Do LLMs Need to Think in One Language? Correlation between Latent Language and Task Performance

Shintaro Ozaki[α,β] **Tatsuya Hiraoka**[γ] **Hiroto Otake**[α,β]
Hirki Ouchi[α,η] **Masaru Isonuma**[β,δ,ϵ] **Benjamin Heinzerling**[η,δ] **Kentaro Inui**[γ,δ,η]
Taro Watanabe[α] **Yusuke Miyao**[ϵ,β] **Yohei Oseki**[ϵ] **Yu Takagi**[θ]

[α]NAIST [β]NII LLMC [γ]MBZUAI [η]RIKEN
[δ]Tohoku University [ϵ]The University of Tokyo [θ]Nagoya Institute of Technology

ozaki.shintaro.ou6@naist.ac.jp takagi.yu@nitech.ac.jp

Abstract

Large Language Models (LLMs) are known to process information using a proficient internal language consistently, referred to as latent language, which may differ from the input or output languages. However, how the discrepancy between the latent language and the input and output language affects downstream task performance remains largely unexplored. While many studies research the latent language of LLMs, few address its importance in influencing task performance. In our study, we hypothesize that thinking in latent language consistently enhances downstream task performance. To validate this, our work varies the input prompt languages across multiple downstream tasks and analyzes the correlation between consistency in latent language and task performance. We create datasets consisting of questions from diverse domains such as translation and geo-culture, which are influenced by the choice of latent language. Experimental results across multiple LLMs on translation and geo-culture tasks, which are sensitive to the choice of language, indicate that maintaining consistency in latent language is not always necessary for optimal downstream task performance. This is because these models adapt their internal representations near the final layers to match the target language, reducing the impact of consistency on overall performance.

1 Introduction

Large Language Models (LLMs) are known to process information using a proficient internal language consistently, referred to as latent language [1, 2], which may differ from the input or output languages. For instance, Llama2 [3], which is primarily trained on English-centric corpora, employs English as a latent language during inference, shaping its internal representations accordingly [1]. Similarly, models like LLM-jp [4], which are trained on both Japanese and English data, tend to think in Japanese during their internal reasoning processes [2]. This pattern suggests that LLMs tend to adopt their proficient latent language shaped by their training data distribution.

However, the correlation between consistency in latent language and downstream task performance remains unexplored. In particular, it is unclear how inputs in a less familiar language affect internal reasoning and, consequently, the accuracy of outputs as shown in Figure 1. This perspective is critical for understanding the linguistic priors and stability of reasoning in multilingual LLMs [5].

Preprint. Under review.

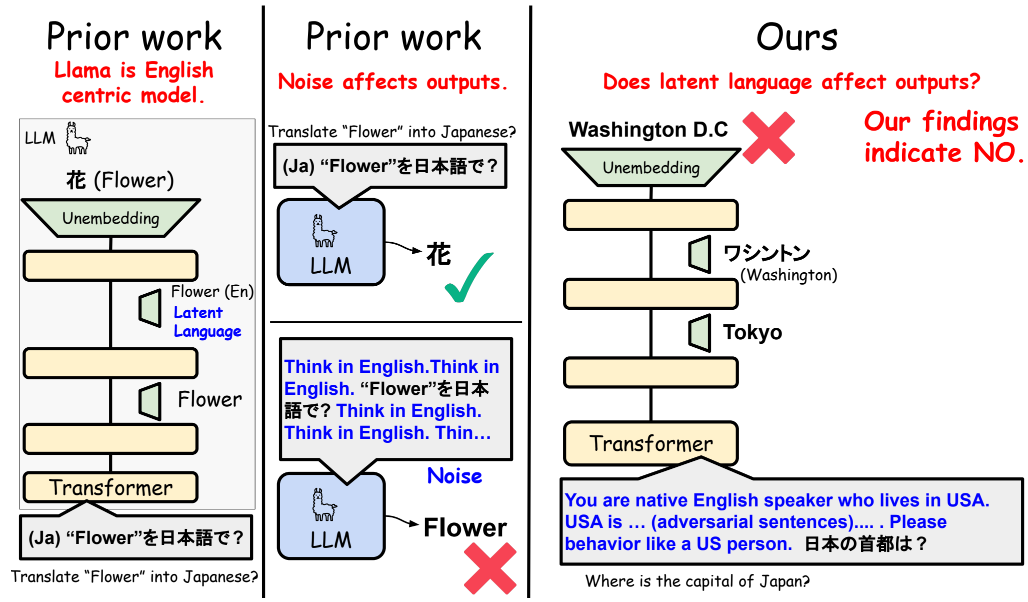


Figure 1: An overview of our research and differences from prior works.

In our study, we hypothesize that disrupting thought in latent language has an influence on downstream task performance. To evaluate this, we deliberately disrupt the internal linguistic consistency of models by injecting adversarial prompts composed of multiple languages into the inputs and quantitatively evaluate how these changes affect downstream task performance. To understand this correlation, we evaluate several LLMs known to possess their proficient latent language, systematically varying both the types and proportions of adversarial prompts' languages. Additionally, we introduce a formal evaluation metric which we define as Latent Language Consistency Score (LLC Score) for quantifying the consistency in latent language within models.

Regarding tasks, there are no existing datasets that quantitatively evaluate our research focus. Thus, our study constructs diverse question datasets inspired by prior research [1, 2, 6, 7], specifically targeting domains where consistency in latent language is greatly affected by their accuracy. In our work, we specifically focus on translation and geo-culture tasks, allowing for precise evaluation.

Experimental results on our created datasets, conducted across multiple LLMs trained in different languages, reveal that models often rely on a consistent latent language during inference. However, even when they do so, they do not always achieve optimal performance, contradicting our initial hypothesis in both tasks. In the translation task, the inclusion of adversarial prompts in less familiar languages from their proficient latent language degrades accuracy. This suggests that linguistic noise in the inputs disrupts the internal reasoning process and destabilizes the model's internal structures.

