sentence into English, only return the translated sentence, the original sentence is:
Question:
{original example['question']}
Choices:
A. {original example['choice A']}
B. {original example['choice B']}
C. {original example['choice C']}
D. {original example['choice D']}
Answer:

<Translated Question Prompts>
Question:
{translated example['question']}
Choices:
A. {translated example['choice A']}
B. {translated example['choice B']}
C. {translated example['choice C']}
D. {translated example['choice D']}
Answer:
```

Furthermore, we report the details of the Q&A prompts used in our experiments as follows:

```
<System Prompts>
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.

<Question Prompts>
Question:
{translated example['question']}
Choices:
A. {translated example['choice A']}
B. {translated example['choice B']}
C. {translated example['choice C']}
D. {translated example['choice D']}
```

---

[7] https://huggingface.co/microsoft/Phi-3-mini-128k-instruct

[8] https://huggingface.co/microsoft/Phi-3-small-128k-instruct

[9] https://huggingface.co/google/gemma-1.1-7b-it

[10] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[11] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

[12] https://huggingface.co/Qwen/Qwen1.5-4B-Chat

[13] https://huggingface.co/Qwen/Qwen2-7B-Instruct

[14] https://huggingface.co/Qwen/Qwen1.5-14B-Chat

[15] API version: gpt-4o-2024-05-13

```
Answer:

<Answer Prompts>
{example['answer']}
```

A.2 Comparative Analysis of Chinese-to-English Translation Cases

As a supplement to Table 2, we report a more detailed comparative analysis of Chinese-to-English translation cases between Google-MT and the proposed NatLan in Table 4.

As shown in Table 4, in the examples from the first two rows, NatLan provides more semantically coherent translations. This coherent semantic description enables Speaker LLMs to more easily understand the relationship between the question and the answer. In the cases presented in the latter two rows, NatLan delivers translations with greater semantic accuracy. For these two questions pertaining to the High School Chemistry discipline, the enriched semantic comprehension of the Transferor LLMs enables NatLan to generate terminology that aligns more closely with domain-specific usage. For instance, it translates to "combusted", which is preferred in chemical contexts, rather than the general term "burned", and "Reactivity" instead of "The intensity of reaction".

This comparative study further confirms the superiority of NatLan over methods using external NMT systems like Google-MT in terms of semantic transfer during translation. The effective semantic conveyance provided by NatLan enhances the understanding of questions by Speaker LLMs and facilitates knowledge elicitation, thereby yielding superior practical performance.

A.3 Sampled Cases Used for Knowledge Activation

As a supplement to §5.4, we report cases used to measure differences in knowledge activation in this experiment, which were sampled from the C-Eval val/test sets. Detailed content is shown in Table 5.

It should be noted that the reason for excluding the comparison of the Self-Translation method in the experiments for Figure 6 is due to its inability to guarantee basic accuracy in the semantic transfer process. This method may generate incomplete translated questions, preventing the Speaker LLMs from accessing complete question information. Such issues can greatly disrupt overall knowledge activation, making comparisons of activation differences with this method meaningless. If complete question information cannot be conveyed to the Speaker LLMs, it is akin to the Speaker LLMs addressing an entirely different question, thereby rendering its knowledge activation incomparable.

Additionally, as the case shown in the second row of Table 5 is mathematical and lacks substantial textual content, and given that our goal is to demonstrate that the knowledge activation provided by NatLan can unlock the limitations posed by different language contexts on the effective application of knowledge in Speaker LLMs, this case may not effectively illustrate the differences between target language (Chinese) and native language (English) prompt contexts.

