NatLan: Native Language Prompting Facilitates Knowledge Elicitation Through Language Trigger Provision and Domain Trigger Retention

Anonymous ACL submission

Abstract

Multilingual large language models (MLLMs) do not perform as well when answering questions in non-dominant languages as they do in their dominant languages. Although existing translate-then-answer methods alleviate this issue, the mechanisms behind their effectiveness remain unclear. In this study, we analogize the dominant language of MLLMs to the native language of humans and use two human cognitive features: the **Language Trigger** (**LT**) and the **Domain Trigger** (**DT**), to interpret the mechanisms behind translate-then-answer methods. This reveals that while sufficient LTs are provided by these methods, there remains a deficiency in DT retention. To mitigate this issue, we propose **Nat**ive **Lan**guage Prompting (**NatLan**), employing a Multi-MLLM collaboration strategy and introducing an additional role-enhanced domain-specific MLLM with stronger multilingual understanding capabilities as the translator. Across five language QA benchmarks, NatLan achieves up to a **31.28%** improvement in accuracy and, compared to existing state-of-the-art methods, provides comparable or greater retention of DTs in up to **87%** of cases. Our code is available at https://github.com/AnonyNLP/NatLan.

1 Introduction

Recent research has observed that multilingual large language models (MLLMs) (Brown et al., 2020; Achiam et al., 2023) fail to answer certain questions in non-dominant languages that they can correctly answer when posed in their dominant languages (i.e., the language with the highest proportion during training[1]) (Zhang et al., 2023; Huang et al., 2023a; Etxaniz et al., 2024). Currently, two main *translate-then-answer* (Schulhoff et al., 2024) methods are employed to resolve this issue. Both involve translating the question into the MLLM's dominant language before it answers, thereby facilitating the use of the MLLM's richer dominant language knowledge. One method allows the MLLM itself to handle the translation (Zhang et al., 2023; Huang et al., 2023a; Etxaniz et al., 2024), while the other uses an external Neural Machine Translation (NMT) system (Shi et al., 2022). We refer to the former as *Self-Translation* and the latter as *NMT-Translation*. However, what are the mechanisms behind the effectiveness of translate-then-answer methods, and how do they affect the elicitation of dominant language knowledge in MLLMs?

To investigate these issues, we analogize the dominant language of MLLMs to the native language of humans and interpret the mechanisms of the translate-then-answer methods through two key features observed in human cognitive processes, which we respectively term as **Language Trigger (LT)** and **Domain Trigger (DT)**. The former reduces the cognitive load of understanding the non-native questions by translating them into the native language (Wu et al., 2022; Zeng et al., 2022; Gao

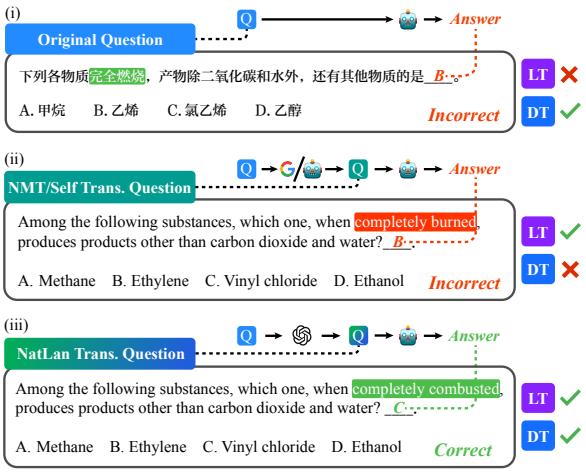

Figure 1: The presence of Language Triggers (LTs) and Domain Triggers (DTs) in questions processed by different methods when addressing non-native language QA. The same icon represents the same question/model.

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

et al., 2023; Del Maschio et al., 2022), while the latter narrows the scope of knowledge elicitation by employing domain-specific terms (Baker and Levin, 2015; Acheampong et al., 2016; Tong et al., 2020; Liu et al., 2022). Together, these cognitive triggers facilitate the elicitation of knowledge when humans answer non-native language questions.

Accordingly, we analyze the occurrences of these cognitive triggers in existing methods. As depicted in Figure 1 (i), when the MLLM directly answers a question in the non-native language (Chinese), it lacks appropriate LTs. Even if correct DTs are present in the human-constructed original question, the method fails to accurately elicit the knowledge within the MLLM. As depicted in Figure 1 (ii), although the existing translate-then-answer methods are capable of translating and providing sufficient LTs (questions in the native language, English) for the MLLM, their inherent deficiencies lead to inadequate translation of fine-grained specialized terms (DTs) (Tu et al., 2017; Zhu et al., 2023; Ai et al., 2023), translating the specialized chemical term "completely combusted" as the more generic "completely burned". Consequently, the absence of DTs leads to the failure of the relevant knowledge elicitation in MLLMs.

To mitigate this issue observed in existing translate-then-answer methods, we propose **Nat**ive **Lan**guage Prompting (**NatLan**), which employs a role-enhanced domain-specific Multi-MLLM collaboration strategy (Talebirad and Nadiri, 2023; Dong et al., 2024), comprising the Translator LLM and the Speaker LLM. The former translates questions from a non-native language to the native language of the Speaker LLM, while the latter answers questions based on the translated questions. As depicted in Figure 1 (iii), as a novel translate-then-answer method, NatLan provides sufficient LTs while maximally retaining DTs during translation, further facilitating the elicitation of relevant knowledge during non-native language QA.

Our contributions are primarily as follows: **(i) Cognition-inspired Interpretation:** We employ two cognitive triggers, LTs and DTs, to interpret the limitations of existing translate-then-answer methods in non-native language QA as the insufficiency of DTs. **(ii) Effective Remediation:** We propose NatLan to mitigate this issue, achieving up to a **31.28%** improvement in accuracy across five non-native language QA benchmarks and surpassing all top-notch methods. **(iii) Multi-level Confirmation:** We confirmed that NatLan provides comparable or greater retention of DTs in up to **87%** of cases at the input level compared to the state-of-the-art NMT-Translation method, and further demonstrate at the activation level that a higher number of DTs results in more salient knowledge activation, which helps correct the MLLM's answers. This validates the rationality of using cognitive triggers to interpret the translate-then-answer process.