2 Related Work

2.1 Latent Language of LLMs

LLMs are known to exhibit behavior strongly influenced by the languages prevalent in their pre-training data. For example, Wendler et al [1] demonstrate that Llama2 [3], which is primarily trained on English data, tends to rely on English for internal reasoning as a latent language, while Zhong et al [2] report that LLM-jp models [4] trained on both Japanese and English exhibit internal representations toward Japanese language. These findings suggest that LLMs often develop a proficient latent language [8] for internal processing, reflecting inherent linguistic representations [9, 10] shaped by their data.

Additionally, the choice of prompt language has been shown to strongly affect model outputs and performance in multilingual settings. For instance, Huang et al [11] found that prompts in different languages lead to substantial variations in generated outputs, i.e., accuracy. Other studies [12, 13] have analyzed the impact of multilingual Chain-of-Thought [14] prompting on reasoning processes and performance.

However, these investigations primarily focus on surface level changes in outputs, such as response accuracy [15], without directly validating the latent language or how shifts in internal reasoning languages might affect downstream task performance.

2.2 Robustness of LLMs to Adversarial Prompts

Prior work has explored the robustness of models to external linguistic noise, i.e., adversarial prompts. For example, Goel et al [16] examined the impact of various syntactic and lexical affections, while Khashabi et al [17] demonstrated that inserting semantically unrelated sentences can significantly reduce the output reliability of question answering models as shown in the center of Figure 1. However, these studies primarily focus on how noise disrupts outputs, while leaving unexplored the potential shifts in the latent languages that models rely on during reasoning.

In contrast to those contributions, our study aims to systematically analyze how the latent language of LLMs influences output robustness and downstream task performance, addressing a critical gap in our understanding of language dependent reasoning. By exploring the correlation between internal linguistic shifts and downstream task performance, we provide a comprehensive framework for evaluating the stability of internal reasoning in multilingual LLMs.

2.3 Visualization of Internal Thought in LLMs

LogitLens [18, 19] enables the calculation of logits, probabilities, and corresponding tokens at each intermediate layer of Transformer [20] decoder models by projecting the hidden states through the unembedding matrix. This approach makes it possible to analyze what the model is "thinking" and which language it is using at each layer of the Transformer [1]. In our study, we utilize this approach to visualize the internal reasoning of the model, describing more details in Appendix B.2. By extracting logits through the application of the unembedding matrix to the outputs of each layer, we can visualize tokens, compute probabilities, and further calculate the KL divergence between adjacent layers to define metrics for language consistency and robustness.

3 Analysis Method

In our study, we analyze how consistently thinking in a latent language affects downstream task performance, which we refer to as robustness. To analyze this correlation, we first identify the latent language used in the model's intermediate layers during inference. We then define specific metrics for both consistency in latent language and robustness, and evaluate whether reasoning in the latent language leads to optimal performance. As we described in Section 1, our hypothesis is disrupting the latent language should lead to a corresponding drop in robustness.

3.1 Latent Language Detection

As described in Section 2.1, we call the language used internally by the model for reasoning as the latent language. Inspired by prior work [1, 2, 8], we use LogitLens [18], introduced in Section 2.3, to extract tokens from the logits obtained at intermediate layers. We then apply a language identification library[1] to classify the extracted tokens as Japanese, Chinese, or English.

3.2 Latent Language Consistency Score (LLC Score)

We define the LLC Score to quantify how consistently a model relies on a latent language throughout its intermediate layers. Our hypothesis is that if the model frequently shifts its internal reasoning across different latent languages, such behavior indicates a lack of coherence in internal processing. In contrast, when the same latent language is consistently used across layers, information flows more smoothly through the model, leading to higher consistency.

[1] `https://github.com/saffsd/langid.py`.

To capture this behavior, we define a score, Score$(v;\theta)$, for each candidate language $v \in \mathcal{V}$, which evaluates how stably the model internally uses that language. This score depends on two key components: The change in internal representations between adjacent layers (KL divergence); The output probability that the model uses language v at each layer. (Applied by LogitLens [18].)

When the model prefers a latent language v at layer l but switches to the different dominant language at layer $l+1$, i.e., $v_{l+1}^{*(\theta)} \neq v$, such a transition suggests a disruption in consistency for v. The function Score$(v;\theta)$ averages the degree of such disruptions:

$$\text{Score}(v;\theta) = \frac{\sum_{l=1}^{L-1} \left(P_{l,v}^{(\theta)} \cdot KL_{l,l+1}^{(\theta)} + KL_{l,l+1}^{(\theta)} \cdot P_{l+1,v}^{(\theta)} \right) \cdot \mathbb{1}(v_{l+1}^{*(\theta)} \neq v)}{\sum_{l=1}^{L-1} \left(P_{l,v}^{(\theta)} + P_{l+1,v}^{(\theta)} \right) \cdot \mathbb{1}(v_{l+1}^{*(\theta)} \neq v)} \tag{1}$$

We define $P_{l,v}^{(\theta)}$ as the probability that the model uses latent language v at layer l. The term $KL_{l,l+1}^{(\theta)}$ denotes the KL divergence between the output distributions at layers l and $l+1$. We denote the most likely language at layer l as $v_l^{*(\theta)} = \arg\max_{u \in \mathcal{V}} P_{l,u}^{(\theta)}$, where u represents a candidate language. The indicator function $\mathbb{1}(v_{l+1}^{*(\theta)} \neq v)$ returns 1 if the most likely language at layer $l+1$ differs from v.

With LogitLens [18] as we described in Section 2.3, $KL_{l,l+1}^{(\theta)}$ enables extraction of logits from each intermediate layer, and $v_l^{*(\theta)} = \arg\max_{u \in \mathcal{V}} P_{l,u}^{(\theta)}$ identifies the language of the token with the highest output probability determined by Section 3.1.