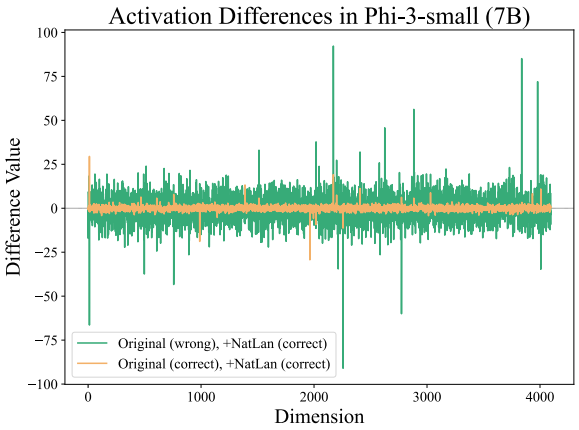

Figure 8: Activation differences between different methods for the same questions. Contents in parentheses indicate the correctness of the Speaker LLMs' responses. This is a supplement to Figure 6 (Top).

Therefore, we provide a supplementary case in Table 6, which contains more extensive textual content (rather than mathematical formulas) to demonstrate more convincingly whether prompts are delivered directly in the target language (Chinese) to utilize implicit Positive Native Language Transfer (PNLT), or explicitly guide PNLT through NatLan at the native language (English) level, when both methods can accurately respond, the patterns of knowledge activation in Speaker LLMs are extremely similar, as depicted in Figure 8. This knowledge activation similarity shows that the activation pattern is independent of the specific language contexts of the prompts, and confirms that our proposed NatLan can effectively simulates PNLT in its performance. Since NatLan is designed to explicitly promote PNLT, this further confirms that NatLan can provide the correct knowledge activation patterns, thus successfully unlocking the

limitations posed by different language contexts in Speaker LLMs and effectively eliciting the corresponding knowledge.

A.4 Analysis of NatLan with Different Transferors in Various Domains

As a supplement to Figure 7, we present a detailed performance analysis of NatLan, employing three different Transferor LLMs applied to various Speaker LLMs, across specific disciplines. These include Phi-3-mini (3.8B) in Figure 9, Gemma-1.1 (7B) in Figure 10, Mistral-0.3 (7B) in Figure 11, and Llama-2 (7B) in Figure 12.

As shown in these figures, NatLan has provided widespread and consistent performance improvements across all Speaker LLMs, with only minor performance declines in a very few disciplines. Furthermore, across each Speaker LLM, performance improvements and the disciplines where declines occur vary due to differences in performance preferences, the proportion of different language data in the training corpora, and variations in data sources and quality. This variation highlights that the knowledge elicitation facilitated by NatLan, aside from the influence of Transferor LLMs, is primarily dependent on the capabilities of the Speaker LLMs in their native languages.

Additionally, it is important to note that since NatLan relies heavily on the collaboration of MLLMs, it also demands a high level of compliance with instructions from the MLLMs. As shown in Figure 12, Llama-2 (7B), compared to other Speaker LLMs, has relatively weaker instruction-following capabilities. Consequently, it is more prone to producing answers that do not conform to the prescribed format during testing. We applied a strict evaluation criterion in these instances, considering any output that did not meet the established format as incorrect. Thus, the performance improvements brought about by NatLan using different Transferor LLMs on Llama-2 (7B) show relatively greater variability. However, from a holistic perspective, disregarding the variations between different Transferor LLMs, NatLan still manages to provide stable performance improvements for Llama-2 (7B). This further confirms the superiority of the proposed NatLan method.