2 Related Work

Language Triggers and Domain Triggers in Cognitive Processes. For human multilinguals, there are two key features that assist in leveraging their native language knowledge to correctly answer questions in a less proficient non-native language. The first one aligns with the human tendency to prioritize thinking in the native **language** by converting non-native language questions into the corresponding native language (Wu et al., 2022; Zeng et al., 2022; Gao et al., 2023), thereby reducing the cognitive load associated with understanding the question (Del Maschio et al., 2022). The second one aligns with the fact that different **domains** of knowledge have their own specialized terminology. Appropriate use of domain-specific terms enables humans to more easily associate knowledge within that field (Baker and Levin, 2015; Acheampong et al., 2016; Tong et al., 2020; Liu et al., 2022), thereby narrowing the scope of knowledge elicitation to a specific domain with greater precision.

In this study, we refer to these cognitive features as the **Language Trigger** (**LT**) and the **Domain Trigger** (**DT**), respectively. Subsequently, we use them to interpret translate-then-answer methods and propose NatLan to address the limitations of existing methods with respect to these two triggers.

Translate-then-answer Prompting. Translate-then-answer Prompting (Schulhoff et al., 2024) aims to leverage the strength of MLLMs in their dominant/native language (English). One category of these methods falls under Self-Translation (Zhang et al., 2023; Huang et al., 2023a; Etxaniz et al., 2024), which requires MLLMs themselves to perform the non-native to native language translation, before answering questions. However, if the model has poor capabilities in the non-native language, it may not capture the DTs in the original questions accurately. 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

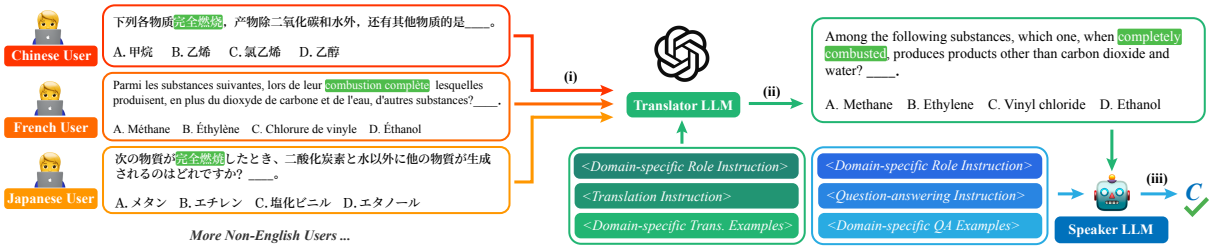

Figure 2: Non-native language question-answering workflow of NatLan. (i) Non-English users issue queries. (ii) The Translator LLM translates the non-native language questions into the native language (English) of the Speaker LLM. (iii) The Speaker LLM answers the native language question. More details are available in Appendix A.1.

et al., 2024), NMT systems are translation-only and lack domain-specific knowledge, resulting in overly generic and literal translations (Tu et al., 2017; Zhu et al., 2023; Ai et al., 2023). Overall, while existing methods can provide sufficient LTs through translation, they inevitably lose some of the DTs in the original questions during the translation process, leading to suboptimal or even detrimental effects on the subsequent answers of MLLMs.

Our proposed NatLan incorporates an additional role-enhanced domain-specific MLLM with stronger multilingual capabilities as the translator. This approach mitigates the limitations of Self-Translation, which is constrained by its own capabilities, and NMT-Translation, which lacks sufficient domain knowledge. As a result, it ensures the provision of LTs while maximizing the retention of DTs during the translate-then-answer process.

3 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 and Nadiri, 2023; Dong et al., 2024). In this study, the translate-then-answer process is inherently divided into two sub-processes: (i) translating and (ii) answering. Accordingly, 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:

(i) **Translator** 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.

(ii) **Speaker** requires MLLMs that excel in their native language (dominant 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.

4 Native Language Prompting

Utilizing our constructed Role-Enhanced Multi-MLLM Collaboration framework, we further proposed **Nat**ive **Lan**guage Prompting (**NatLan**) to provide rich Language Triggers (LTs) while maximally retaining Domain Triggers (DTs) during the translate-then-answer process.

As depicted in Figure 2, in addition to constructing the domain-specific role instruction, we also injected domain-specific context through few-shot (5-shot) examples to achieve two objectives: **(i) Individual Enhancement:** facilitating the recall of domain-specific knowledge by each MLLM in this framework through domain-specific role prompting. **(ii) Joint Enhancement:** maximizing the retention of DTs by the Translator LLM during the translation process, allowing the rich DTs in the translated question to be passed on to the Speaker LLM in a more easily understandable form, i.e. in the native language of the Speaker LLM. In this process, the domain-specific information captured by the Translator LLM[2] is explicitly conveyed to the Speaker LLM through the combination[3] of DTs and LTs, further eliciting the relevant knowledge in the Speaker LLM and thereby achieving a collaborative joint enhancement effect.

[2]Such information-capturing capability is the strength of the Translator LLM, as this information cannot be captured independently by the Speaker LLM.

[3]Domain-specific terms (DTs) that are translated into the native language (LTs).

5 Experiments

To explore the improvements that NatLan brings to knowledge elicitation, we selected question-answering (QA) as the evaluation task because it clearly indicates whether the relevant knowledge in the MLLMs has been correctly elicited. Since the native language (dominant langauge) of nearly all mainstream MLLMs is English, we have selected *English (en)* as the native language in this study.

Considering the need to ensure diversity and distinctiveness among languages, we have selected five representative languages as our non-native target languages (hereafter referred to as the target language): *Arabic (ar), Chinese (zh), French (fr), German (de), and Japanese (ja)*. Among these, French and German belong to the Indo-European language family, similar to English, while Arabic (Afro-Asiatic language family), Chinese (Sino-Tibetan language family), and Japanese (Japonic language family) are from different language families, showing significant differences from English.

Dataset. Based on the aforementioned five target languages, we have selected the Multilingual MMLU (MMMLU) benchmark[4] of QA to assess the knowledge elicited from MLLMs. MMMLU consists of expert-translated versions of the MMLU benchmark (Hendrycks et al., 2021) in different languages. Each language version comprises 14,079 multiple-choice questions from 57 disciplines.

Additionally, we have also selected C-Eval Chinese benchmark (Huang et al., 2023b) to conduct further ablation studies and case studies, for the linguistic differences between Chinese and English are significant and both languages are sufficiently representative. C-Eval is also a multidisciplinary QA benchmark, containing 13,948 multiple-choice questions from 52 disciplines.