The score aggregates internal disruptions related to language v, weighted by the model's output probability in using that language. Then, the overall LLC Score for model θ is defined as the minimum of the consistency scores across all candidate languages:

$$\text{LLC Score}(\theta) = \min_{v \in \mathcal{V}} \text{Score}(v;\theta) \tag{2}$$

A lower LLC Score suggests that the model maintains stable internal representations using at least one latent language, just as lower KL divergence indicates better layerwise consistency.

Following prior work [18, 21], we restrict our analysis to the latter half of the model's intermediate layers, where semantic reasoning tends to consolidate.

3.3 Robustness against Adversarial Prompts

We define accuracy, i.e., the model's outputs, as a measure of robustness. Our hypothesis, as described in Section 3, is that when the latent language of the model is disrupted, the robustness in downstream tasks should also be similarly affected.

3.4 Adversarial Prompts

To guide the latent language processing of models, we introduce adversarial prompts. Drawing on prior work [16, 17] that highlights the accuracy degradation caused by adversarial inputs, we aim to disrupt consistency in latent language as well. Specifically, we create prompts that describe the culture, background, and history of Japan, English, Chinese-speaking regions, and are followed by an instruction like, "Based on the above, act as if you are in Chinese." as shown in Appendix B.7. This approach tests whether inducing models to think in the prompt language can disrupt consistency in latent language and affect robustness.

4 Dataset Creation

We choose cloze style tasks [1, 2, 22], where the target is to predict a specific word or phrase that fills a blank space ('_') in a sentence [23]. This format is particularly suitable for LogitLens

Table 1: All of settings we used in our experiments.

Option	Candidates
Task	Translation, Geo-culture
Language	English, Japanese, Chinese
Data size	2,000 for each
Model	Qwen2.5, Gemma3, LLM-jp-3
Translation Task Language	{En-Ja}, {En-Zh}, {Zh-En}, {Ja-En}
Geo-culture Task Language	{En}, {Ja}, {Zh}
Adversarial Prompts Language	{En}, {Ja}, {Zh}
Adversarial Prompts Ratio	{None, 0.2, 0.4, 0.6, 0.8, 1.0}

analysis [18], as it focuses on specific token predictions, allowing us to more precisely interpret intermediate representations. In prior research, no cloze task datasets specifically designed to evaluate the impact of linguistic consistency on model robustness have been developed. To address this issue, we construct a dataset, following prior studies [6, 7] that demonstrate the feasibility of semi-automatically generating QA pairs using LLMs. Our work focuses on translation and geo-culture domains, which are particularly sensitive to the choice of language. We split into two distinct steps, following the method described in previous work [6, 7], which emphasizes the importance of separating question generation from model inference. This approach ensures that the generated questions are not biased by the models being evaluated, providing a cleaner test.

4.1 STEP 1: Generate Questions

Our work uses prompts like "Generate cloze task questions and their answers for translation." to generate cloze task questions [23]. To ensure diversity and avoid overfitting to a single prompt structure [24], we prepare a list of several dozen category labels as shown in Appendix B.1. When generating the questions, these categories are selected at random, encouraging the model to produce a wide range of questions and reducing the risk of repetitive patterns. We also calculate Self-BLEU [25] as a diversity metric to quantitatively evaluate; the results of which are presented in Table 3 in the Appendix B.1. These results confirm that our dataset exhibits diverse questions, ensuring its reliability. For question generation, we use GPT-4o [26], the highest performing model. Appendix B.7 provides the specific prompts used for this process.

4.2 STEP 2: Filter Questions

After generating the initial set of cloze task questions in STEP 1, we perform an additional filtering step to ensure the quality and consistency of the dataset. Specifically, we instruct GPT-4o to answer the generated questions and then compare the predicted answers with the original answers created during question generation. Only the questions for which the model's predicted answer exactly matches the original gold answer are retained in the final dataset.

Additionally, we remove any questions that do not meet the following criteria. (1) Single-Token Answers: The answer must be a single token, following the approach used in prior studies [22]; (2) Cloze Task Format: The question must adhere to the cloze task format, as defined in previous work [6], meaning the prompt must clearly indicate a single missing word or phrase.

After this filtering process, we obtain a final dataset of 2,000 questions for each task. This strict filtering ensures that the resulting dataset effectively captures whether consistency in latent language is disrupted. Examples from the final dataset include translation tasks like "Please translate flower into Japanese._, answer: " and geo-culture tasks like "Who is the prime minister of Japan?_, answer: " Appendix B.8 provides further details on the dataset construction and filtering criteria.

5 Experiments

The goal of our experiments is to examine how the consistency of latent language affects downstream task performance across multilingual LLMs. We specifically aim to test the hypothesis that maintaining internal consistency in latent language contributes to higher task accuracy, especially when

Table 2: Correlation (r) between LLC Score and robustness in each ratio by geo-culture tasks.