Furthermore, we have reported the detailed performance evaluation scores of NatLan and top-notch related methods in Table 7 for all settings, as a supplement to Table 1 and Table 3

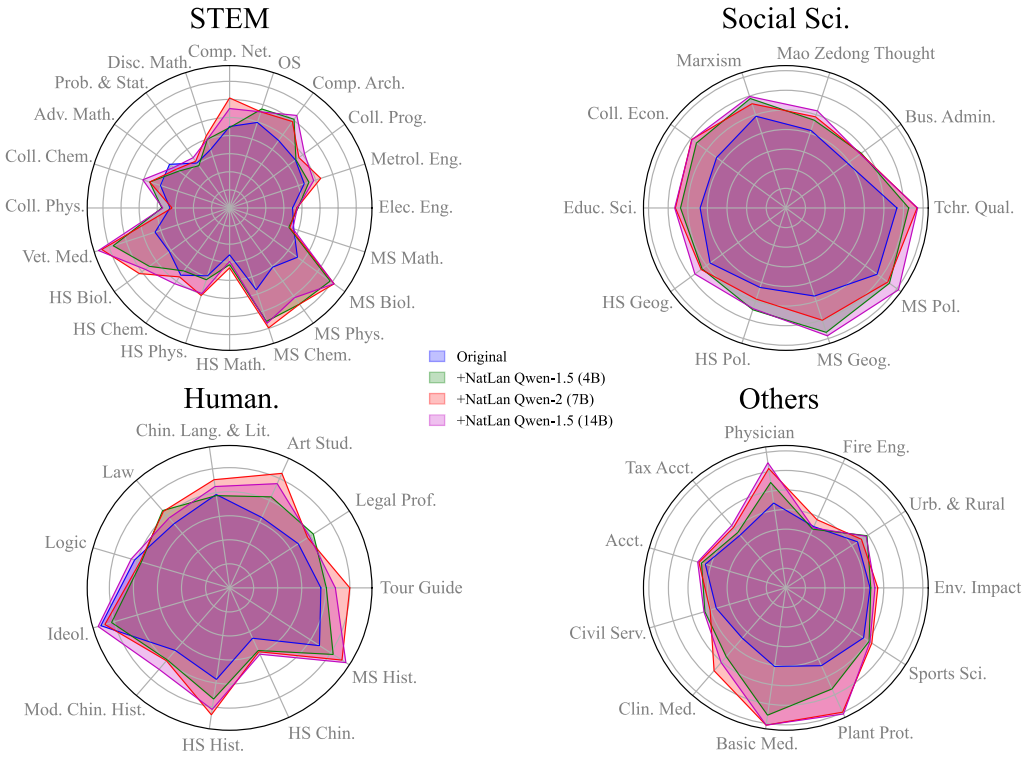

Figure 9: Performance comparison of NatLan using different Transferor LLMs in the C-Eval test sets, divided into 52 distinct disciplines, with Phi-3-mini (3.8B) as the Speaker LLMs.

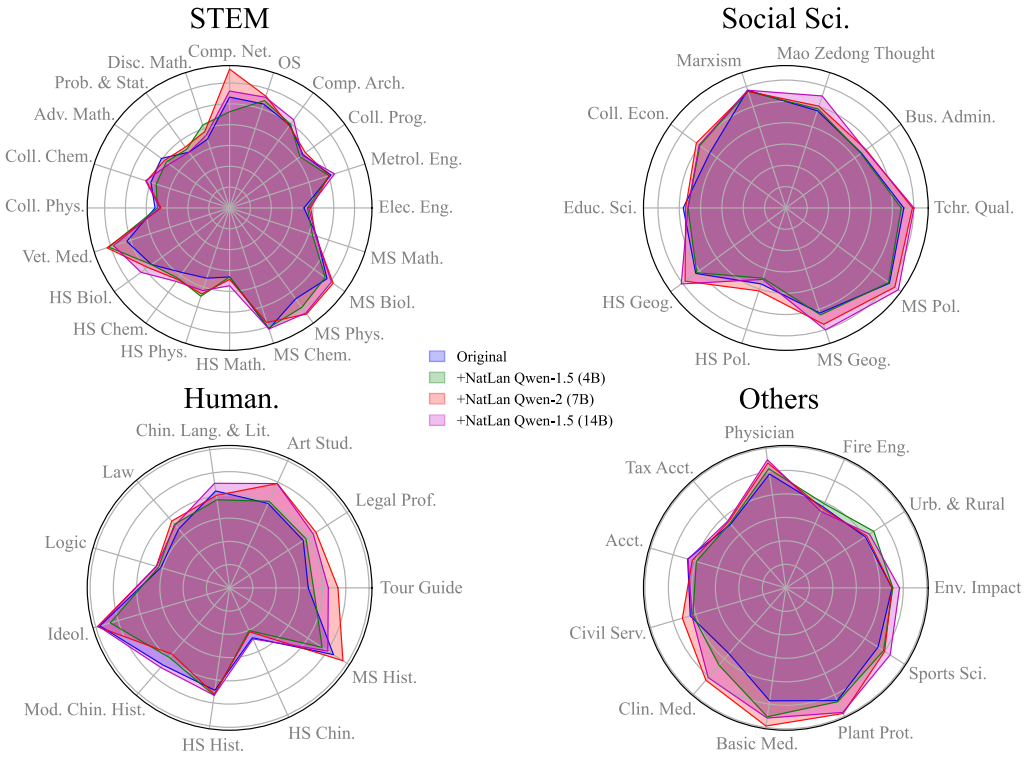

Figure 10: Performance comparison of NatLan using different Transferor LLMs in the C-Eval test sets, divided into 52 distinct disciplines, with Gemma-1.1 (7B) as the Speaker LLMs.