NatLan Setup. In the proposed NatLan, we selected GPT-4o-mini as a universal translator to accomplish translations from the five target languages to the native language, for its comprehensive multilingual understanding capabilities. Additionally, to analyze the effects of the Translator LLM with varying capabilities on NatLan, we chose the Qwen MLLMs (Bai et al., 2023) as Chinese-to-English translators, with 4B, 7B, and 14B parameters, for their leading Chinese comprehension capabilities.

Furthermore, we selected five representative MLLMs with the capability to understand the five target languages to serve as Speakers, including 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 Translator 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** (Zhang et al., 2023; Huang et al., 2023a; Etxaniz et al., 2024), 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[5] (API) as the translator and MLLMs as the Speaker LLM.

It is important to note that the requirement for Speaker LLMs to possess the five target languages comprehension abilities is crucial for conducting Self-Translation and direct evaluations on the target language, ensuring fair performance comparisons. More details are available in Appendix A.1.

5.1 Overall Performance Results

As shown in Table 1, we demonstrated the performance of the proposed NatLan, along with top-notch related methods, on the MMMLU benchmark. Overall, NatLan 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.

More specifically, the performance of different methods exhibited a particular incremental pattern, i.e. *Self-Translation* < *Google-MT* < *NatLan*. We confirmed in Section §5.2 that such performance differences are closely associated with the incremental retention of DTs in the questions translated by each method, where the retention follows the same order: *Self-Translation* < *Google-MT* < *NatLan*. Moreover, virtually all translate-then-answer methods outperformed direct answers in the non-native language (reflecting the impact of LTs), validating the rationality of using LTs and DTs to interpret the translate-then-answer methods.

However, it is noteworthy that on a few MLLMs, particularly Llama-2 (7B), due to the limitations of the Speaker LLM's own instruction-following ca-

[4] https://huggingface.co/datasets/openai/MMMLU

[5] Translation-only NMT model, unlike the general-purpose instruction-following MLLMs mentioned in this study.

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
+NatLan	64.18	64.23	65.55	65.58	65.00
+Human (Gold)	68.10	68.10	68.10	68.10	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
+NatLan	70.52	70.30	72.16	71.92	71.46
+Human (Gold)	74.67	74.67	74.67	74.67	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
+NatLan	56.10	56.03	56.84	56.72	56.55
+Human (Gold)	58.12	58.12	58.12	58.12	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
+NatLan	56.08	56.17	57.34	56.93	56.59
+Human (Gold)	58.70	58.70	58.70	58.70	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
+NatLan	32.54	32.34	33.18	32.79	31.40
+Human (Gold)	34.97	34.97	34.97	34.97	34.97

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

pabilities, the performance of Self-Translation was worse than answering directly in the non-native language. This degradation is caused by the propagation of errors during the translation phase, further emphasizing the necessity of incorporating additional models better suited for multilingual translation tasks to achieve more stable performance improvements from the translate-then-answer process. Further performance demonstrations on the C-Eval benchmark are available in Appendix A.4.

5.2 NatLan Retains More Domain Triggers

To effectively assess the superiority of the proposed NatLan over other translate-then-answer baselines in DT retention, we developed a pairwise comparison evaluation method. Under the supervision of GPT-4o-mini, this evaluation method examines the advantage ratios of different approaches in DT retention across all samples in the MMMLU benchmark. Specifically, through few-shot prompting, GPT-4o-mini selects the translation with better DT retention from two given options, determining which method is superior for that sample.

As depicted in Figure 3, the results of the pairwise comparison clearly demonstrate the same incremental trend reported in §5.1, i.e., in terms of DT retention, *Self-Translation* (Phi-3-small (7B)) < *Google-MT* < *NatLan*. In other words, the DT retention achieved during the translation process is **positively** correlated with the accuracy of the answers based on the translated questions. This confirms the validity of using DTs to interpret the mechanisms of the translate-then-answer methods and emphasizes the effectiveness of NatLan's improvements aimed at enhancing DT retention.

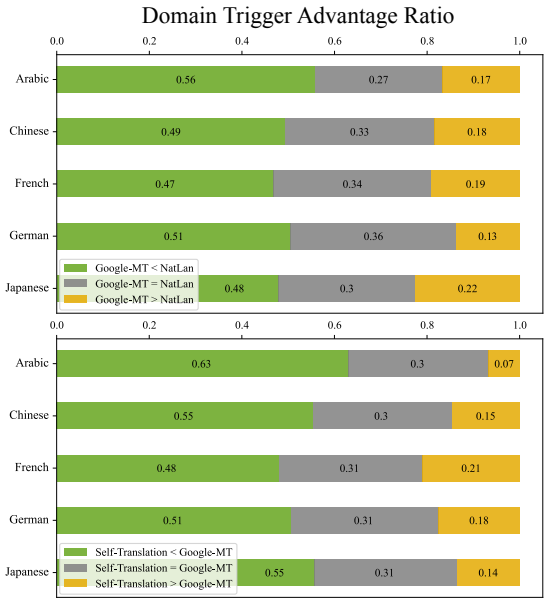

Figure 3: Pairwise compared DT advantage ratios on the MMMLU benchmark under GPT-4o-mini supervision. The gray bars indicate that the retention of DTs in the questions translated by the two methods is nearly equivalent. More details are available in Appendix A.1.

5.3 Two-Aspect Domain Trigger Retention

We conducted a comprehensive analysis to determine how NatLan retains DTs from the original questions in translated cases, summarizing 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 the first row of Table 2, Google-MT produces a literal translation such as "Daughter Country" without considering the folkloric context of the Chinese expression, whereas NatLan accurately uses "the Kingdom of Women". Similarly, NatLan uses

Original Question	Google-MT Trans. Question	NatLan 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 +NatLan : C True Label: C
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

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

“literacy” instead of “quality”, etc., rendering the descriptions of the translated entities semantically more accurate in the folkloric context of this case.

(ii) Explicit descriptions of relationships between domain-specific entities. Although it is possible to approximately infer the implicit relationships between domain-specific entities through some deliberation, this clearly increases cognitive load and does not guarantee the accuracy of such inferences. Explicitly describing domain-specific operations between entities can effectively avoid ambiguity from implicit descriptions. As shown in the first row of Table 2, NatLan explicitly renders the “is accessed” operation between the operand and the addressing method, conveying a more complete translation of the domain-specific relationships implied in the original question.