Model	Question	Adversarial	LLC Score (↓)					Robustness (↑)					r
			0.2	0.4	0.6	0.8	1.0	0.2	0.4	0.6	0.8	1.0	
LLM-jp-3	Ja	Ja	Ja$_{0.06}$	En$_{0.08}$	Ja$_{0.09}$	Ja$_{0.10}$	Ja$_{0.11}$	0.27	0.26	0.27	0.24	0.24	-0.82
		En	Ja$_{0.09}$	Ja$_{0.07}$	Ja$_{0.11}$	Ja$_{0.10}$	Ja$_{0.11}$	0.13	0.22	0.13	0.06	0.04	-0.83
	En	Ja	Ja$_{0.06}$	Ja$_{0.07}$	En$_{0.12}$	En$_{0.12}$	En$_{0.13}$	0.15	0.11	0.10	0.15	0.13	-0.17
		En	En$_{0.10}$	En$_{0.11}$	En$_{0.11}$	En$_{0.12}$	En$_{0.13}$	0.22	0.20	0.20	0.18	0.17	-0.97
		Zh	Ja$_{0.04}$	En$_{0.12}$	En$_{0.13}$	En$_{0.13}$	En$_{0.13}$	0.10	0.10	0.09	0.10	0.12	-0.05
	Zh	En	En$_{0.11}$	En$_{0.12}$	En$_{0.12}$	Zh$_{0.09}$	En$_{0.13}$	0.04	0.05	0.04	0.03	0.02	0.21
		Zh	Ja$_{0.05}$	Ja$_{0.06}$	Ja$_{0.10}$	En$_{0.11}$	En$_{0.10}$	0.07	0.07	0.06	0.05	0.02	-0.61
Qwen2.5	Ja	Ja	En$_{0.10}$	En$_{0.09}$	En$_{0.10}$	En$_{0.09}$	En$_{0.10}$	0.00	0.00	0.00	0.00	0.00	N.A.
		En	En$_{0.09}$	En$_{0.09}$	En$_{0.09}$	En$_{0.10}$	En$_{0.10}$	0.00	0.00	0.00	0.00	0.00	N.A.
	En	Ja	En$_{0.09}$	En$_{0.09}$	En$_{0.09}$	En$_{0.09}$	En$_{0.09}$	0.31	0.30	0.27	0.25	0.27	-0.94
		En	En$_{0.09}$	En$_{0.09}$	En$_{0.09}$	En$_{0.10}$	En$_{0.10}$	0.33	0.27	0.25	0.29	0.32	-0.25
		Zh	En$_{0.09}$	En$_{0.09}$	En$_{0.09}$	En$_{0.10}$	En$_{0.09}$	0.32	0.32	0.31	0.25	0.28	-0.87
	Zh	En	En$_{0.09}$	En$_{0.08}$	En$_{0.08}$	En$_{0.08}$	En$_{0.09}$	0.00	0.00	0.00	0.00	0.00	0.21
		Zh	En$_{0.10}$	En$_{0.10}$	Zh$_{0.09}$	Zh$_{0.09}$	En$_{0.10}$	0.01	0.01	0.00	0.00	0.01	0.83
Gemma3	Ja	Ja	Zh$_{0.03}$	Zh$_{0.01}$	Zh$_{0.03}$	En$_{0.03}$	Zh$_{0.02}$	0.00	0.01	0.01	0.01	0.02	-0.29
		En	En$_{0.02}$	En$_{0.02}$	Zh$_{0.02}$	Zh$_{0.02}$	En$_{0.02}$	0.00	0.01	0.01	0.00	0.00	0.90
	En	Ja	En$_{0.02}$	En$_{0.03}$	En$_{0.03}$	En$_{0.02}$	En$_{0.02}$	0.31	0.29	0.25	0.19	0.19	0.85
		En	En$_{0.02}$	En$_{0.02}$	En$_{0.02}$	En$_{0.02}$	En$_{0.02}$	0.29	0.27	0.17	0.14	0.12	0.98
		Zh	En$_{0.03}$	En$_{0.03}$	En$_{0.02}$	En$_{0.02}$	En$_{0.02}$	0.31	0.31	0.25	0.23	0.23	0.98
	Zh	En	En$_{0.02}$	En$_{0.02}$	En$_{0.02}$	En$_{0.02}$	En$_{0.02}$	0.00	0.01	0.00	0.00	0.00	0.83
		Zh	Zh$_{0.02}$	Zh$_{0.02}$	Zh$_{0.02}$	Zh$_{0.02}$	Zh$_{0.02}$	0.01	0.01	0.00	0.01	0.00	-0.15

models encounter prompts containing unfamiliar languages. Ideally, models that consistently think in their proficient latent language should maintain good robustness.

5.1 Models

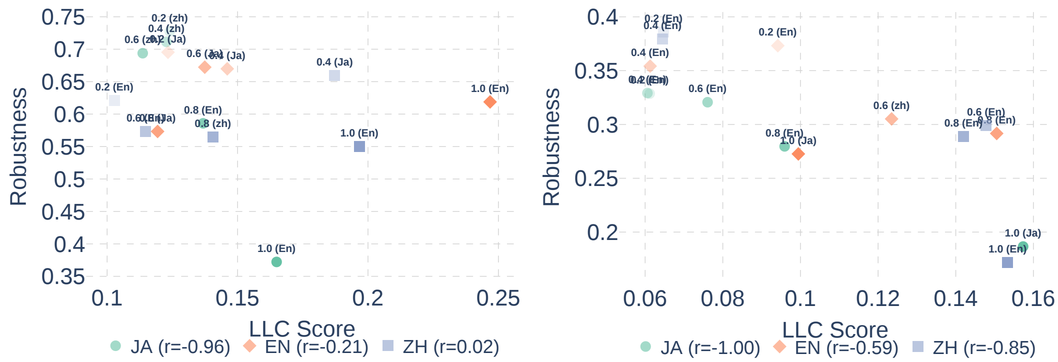
For our experiments, we select models that are primarily pre-trained on English, Chinese, and Japanese data, including Gemma3 [27], Qwen2.5 [28], and LLM-jp [4] respectively. We select them to evaluate the impact of performing downstream tasks in languages that differ from each model's dominant latent language, providing a more comprehensive understanding of consistency and robustness. Appendix B.3 provides detailed settings and model configurations.