Original Question	Google-MT Trans. Question	NatLan Trans. Question	Answers
某计算机的指令系统中共有101条不同的指令，采用微程序控制方式时，控制存储器中具有的微程序数目至少是____。 A. 100 B. 102 C. 103 D. 104	There are 101 different instructions in the instruction system of a certain computer. When using microprogram control, the number of microprograms in the control memory is at least ____. A. 100 B. 102 C. 103 D. 104	In a computer's instruction set with a total of 101 different instructions, the minimum number of microprograms required in the control memory when using microprogram control is ____. A. 100 B. 102 C. 103 D. 104	Original: B +Google-MT: C +NatLan : B True Label : B
迁都后对帕朗卡拉亚的影响有____。 A. 有利于缓解住房紧张问题 B. 有利于缓解交通拥堵状况 C. 有利于环境污染的治理 D. 基础设施的完善	The impact of the capital relocation on Palangkaraya is____. A. It is conducive to alleviating housing shortages B. It is conducive to alleviating traffic congestion C. It is conducive to the control of environmental pollution D. The improvement of infrastructure	The impact of the capital relocation on Palangkaraya would include____. A. Alleviating housing shortages B. Alleviating traffic congestion C. Facilitating environmental pollution control D. Improvement of infrastructure	Original: A +Google-MT: C +NatLan : D True Label: D
下列各物质完全燃烧，产物除二氧化碳和水外，还有其他物质的是____。 A. 甲烷 B. 乙烯 C. 氯乙烯 D. 乙醇	When the following substances are completely burned, the products include carbon dioxide and water, and other substances____. A. Methane B. Ethylene C. Vinyl chloride D. Ethanol	Among the following substances, which one, when completely combusted, produces products other than carbon dioxide and water?____. A. Methane B. Ethylene C. Vinyl chloride D. Ethanol	Original: C +Google-MT: D +NatLan : C True Label : C
下列有关$NaHCO_3$与Na_2CO_3的说法中不正确的是____。 A. 在水中溶解性：$Na_2CO_3 < NaHCO_3$ B. 与相同浓度酸反应的剧烈程度：$Na_2CO_3 < NaHCO_3$ C. 热稳定性：$Na_2CO_3 < NaHCO_3$ D. 二者间在一定条件下可相互转化	Which of the following statements about $NaHCO_3$ and Na_2CO_3 is incorrect____. A. Solubility in water: $Na_2CO_3 < NaHCO_3$ B. The intensity of the reaction with the same concentration of acid: $Na_2CO_3 < NaHCO_3$ C. Thermal stability: $Na_2CO_3 < NaHCO_3$ D. The two can be converted into each other under certain conditions	Which of the following statements about $NaHCO_3$ and Na_2CO_3 is incorrect?____. A. Solubility in water: $Na_2CO_3 < NaHCO_3$ B. Reactivity with equal concentration acids: $Na_2CO_3 < NaHCO_3$ C. Thermal stability: $Na_2CO_3 < NaHCO_3$ D. They can transform into each other under certain conditions	Original: B +Google-MT: B +NatLan : C True Label : C

Table 4: Supplementary comparative analysis of Chinese-to-English translation cases, with cases sampled from the C-Eval test sets. The contents marked in green indicate semantic accuracy/coherence in the translation or correctness in the response of Phi-3-mini (3.8B) , while those marked in red indicate errors.