It should be noted that since Self-Translation is inferior to Google-MT, to fairly demonstrate the advantages of NatLan in DT retention, we have only compared NatLan with the Google-MT.

5.4 Knowledge Activation Rectification

We further analyzed how increased retention of DTs in translation results impacts knowledge elicitation during the answering process at a more fine-grained activation level. In our QA 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. Therefore, we extracted it for knowledge activation analysis. Additionally, on the MMMLU benchmark, we used the knowledge activation generated by Speaker LLMs when answering on the human-constructed English version MMLU benchmark as the gold knowledge activation standard[6] for evaluation.

We initially visualized the knowledge activation distribution of Speaker LLMs using t-SNE (Van der Maaten and Hinton, 2008), as depicted in Figure 4. The blue dots represent the knowledge activation distributions obtained from various methods on the MMMLU. From left to right, the methods used are: answering directly on non-native language questions (Original), Self-Translation, Google-MT, and our proposed NatLan. Qualitatively, compared to the knowledge activation distribution obtained when answering on the human gold standard (Green), NatLan shows a higher degree of overlap.

Moreover, on the MMMLU benchmark, we measured the average Euclidean distance between the activation vectors obtained through various methods and those obtained when answering using the human gold standard. As shown in Table 3, NatLan achieved the closest approximation to the human gold standard in terms of knowledge activation.

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

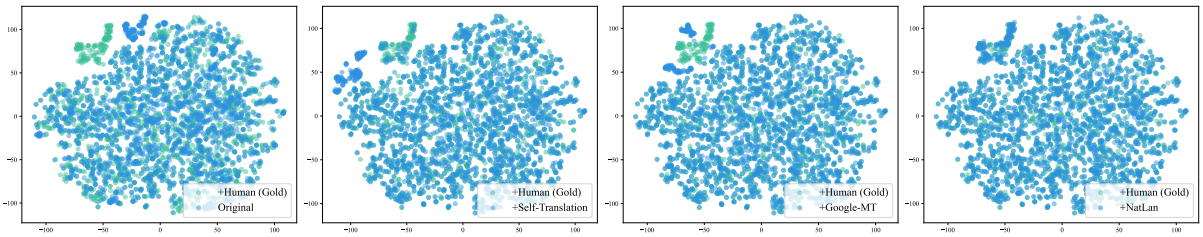

Figure 4: The visualized knowledge activation distributions on the French version of the MMMLU benchmark, with the Speaker LLM: Phi-3-small (7B). The greater the overlap with the green dots (human gold standard), the more accurate the knowledge activation is considered. More cases are available in Appendix A.3.

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
+NatLan	9.84	11.24	7.46	8.38	11.28

Table 3: Average Euclidean distances between the knowledge activation (extracted from Phi-3-small (7B)) obtained using various methods and that obtained using the human gold standard. A smaller distance indicates a closer approximation to the human gold standard.

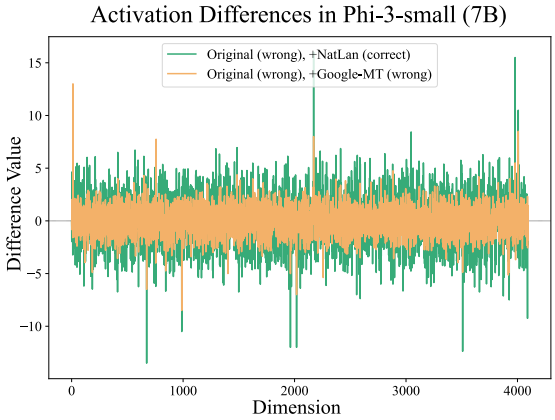

Figure 5: Activation differences between different methods for the same questions. Contents in parentheses indicate the correctness of the Speaker LLMs' responses.

Interestingly, the performance evaluation of activation distance followed the same incremental trend previously mentioned, namely: *Self-Translation* < *Google-MT* < *NatLan*. This uniformly confirms that more complete DT retention after translation can rectify the activation in Speaker LLMs when answering questions, eliciting relevant knowledge.

Furthermore, from Figure 5, we can observe that Google-MT, the best-performing related methods, also provides some degree of knowledge activation rectification. However, due to its inferior DT retention compared to our proposed NatLan, the rectification effect brought by Google-MT is insufficient to correct the answers. In contrast, NatLan, through more complete DT retention, delivers a higher level of rectification effect, thereby successfully correcting originally incorrect answers.

5.5 Impact of Translator LLMs' Semantic Capabilities on NatLan

To analyze the correlation between the semantic capabilities of the Translator LLM and the extent of improvements it can provide within NatLan, we employed three Qwen MLLMs with varying Chinese semantic understanding abilities as Chinese-to-English Translator LLMs and conducted a comparative analysis of performance gains on the C-Eval Chinese benchmark. As shown in Table 4, 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 Translator 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 Translator LLMs in the effectiveness of NatLan. Intuitively, this impact applies to all translate-then-answer methods.

5.6 Impact of Domain-Language Characteristics on NatLan

More comprehensively, we explored the characteristics that exist between specific domains and languages and their impact on NatLan. As shown in Figure 6, although using NatLan to translate questions from a non-native language (Chinese) to the

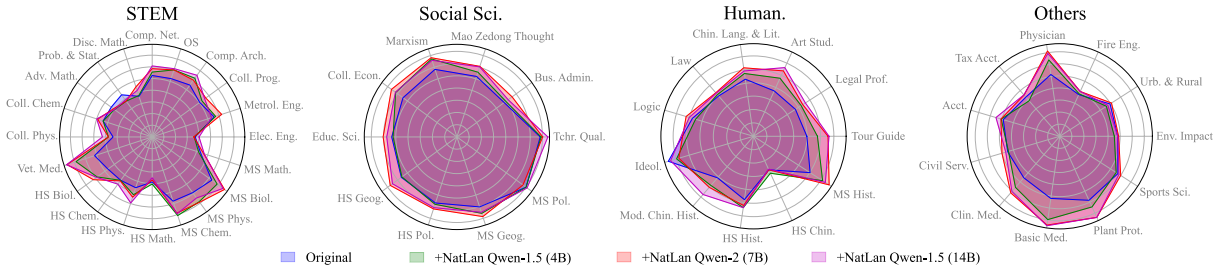

Figure 6: Performance of NatLan across 52 disciplines on the C-Eval test sets under four configurations, with Phi-3-small (7B) as the Speaker LLM. More details are available in Appendix A.5.