5.2 Datasets

We utilize the dataset described in Section 4. For translation tasks, possible Source-Target language pairs include {En, Ja, Zh}-{En, Ja, Zh}, but we exclude tasks where the source and target languages are the same like {En}-{En} or where the token language cannot be clearly identified, such as {Ja}-{Zh} or {Zh}-{Ja} translations. We choose these three languages because English has the largest share of training data [29, 30]. Japanese, as a typologically distinct language, exhibits substantial orthographic and structural differences from English [31, 32]. Meanwhile, Chinese, with a data volume comparable to English, is also included for this analysis. For each language pair, we create 2,000 samples, and for the geo-culture task, we prepare 2,000 samples for each language as well.

5.3 Adversarial Prompts Settings

As described in Section 3.4, we attempt to disrupt consistency by injecting adversarial prompts. We systematically examine the proportion of the input length occupied by these adversarial prompts, as detailed in Appendix B.4, using ratios of 20%, 40%, 60%, 80%, and 100% of the model's maximum input token length (e.g., 32,768 tokens for Qwen2.5). By varying these ratios incrementally, our work makes it possible to systematically analyze the impact of adversarial prompts. The input languages include Japanese, Chinese, and English. Appendix B.7 provides examples of the prompts, and Table 1 shows all experimental settings.



(a) Src: En, Trg: Ja

(b) Src: En, Trg: Zh

Figure 2: Correlation (r) between language consistency and robustness in translation tasks using LLM-jp-3. For example, Ratio (Lang) indicates when inserting Japanese adversarial prompts with (Ratio)%, the model thinks in the latent language (Lang).

5.4 Templates

The model input consists of the problem statements created in Section 4. Following prior work [1], we conduct experiments using 4-shot to improve the model's ability to accurately fill in the blanks.

6 Results

Figures 2 and 3 present the experimental results for the translation task, while Figures 4 and 5 show the results for the geo-culture task. In these plots, the x-axis represents the LLC Score, and the y-axis represents the robustness. Each point represents the result on a specific task. 1.0 (Ja) indicates the result when adversarial prompts in English make up 100% of the input, resulting in the latent language being Japanese (Ja). We denote the correlation between Robustness and the LLC Score as r. If our hypothesis is correct, the data points should align along the line $y = -x$, indicating that LLC Score should lead to a corresponding drop in robustness.

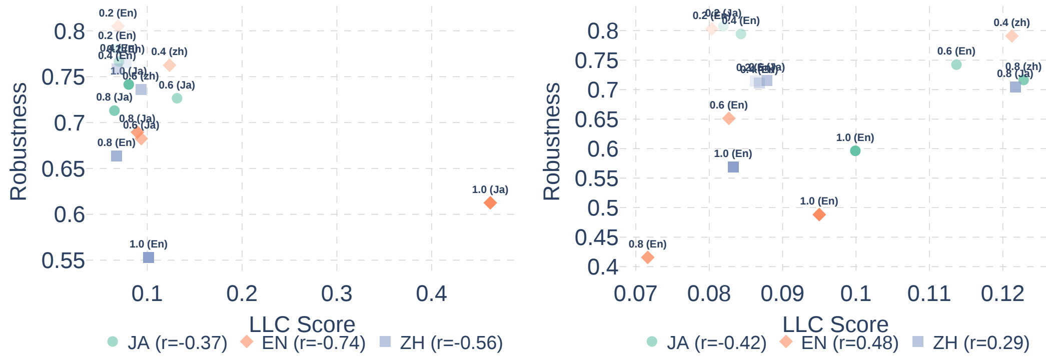
In Figure 2 (b), when looking at the results for Chinese adversarial prompts, we observe that lower ratios of adversarial prompts, e.g., 0.2 (En) or 0.4 (En), lead to higher Robustness and lower LLC Score. As the ratio increases, both metrics degrade, indicating a negative correlation ($r = -0.85$). Similar behavior appears with adversarial prompts in other languages as well ($r = -1.0$ for Japanese, and $r = -0.59$ for English). In contrast, Figures 3 and 5 show different trends, i.e., ($r = 0.29$ for Chinese in Figure 3, and $r = 0.98$ for English and Chinse in Figure 5), making the overall pattern less clear. Looking at each point, we observe that the latent language remains unchanged regardless of whether a correlation exists or not. This observation suggests that models do not necessarily need to operate in their proficient latent language to perform downstream tasks effectively. These results contradict the initial hypothesis that disrupting thought in latent language affects downstream task performance.

7 Discussion

7.1 Correlation Between Robustness and Latent Language

Figures 2, 3, and 4 illustrate the correlation between consistency and robustness in translation and geo-culture tasks. The results tend to align with the ideal line $y = -x$, indicating that a loss of robustness is often accompanied by a disruption in consistency. For example, in Figure 4, when the adversarial prompts are in English, the points align well along the line $y = -x$.

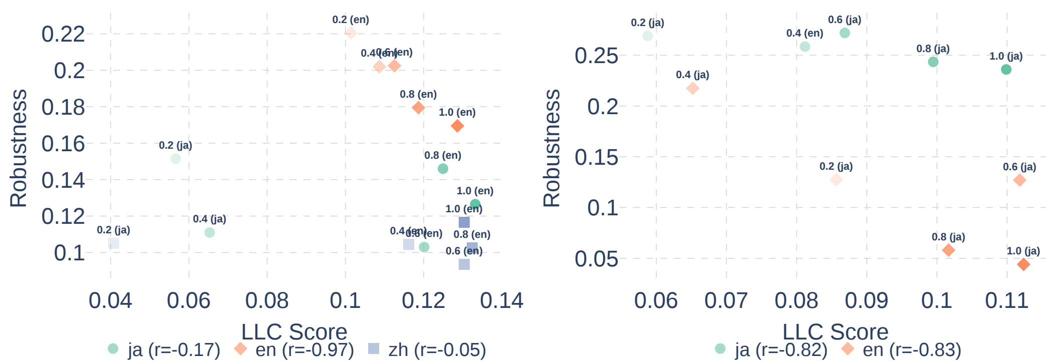
Focusing on the y-axis, we observe that when the question language or the source language matches the adversarial prompt, the model tends to maintain higher robustness. In Figure 4 (b), where



(a) Src: Ja, Trg: En

(b) Src: Zh, Trg: En

Figure 3: Same as Figure 2, but (a) is {Ja}-{En} and (b) is {Zh}-{En}.