Original Question	Google-MT Trans. Question	NatLan Trans. Question	Answers
《尼伯龙根的指环》是____的作品。 A. 布拉姆斯 B. 肖邦 C. 威尔第 D. 瓦格纳	–	"Der Ring des Nibelungen" is the work of ____. A. Brahms B. Chopin C. Verdi D. Wagner	Original: C +NatLan : D True Label : D
求极限: $\lim_{x\to 0}\frac{\int_{x^2}^{x}\frac{\sin(xt)}{t}\mathrm{d}t}{x^2}$ =____ A. $\frac{5}{6}$ B. 1 C. $\frac{7}{6}$ D. $\frac{4}{3}$	–	Find the limit: $\lim_{x\to 0}\frac{\int_{x^2}^{x}\frac{\sin(xt)}{t}\mathrm{d}t}{x^2}$ =____ A. $\frac{5}{6}$ B. 1 C. $\frac{7}{6}$ D. $\frac{4}{3}$	Original: B +NatLan : B True Label : B
间址寻址第一次访问内存所得到的信息经____传送到MDR。 A. 数据总线 B. 地址总线 C. 控制总线 D. 总线控制器	The information obtained by indirect addressing when accessing the memory for the first time is transmitted to MDR via____. A. Data bus B. Address bus C. Control bus D. Bus controller	The information obtained from the first memory access using indirect addressing is transmitted to the MDR via____. A. data bus B. address bus C. control bus D. bus controller	Original: B +Google-MT: B +NatLan : A True Label : A

Table 5: Cases sampled from the C-Eval val/test sets for knowledge activation analysis in §5.4.