Model	Lang.	Avg.	Avg. (Hard)
Translator 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 4: Performance comparison of NatLan using different Translator LLMs on the C-Eval test sets.

native language (English) greatly enhances answer accuracy across most disciplines, due to the richer knowledge available in the native language across most domains, there are exceptions. In a few disciplines, translating languages using NatLan does not result in performance gains. We categorized such disciplines with specific domain-language characteristics 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 during translation, impacting the correct understanding of the questions.

(ii) Language-Knowledge Bound Disciplines: Disciplines like Ideological and Moral Cultivation (Ideol.), which are closely tied to specific languages due to cultural and other differences between countries. In these disciplines, training materials in different languages may contain distinctly different relevant knowledge about similar topics. Uniformly translating such questions into the native language (English) could lead to confusion and conflicts in the relevant knowledge, impacting the correct elicitation of knowledge.

It should be noted that while there is a possibility that applying NatLan 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.

6 Conclusion

MLLMs often fail to answer questions posed in non-dominant languages, which they can successfully address when presented in their dominant languages. Although existing translate-then-answer methods can mitigate this issue, the mechanisms behind their effectiveness remain unclear. To clarify, we employ human cognitive features: Language Triggers (LTs) and Domain Triggers (DTs) to interpret the mechanisms behind existing translate-then-answer methods, highlighting issues of DT insufficiency in existing approaches and proposing NatLan as a remediation. Our proposed NatLan achieves up to a **31.28%** average accuracy improvement across five non-native language QA benchmarks and provides comparable or greater retention of DTs in up to **87%** of cases compared to Google-MT, advancing the understanding of translate-then-answer process through the lens of LTs and DTs.

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, such as the enhanced proficiency in Chinese of Qwen MLLMs, they still cannot outperform the performance under English prompting. In other words, English corpora consistently dominate their training data. In our preliminary investigation, Qwen-2 (7B) achieved only **47.70%** accuracy on the human expert-constructed Chinese version of MMLU, significantly lower than the **58.19%** accuracy attained on the English version of MMLU. Therefore, we encourage future research to develop MLLMs with different native languages other than English, or investigate other advanced language transfer techniques. 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 Translator 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 Translator 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. The C-Eval Chinese benchmark have CC-BY-NC-SA-4.0 License [7], the MMMLU benchmark and Phi-3-* models have MIT License [8], Qwen-1.5-* models have Tongyi-Qianwen-Research License [9], Qwen-2-* and Mistral-0.3-* models have Apache-2.0 License [10], Llama-2-* models have Llama 2 Community License [11] and Gemma-1.1-* models have Gemma Terms of Use [12].

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*.

Kingsley N Acheampong, Zhen-Hao Pan, Er-Qiang Zhou, and Xiao-Yu Li. 2016. Answer triggering of factoid questions: A cognitive approach. In *2016 13th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 33–37. IEEE.

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*.

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

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

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

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

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

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

Lewis J Baker and Daniel T Levin. 2015. The role of relational triggers in event perception. *Cognition*, 136:14–29.

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.

Nicola Del Maschio, Federico Crespi, Francesca Peressotti, Jubin Abutalebi, and Simone Sulpizio. 2022. Decision-making depends on language: A meta-analysis of the foreign language effect. *Bilingualism: Language and Cognition*, 25(4):617–630.

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

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.

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.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023a. 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*, pages 12365–12394.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Jiayi Lei, Yao Fu, et al. 2023b. 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.

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.

Jian Liu, Yufeng Chen, and Jinan Xu. 2022. Saliency as evidence: Event detection with trigger saliency attribution. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 4573–4585.

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*.

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*.

Meihan Tong, Bin Xu, Shuai Wang, Yixin Cao, Lei Hou, Juanzi Li, and Jun Xie. 2020. Improving event detection via open-domain trigger knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5887–5897.

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.

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

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

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 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*, pages 7915–7927.

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 Translator and Speaker LLMs. Furthermore, all MLLMs involved in the experiments are open-source models of the Instruct/Chat version: Phi-3-mini (3.8B) [13], Phi-3-small (7B) [14], Gemma-1.1 (7B) [15], Mistral-0.3 (7B) [16], Llama-2 (7B) [17], Qwen-1.5 (4B) [18], Qwen-2 (7B) [19], and Qwen-1.5 (14B) [20].

At the same time, as we deployed Translator LLMs within NatLan that required designing translation prompts, we used GPT-4o [21] 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. As for the MMMLU benchmark, we directly used official samples translated by human experts as examples in our prompts and employed GPT-4o-mini[22] as a cost-effective and powerful universal translator.

Since the Translator 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 {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.

<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>
```

[13] https://huggingface.co/microsoft/Phi-3-mini-128k-instruct
[14] https://huggingface.co/microsoft/Phi-3-small-128k-instruct
[15] https://huggingface.co/google/gemma-1.1-7b-it
[16] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
[17] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
[18] https://huggingface.co/Qwen/Qwen1.5-4B-Chat
[19] https://huggingface.co/Qwen/Qwen2-7B-Instruct
[20] https://huggingface.co/Qwen/Qwen1.5-14B-Chat
[21] API version: gpt-4o-2024-05-13
[22] API version: gpt-4o-mini-2024-07-18

```
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:

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

Additionally, the disciplines used for evaluating the advantages of DT retention, discussed in §5.2, encompass ten domains rich in domain-specific terms, including Anatomy, Astronomy, College Biology, College Chemistry, College Computer Science, College Medicine, High School Biology, High School Chemistry, High School Computer Science, and Virology.

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 5.

As shown in Table 5, in the examples from the first two rows, NatLan provides more explicit operation descriptions in translations, enabling Speaker LLMs to more easily understand the relationship between the domain-specific entities in the questions and answer candidates. In the cases presented in the latter two rows, NatLan delivers translations with more accurate domain-specific entity descriptions. For these two questions pertaining to the High School Chemistry discipline, the enriched semantic comprehension of the Translator LLMs enables NatLan to generate terms 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".

A.3 More Cases in Knowledge Activation

As a supplement to §5.4, we report the visualized knowledge activation distributions for the Arabic, Chinese, German, Japanese version of the MMMLU benchmark, with the Speaker LLM: Phi-3-small (7B), in Figure 7, Figure 8, Figure 9 and Figure 10, respectively. Moreover, we report the case 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 6.