(a) Question in English

(b) Question in Japanese

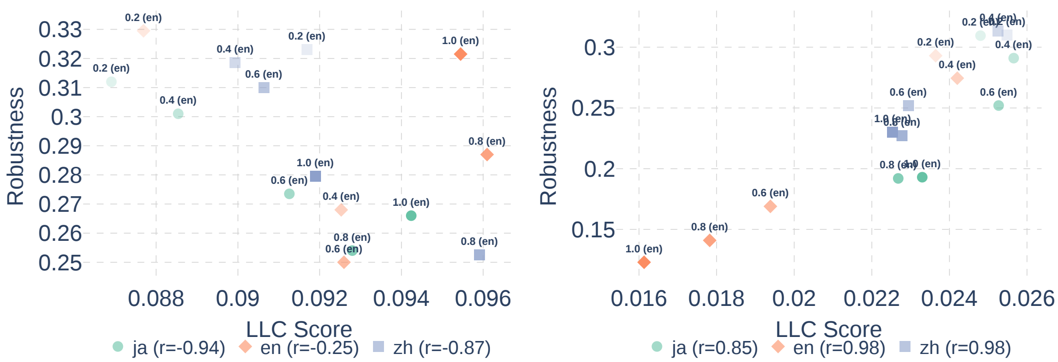
Figure 4: Correlation (r) between consistency and robustness in geo-culture tasks using LLM-jp-3. For example, Ratio (Lang) indicates when inserting English adversarial prompts with (Ratio)%, the model thinks in the latent language (Lang).

the question is in English, the highest accuracy, i.e., the highest robustness, is achieved when the adversarial prompt is also in English. We can observe a similar tendency in (a) as well.

In contrast, adversarial prompts in different languages generally lead to lower robustness. This suggests that mismatches between the question language and the adversarial prompt can largely disrupt downstream task performance. However, examining the x-axis reveals a different pattern: even when the adversarial prompt language differs from the question language, the expected consistency breakdown does not always occur. This finding indicates that models can flexibly adjust to different latent languages depending on the input language.

7.2 Impact of Adversarial Prompt Ratio on Consistency

We analyze the impact of adversarial prompt ratio on consistency. In many cases, increasing this ratio of adversarial prompts clearly affects internal consistency as shown in Figures 2, 3, and 4. However, when the adversarial prompt language differs from the question language, the latent language processing appears unaffected. For example, in Figure 4 (a), since the question is in English, we would ideally expect the LLC Score to be low when the latent language is English. However, the results show that the latent language is Japanese instead. This result implies that consistency in latent language is resilient to adversarial prompts, supporting the view that models do not necessarily need to operate in their preferred latent language to maintain internal consistency.



(a) Question in English

(b) Question in English

Figure 5: Same as Figure 2, but with different models. (a) shows Qwen2.5 and (b) shows Gemma3.

(a) Question in Japanese w/o adversarial

(b) Question and adversarial in Japanese

Figure 6: Results of the LLM-jp-3 model on the geo-culture task. The left panel shows results without noise, while the right panel shows results with 100% noise. Out of 2,000 samples, green points indicate correct predictions, while red points indicate incorrect ones.

7.3 Impact of Adversarial Prompt Ratio on Robustness

Increasing the proportion of adversarial prompts largely impacts model robustness across most models. This observation aligns with prior studies [16, 17], which have shown that increasing noise in the input degrades downstream task performance.

7.4 Data Distribution Analysis

To investigate how latent language disruptions relate to internal consistency and model confidence, we analyze the correlation between KL divergence in each adjacent layer and output probabilities at the question level. Figure 6 presents a scatter plot of KL values and output probabilities for each question in the geo-culture task. The x-axis represents KL divergence values (lower is better), while the y-axis shows output probabilities (higher is better). The KL values are computed for each layer using the LogitLens, calculating the KL divergence between consecutive layers ($KL_{l,l+1}$ in Section 3). Green circles (●) indicate correct responses, while red crosses ($\times$) represent incorrect ones. Figure (a) includes no adversarial prompts, while (b) includes 100% adversarial prompts. According to Table 2, the accuracy for these two settings is 0.25 in left and 0.24 in right respectively. This analysis demonstrates that inserting adversarial prompts reduces model confidence and robustness, as indicated by the overall downward shift in output probabilities.

8 Conclusion

In our study, we systematically analyze the impact of consistency in latent language on downstream task performance. Specifically, we focus on understanding how disruptions in consistency of the latent language influence their reasoning processes and overall task performance. To achieve this, we insert adversarial prompts containing multiple languages, deliberately disrupting the models' internal representations and quantitatively measuring the resulting performance shifts.

Our findings indicate that, contrary to our initial hypothesis, maintaining a consistent latent language is not always necessary for achieving best task performance. Models can flexibly adapt to varying input languages near the final layers, suggesting that their internal reasoning processes are more resilient to linguistic perturbations than previously thought. However, tasks may be heavily dependent on precise linguistic alignment, such as translation, are particularly sensitive to disruptions by adversarial prompts. In these cases, the introduction of adversarial prompts from less familiar languages largely degrades performance, likely due to destabilization of the models' internal structures.

References

[1] Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in English? on the latent language of multilingual transformers. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[2] Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. Beyond english-centric llms: What language do multilingual language models think in? *arXiv preprint arXiv:2408.10811*, 2024.

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

[4] Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, et al. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms. *arXiv preprint arXiv:2407.03963*, 2024.