Original Question	Google-MT Trans. Question	NatLan Trans. Question	Answers
某应急避难场所安装消防应急照明和疏散指示系统等消防设施，对于面积大于____的防火分区应单独设置应急照明配电箱或应急照明分配电装置。 A. $1000m^2$ B. $2000m^2$ C. $2500m^2$ D. $3000m^2$	–	In a certain emergency shelter, if fire safety lighting and evacuation sign systems are installed, separate emergency lighting distribution boxes or emergency lighting distribution devices should be provided for the fire protection zone with an area greater than____. A. $1000m^2$ B. $2000m^2$ C. $2500m^2$ D. $3000m^2$	Original: B +NatLan : B True Label : B

Table 6: The supplemental case in Figure 8, which is provided to further elucidate §5.4.

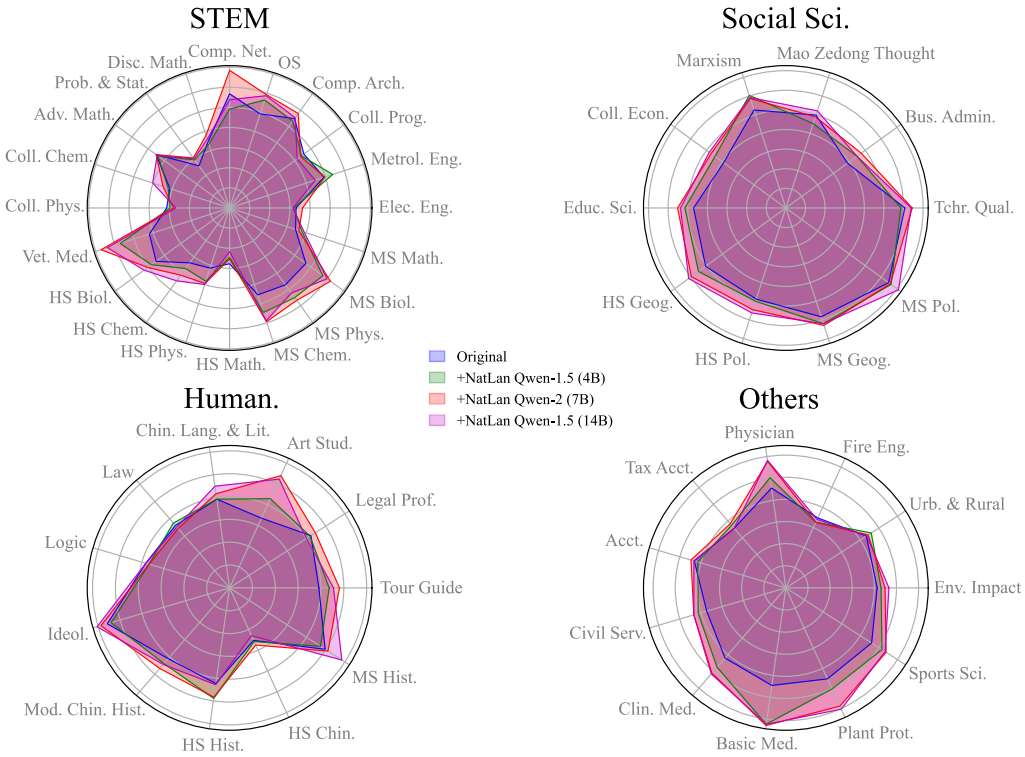

Figure 11: Performance comparison of NatLan using different Transferor LLMs in the C-Eval test sets, divided into 52 distinct disciplines, with Mistral-0.3 (7B) as the Speaker LLMs.

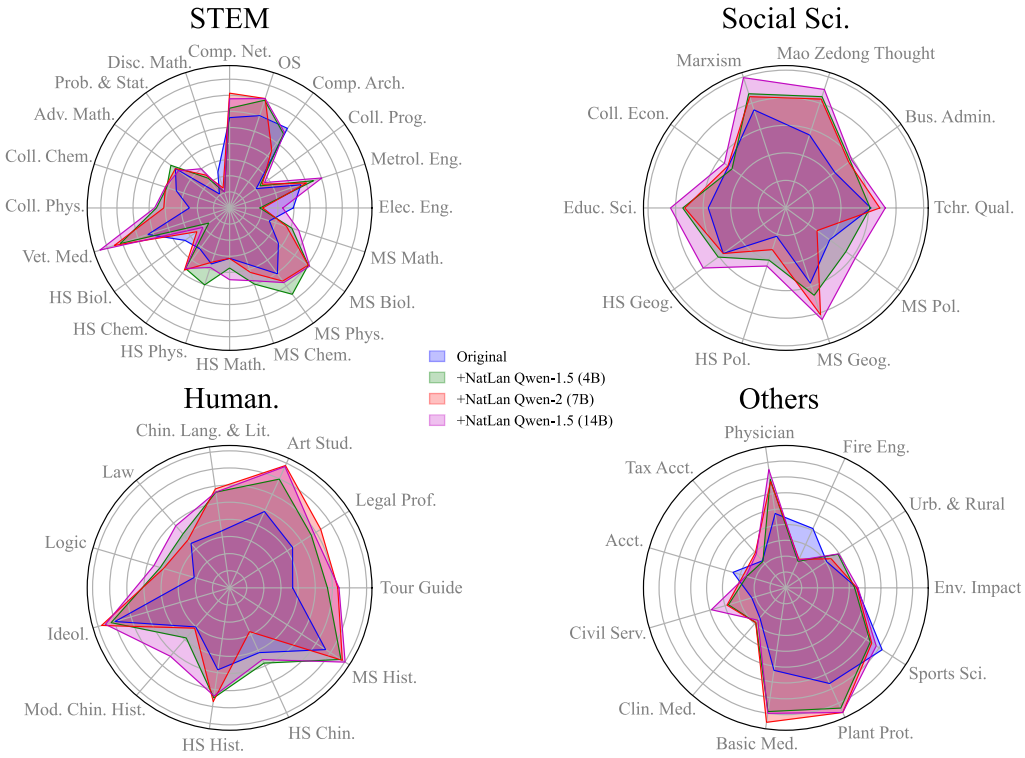

Figure 12: Performance comparison of NatLan using different Transferor LLMs in the C-Eval test sets, divided into 52 distinct disciplines, with Llama-2 (7B) as the Speaker LLMs.