A.4 Overall Performance Results on the C-Eval benchmark

As shown in Table 7, our proposed NatLan also achieved optimal performance on the C-Eval benchmark. However, unlike MMMLU, the C-Eval benchmark does not have a human expert-constructed gold standard English version and only includes Chinese. Therefore, we primarily used it for ablation studies and case studies.

A.5 Analysis of NatLan with Different Translators in Various Domains

As a supplement to Figure 6, we present a detailed performance analysis of NatLan, employing three different Translator LLMs applied to various Speaker LLMs, across specific disciplines. These include Phi-3-mini (3.8B) in Figure 12, Gemma-1.1 (7B) in Figure 13, Mistral-0.3 (7B) in Figure 14, and Llama-2 (7B) in Figure 15.

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 Translator 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 15, 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 Translator LLMs on Llama-2 (7B) show rel-

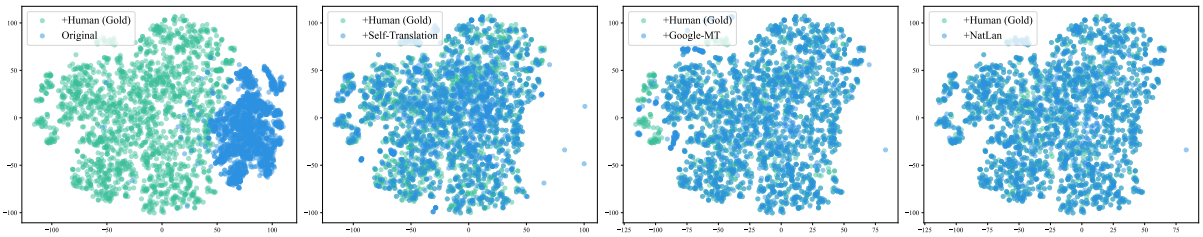

Figure 7: The visualized knowledge activation distributions on the Arabic version of the MMMLU benchmark, with the Speaker LLM: Phi-3-small (7B). The greater the overlap with green (human gold standard), the more accurate the knowledge activation is considered.

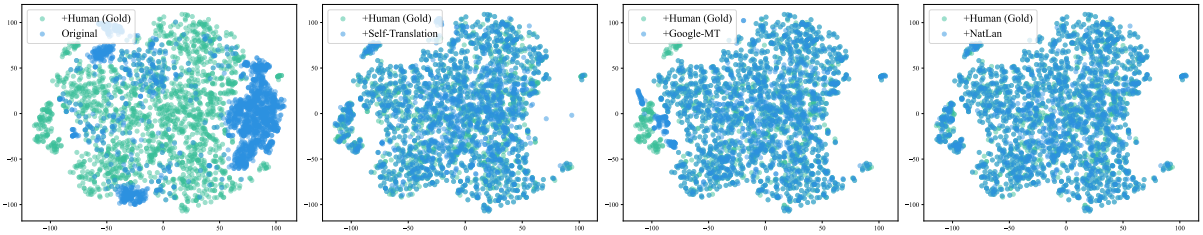

Figure 8: The visualized knowledge activation distributions on the Chinese version of the MMMLU benchmark, with the Speaker LLM: Phi-3-small (7B). The greater the overlap with green (human gold standard), the more accurate the knowledge activation is considered.

atively greater variability. However, from a holistic perspective, disregarding the variations between different Translator 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 8 for all settings, as a supplement to Table 7 and Table 4

A.6 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 11, 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.

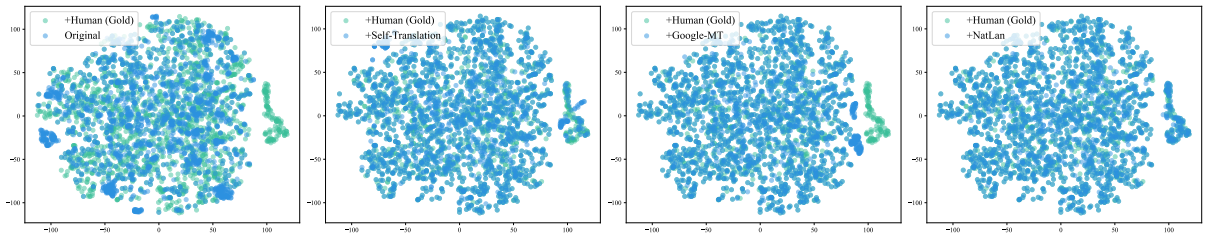

Figure 9: The visualized knowledge activation distributions on the German version of the MMMLU benchmark, with the Speaker LLM: Phi-3-small (7B). The greater the overlap with green (human gold standard), the more accurate the knowledge activation is considered.

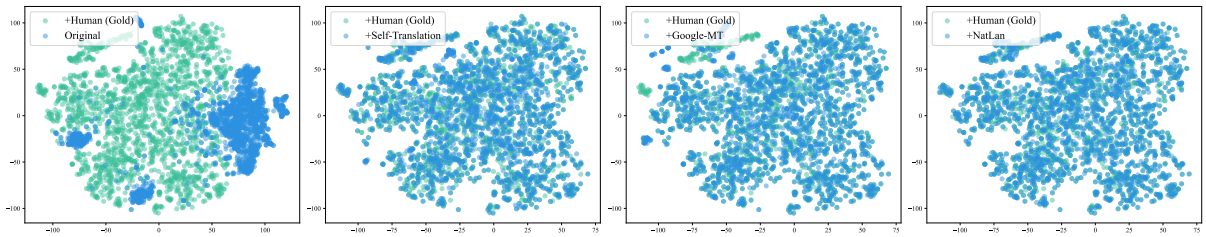

Figure 10: The visualized knowledge activation distributions on the Japanese version of the MMMLU benchmark, with the Speaker LLM: Phi-3-small (7B). The greater the overlap with green (human gold standard), the more accurate the knowledge activation is considered.