[5] Zheng-Xin Yong, M. Farid Adilazuarda, Jonibek Mansurov, Ruochen Zhang, Niklas Muennighoff, Carsten Eickhoff, Genta Indra Winata, Julia Kreutzer, Stephen H. Bach, and Alham Fikri Aji. Crosslingual reasoning through test-time scaling, 2025.

[6] Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. mCSQA: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14182–14214, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[7] Shintaro Ozaki, Kazuki Hayashi, Miyu Oba, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. Bqa: Body language question answering dataset for video large language models. *arXiv preprint arXiv:2410.13206*, 2024.

[8] Lisa Schut, Yarin Gal, and Sebastian Farquhar. Do multilingual llms think in english? *arXiv preprint arXiv:2502.15603*, 2025.

[9] Katharina Hämmerl, Bjoern Deiseroth, Patrick Schramowski, Jindřich Libovický, Constantin Rothkopf, Alexander Fraser, and Kristian Kersting. Speaking multiple languages affects the moral bias of language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2137–2156, Toronto, Canada, July 2023. Association for Computational Linguistics.

[10] IO Gallegos, RA Rossi, J Barrow, MM Tanjim, S Kim, F Dernoncourt, T Yu, R Zhang, and NK Ahmed. Bias and fairness in large language models: A survey. arxiv, 2023.

[11] Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore, December 2023. Association for Computational Linguistics.

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

[13] Huiyuan Lai and Malvina Nissim. mCoT: Multilingual instruction tuning for reasoning consistency in language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12012–12026, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[14] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[15] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June 2021. Association for Computational Linguistics.

[16] Karan Goel, Nazneen Fatema Rajani, Jesse Vig, Zachary Taschdjian, Mohit Bansal, and Christopher Ré. Robustness gym: Unifying the NLP evaluation landscape. In Avi Sil and Xi Victoria Lin, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 42–55, Online, June 2021. Association for Computational Linguistics.

[17] Daniel Khashabi, Tushar Khot, and Ashish Sabharwal. More bang for your buck: Natural perturbation for robust question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 163–170, Online, November 2020. Association for Computational Linguistics.

[18] nostalgebraist. Interpreting gpt: the logit lens. `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`, August 2020. LessWrong.

[19] Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens, 2023.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[21] Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

[22] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics.

[23] Wilson L Taylor. “cloze procedure”: A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433, 1953.

[24] Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. Personagym: Evaluating persona agents and llms. *arXiv preprint arXiv:2407.18416*, 2024.

[25] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100, 2018.

[26] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[27] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[28] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

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

[30] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. arxiv 2022. *arXiv preprint arXiv:2204.02311*, 10:1, 2022.

[31] M. Haspelmath, M.S. Dryer, D. Gil, and B. Comrie. *The World Atlas of Language Structures*. Number v. 1 in Oxford linguistics. OUP Oxford, 2005.

[32] Hiroyuki Deguchi, Go Kamoda, Yusuke Matsushita, Chihiro Taguchi, Masaki Waga, Kohei Suenaga, and Sho Yokoi. Softmatcha: A soft and fast pattern matcher for billion-scale corpus searches. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*, 2025.

[33] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[34] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems*, 35:30318–30332, 2022.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We stated our contributions and scope of in the abstract and introduction, and the statements are supported by the experimental results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed our limitations in Appendix A.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This is an interpretation paper without any theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provided detailed model settings in Appendix B.3, detailed model name we used in Appendix 4, and detailed prompts which we instructed in Appendix B.7.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: As we mentioned in Appendix B.3, due to double blind anonymized policy, we will release upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We mentioned in Section 4, Section 5.1, Appendix B.3, and Appendix 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not report the error bars for the measurement of effectivity and retainability, as it is required to run the pretraining experiments multiple times with different seeds, which is not feasible due to the restriction of computational resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We mentioned in Appendix B.5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have confirmed that this work is done following the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discussed our impact in Abstract, Introduction, and Conclusion.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work poses no risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cited and explicitly mentioned the original papers that produced the pretrained models or datasets that are used in this work.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We mentioned carefully in Section 4 and Appendix B.7.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We mentioned in Appendix B.6.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Limitations

Our study has several limitations. First, there are various approaches for determining the dominant language. In our study, we defined the LLC Score and identified the language with the lowest score as the dominant language, treating it as the latent language. However, other, potentially better, approaches may exist. Our approach was selected because it showed the strongest correlation with other potential methods, such as averaging KL scores which we mentioned in 7.4.

Moreover, our study investigates the correlation between robustness and consistency through the use of adversarial prompts. This raises the possibility that the observed effects may depend on the specific adversarial prompts used.

Lastly, it would be valuable to extend this analysis to larger models as we mentioned in Appendix B.10 partially and datasets.

B Appnendix

B.1 Questions Diversity in Datatasets

Self-BLEU [25]is a metric commonly used to evaluate the diversity of generated text within a dataset. It measures the extent to which each sample in a set is distinct from the others, providing insight into the range of expressions present in the generated outputs. A lower Self-BLEU score indicates higher diversity, while a higher score suggests redundancy or repetition in the dataset.

The Self-BLEU score for a single question q_i is calculated as:

$$\text{Self-BLEU}(q_i) = \text{BLEU}(q_i, \mathcal{Q} \setminus \{q_i\})$$

Here, q_i represents a single question from the dataset, $\mathcal{Q}$ denotes the complete set of questions in the dataset, and $\mathcal{Q} \setminus \{q_i\}$ is the set of all questions excluding q_i. The overall Self-BLEU for the entire dataset is then the average of the individual Self-BLEU scores:

$$\text{Self-BLEU}(\mathcal{Q}) = \frac{1}{|\mathcal{Q}|} \sum_{q_i \in \mathcal{Q}} \text{Self-BLEU}(q_i)$$

where $|\mathcal{Q}|$ is the total number of questions in the dataset. This formulation effectively captures the overlap between questions, providing a quantitative measure of diversity. A more diverse dataset will naturally yield lower Self-BLEU scores, reflecting the presence of a broader range of expressions and phrasing within the question set.