Model	Lang.	STEM	Social Sci.	Human.	Others	Avg.	Avg. (Hard)
Transferor LLMs							
Qwen-1.5 (4B)	zh	55.2	73.7	62.0	54.9	60.1	42.3
Qwen-2 (7B)	zh	71.4	88.7	80.9	81.8	78.9	56.7
Qwen-1.5 (14B)	zh	69.9	86.7	76.3	71.6	74.9	58.9
Speaker LLMs							
Phi-3-mini (3.8B)	zh	40.5	46.9	37.8	40.5	41.2	36.3
+Self-Translation	en	44.8	48.9	37.4	43.7	43.8	37.7
+Google-MT	en	50.1	56.3	46.7	51.4	50.9	40.4
+NatLan Qwen-1.5 (4B)	en	47.6	56.5	41.8	47.7	48.1	37.9
+NatLan Qwen-2 (7B)	en	50.5	56.1	45.4	51.7	50.8	39.9
+NatLan Qwen-1.5 (14B)	en	50.6	59.2	45.1	51.7	51.3	41.3
Phi-3-small (7B)	zh	47.9	57.7	43.4	48.8	49.0	41.6
+Self-Translation	en	51.4	59.6	46.4	51.8	52.0	42.1
+Google-MT	en	54.0	63.5	51.0	56.5	55.7	42.7
+NatLan Qwen-1.5 (4B)	en	51.8	60.5	47.8	52.1	52.7	41.9
+NatLan Qwen-2 (7B)	en	54.1	64.6	50.5	57.1	56.0	43.5
+NatLan Qwen-1.5 (14B)	en	54.3	63.4	51.6	56.4	55.9	44.7
Gemma-1.1 (7B)	zh	44.6	49.9	40.1	43.6	44.4	36.3
+Self-Translation	en	42.3	44.9	38.2	42.3	41.9	33.9
+Google-MT	en	47.5	50.4	41.9	46.5	46.7	38.2
+NatLan Qwen-1.5 (4B)	en	45.5	49.9	39.1	45.4	45.0	38.2
+NatLan Qwen-2 (7B)	en	47.5	53.3	43.0	47.5	47.7	38.6
+NatLan Qwen-1.5 (14B)	en	47.1	53.7	43.1	47.5	47.6	38.0
Mistral-0.3 (7B)	zh	40.5	51.1	40.3	41.7	42.8	32.6
+Self-Translation	en	35.5	36.1	31.6	35.6	34.8	30.9
+Google-MT	en	44.5	55.9	45.8	49.2	48.0	33.3
+NatLan Qwen-1.5 (4B)	en	43.4	53.9	42.0	45.8	45.6	33.6
+NatLan Qwen-2 (7B)	en	46.5	56.5	44.7	48.4	48.4	35.3
+NatLan Qwen-1.5 (14B)	en	44.8	57.3	44.1	48.4	47.8	35.5
Llama-2 (7B)	zh	18.9	25.9	21.6	20.9	21.3	14.7
+Self-Translation	en	8.7	8.7	11.5	9.6	9.6	10.3
+Google-MT	en	19.9	31.9	29.9	24.9	25.4	15.1
+NatLan Qwen-1.5 (4B)	en	22.3	31.8	28.4	23.2	25.6	18.7
+NatLan Qwen-2 (7B)	en	21.4	30.8	28.3	24.0	25.2	17.3
+NatLan Qwen-1.5 (14B)	en	23.3	36.3	30.4	24.8	27.6	18.6

Table 7: Detailed performance scores (accuracy) of NatLan and top-notch related methods under different configurations on the C-Eval test sets. The meanings assigned to the different colors correspond to those in Table 1.