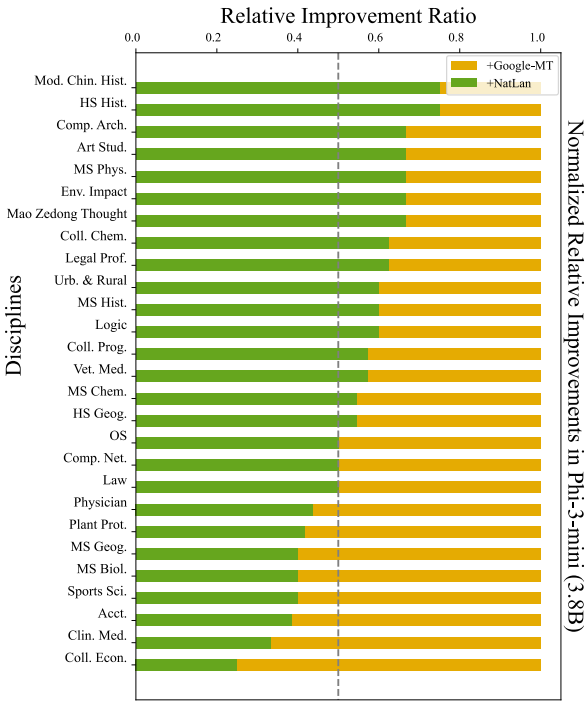

Figure 11: Normalized relative improvements in specific disciplines, with the dashed grey line indicating where their respective relative improvements are equivalent.

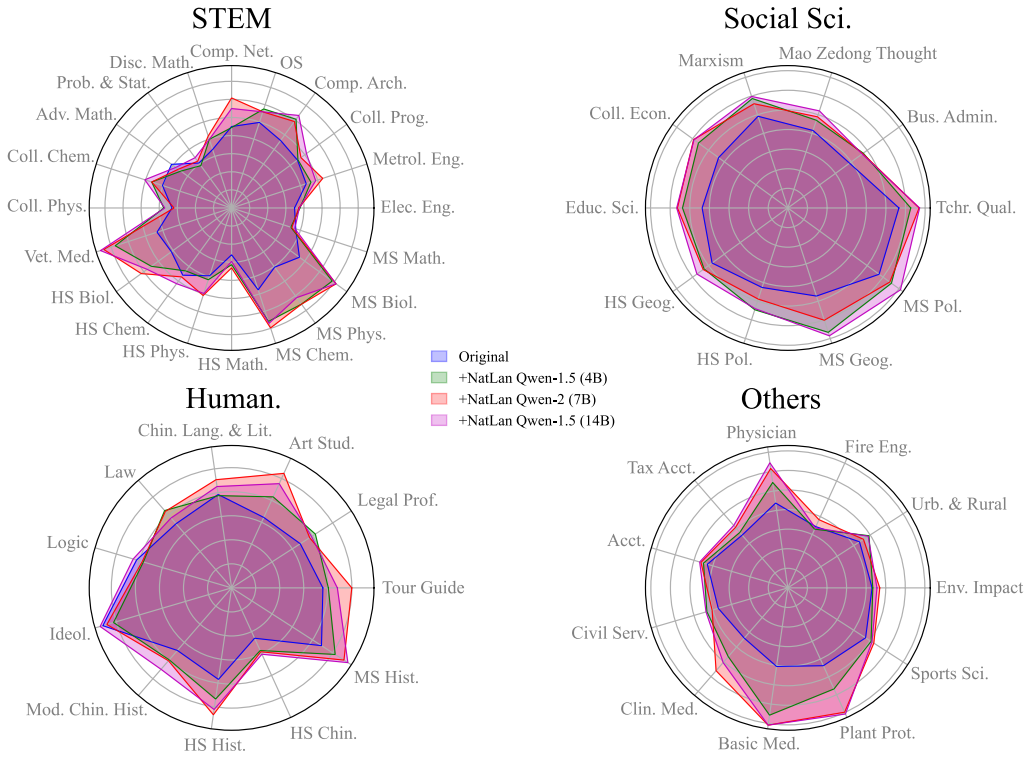

Figure 12: Performance comparison of NatLan using different Translator LLMs in the C-Eval test sets, divided into 52 distinct disciplines, with Phi-3-mini (3.8B) as the Speaker LLMs.

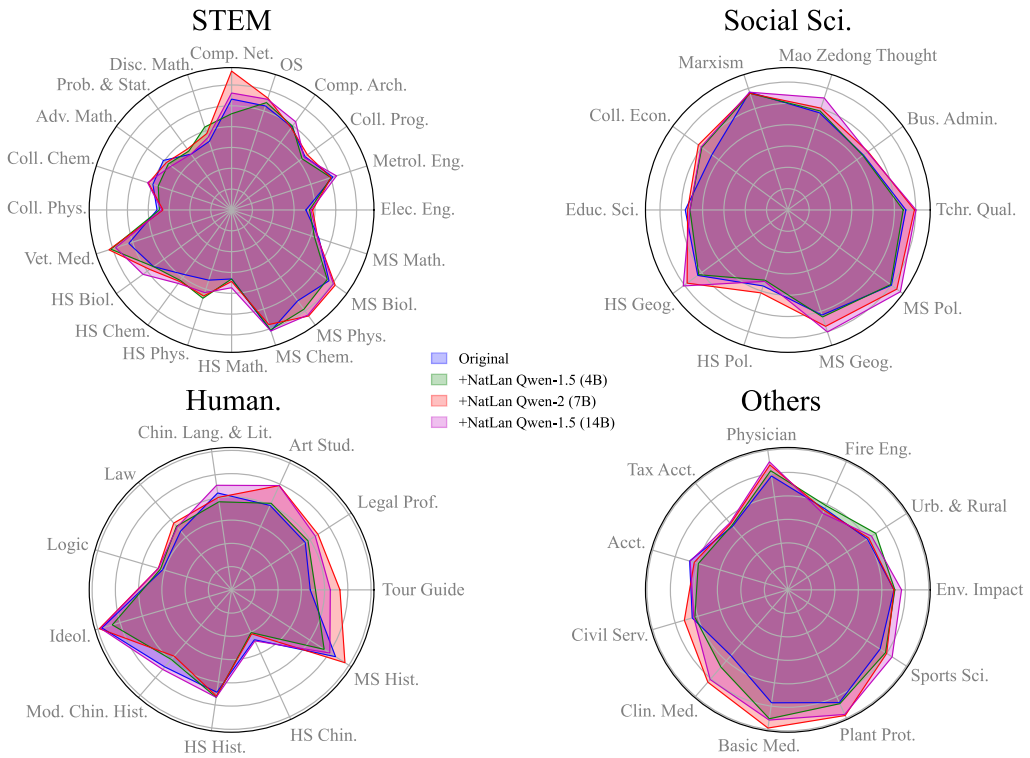

Figure 13: Performance comparison of NatLan using different Translator LLMs in the C-Eval test sets, divided into 52 distinct disciplines, with Gemma-1.1 (7B) as the Speaker LLMs.

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 5: Supplementary Chinese-to-English translation cases, with cases sampled from the C-Eval test sets. The Speaker LLM is Phi-3-mini (3.8B) and the Translator LLM is Qwen-1.5 (14B) for this case study.

Original Question	Google-MT Trans. Question	NatLan Trans. Question	Answers
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 6: Cases sampled from the C-Eval val/test sets for knowledge activation analysis in §5.4.

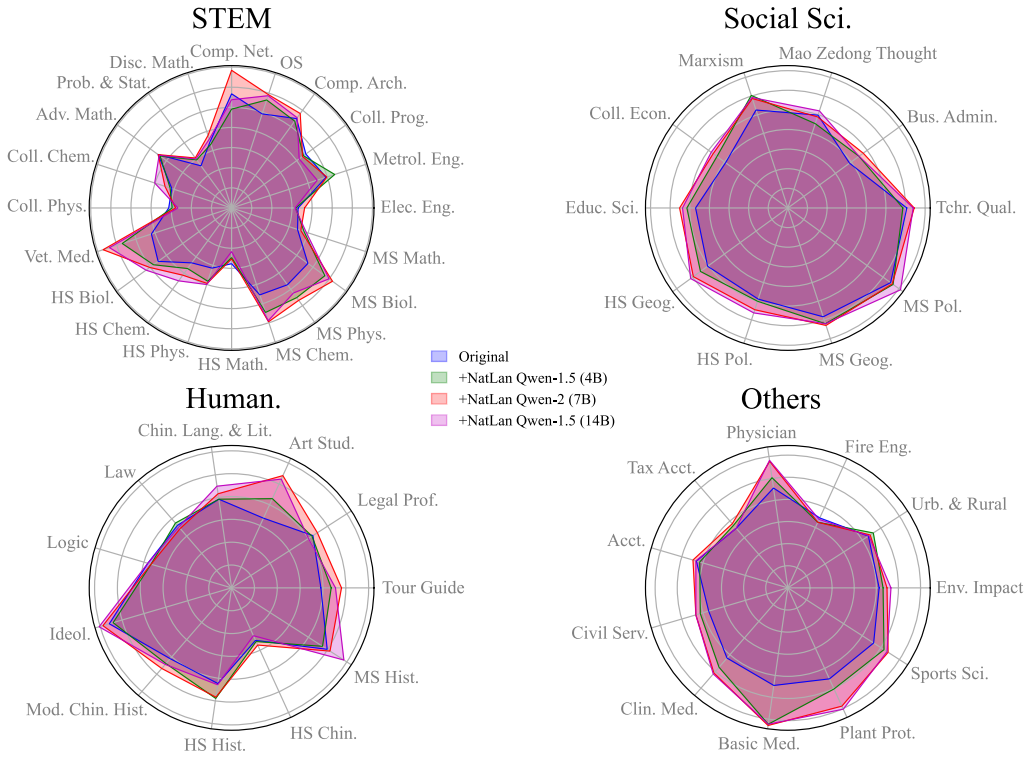

Figure 14: Performance comparison of NatLan using different Translator LLMs in the C-Eval test sets, divided into 52 distinct disciplines, with Mistral-0.3 (7B) as the Speaker LLMs.

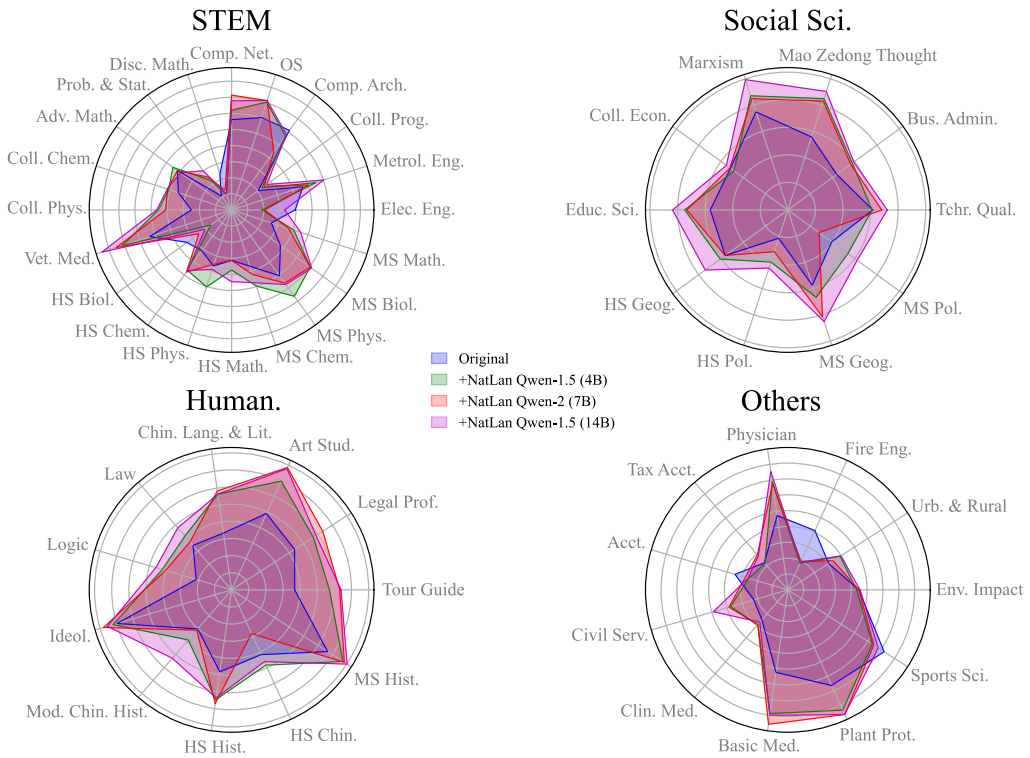

Figure 15: Performance comparison of NatLan using different Translator LLMs in the C-Eval test sets, divided into 52 distinct disciplines, with Llama-2 (7B) as the Speaker LLMs.

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 7: Performance on the C-Eval Chinese benchmark, measured by accuracy. *Lang.* indicates the language of the questions. The meanings of these colors are the same as in Table 1. NatLan employs Qwen MLLMs as the Chinese-to-English translators, selecting the Qwen configurations reported as optimal in §5.5.

Model	Lang.	STEM	Social Sci.	Human.	Others	Avg.	Avg. (Hard)
Translator 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 8: 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 7.