To ensure more diverse problem statements generated by GPT-4, a category list was created, from which approximately 20 more specific categories were randomly selected to increase variability. The results are presented in Table 3.

Categories
Capital City, Official Language, National Currency, Head of State, Independence Day, Major Religion, National Anthem, Famous Landmark, National Dish, Traditional Clothing, Flag Colors, Historical Figure, Major River, Mountain Range, Neighboring Countries, Climate, Population, Area (km^2), UNESCO World Heritage Site, Government Type

Table 3: A result of Self-BLEU by 4-gram. This result shows that our created datasets have more diverse than preliminary ones which we ask GPT-4o to generate questions by same prompts.

	Self-BLEU (↓)			
Task	**Preliminary**		**Ours**	
	Qustion Language	**Score**	**Question Language**	**Score**
Culture	English	0.9976	English	**0.7336**
	Japanese	0.9707	Japanese	**0.6007**
	Chinese	0.9653	Chinese	**0.6368**
Translation	Japanese → English	0.9559	Japanese → English	**0.6180**
	Chinese → English	0.9273	Chinese → English	**0.7807**
	English → Japanese	0.9940	English → Japanese	**0.6925**
	English → Chinese	0.9950	English → Chinese	**0.6887**

B.2 Mathmetical Approach for LogitLens

We provide a detailed explanation of LogitLens [18] using mathematical notation. Here, we define each variable used in Equations (1)–(3) for the LogitLens analysis. In Equation (1),

$$\mathbf{h}_{\ell+1} = \mathbf{h}_\ell + \mathcal{F}_\ell(\mathbf{h}_\ell)$$

$\mathbf{h}_\ell \in \mathbb{R}^d$ is the hidden state at layer ℓ, and $\mathcal{F}_\ell(\cdot)$ is the residual streams. This equation describes the standard residual update in Transformer models.

In Equation (2),

$$\mathcal{M}_{>\ell}(\mathbf{h}_\ell) = \mathbf{W}_U^\top \ \mathrm{LayerNorm}\left[\mathbf{h}_\ell + \sum_{\ell'=\ell}^{L} \mathcal{F}_{\ell'}(\mathbf{h}_{\ell'})\right]$$

$\mathcal{M}_{>\ell}(h_\ell)$ denotes the final logits if forward computation continues from layer ℓ. $W_U \in \mathbb{R}^{d\times|V|}$ is the unembedding matrix, and $|V|$ is the vocabulary size. The sum represents all residual updates after layer ℓ.

In Equation (3),

$$\mathrm{LogitLens}(\mathbf{h}_\ell) = \mathbf{W}_U^\top \, \mathrm{LayerNorm}(\mathbf{h}_\ell)$$

The LogitLens removes all future residuals and directly maps h_ℓ to vocabulary logits. It shows what the model "knows" at layer ℓ, independently of later layers.

B.3 Detailed Models Settings

Details of the models used in the experiments are listed in Table 4. To ensure reproducibility, the seed was set to 42, and top_p was fixed at 0.0. The implementation utilized the Transformers library [33] and bitsandbytes [34]. Note that all code and datasets will be available upon acceptance.

B.4 Details of Input Token Length

As mentioned in Section 5.3, our study aimed to fill the maximum input length of the model with adversarial prompts. The number of tokens used for this purpose is listed in Table 5 below. In addition to these token counts, the input also includes 4-shot examples and the question text.

Table 4: Detailed model names, parameters, and max input tokens.

Model	Params	Max Tokens	HuggingFace Name / OpenAI Name
Gemma3	1B	32,768	google/gemma-3-1b-it
Qwen2.5	1.5B	32,768	Qwen/Qwen2.5-1.5B-Instruct
LLM-jp-3	1.8B	4,096	llm-jp/llm-jp-3-1.8b-instruct
Qwen2.5 (3B)	3B	32,768	Qwen/Qwen2.5-3B-Instruct
LLM-jp-3 (7.2B)	7.2B	4,096	llm-jp/llm-jp-3-7.2b-instruct
GPT-4o	–	128k	gpt-4o-2024-11-20

Table 5: Adversarial Prompt Token Counts for Different Ratios

Model	Lang	Ratio					Max Token Length
		0.2	0.4	0.6	0.8	1.0	
LLM-jp-3	Ja	845	1,664	2,483	3,301	3,925	4,096
	En	842	1,661	2,479	3,298	3,921	
	Zh	848	1,667	2,485	3,304	3,928	
Qwen2.5	Ja	6,584	13,138	19,691	26,245	32,631	32,768
	En	6,570	13,119	19,670	26,220	32,602	
	Zh	65,71	13,125	19,678	26,232	32,618	
Gemma3	Ja	6,578	13,130	19,681	26,233	32,617	32,768
	En	6,571	13,121	19,671	26,221	32,604	
	Zh	6,571	13,122	19,672	26,222	32,605	

B.5 Inference Settings

In our study, as far as inference which needs to use GPUs, all experiments were conducted on a single NVIDIA RTX A6000 and NVIDIA A100 80GB PCIe.

B.6 AI Assistant Tools

We used ChatGPT [2] and DeepL [3] to translate sentences to English and accelerate our research.

B.7 Detailed Prompts

Below are examples of the prompts used in the experiments. As previously mentioned, the experiments were conducted with 4-shots, following prior studies. The placeholder {variable} will be replaced with actual questions. Due to their large size, the overall adversarial prompts will be published as URLs upon acceptance.

B.7.1 Question Generation

Prompt for geo-culture and translation task

```
{Adversarial prompts} (if needed)


「晴れ」を中国語に翻訳してください。, answer:  晴天
What type of government does the USA have?, answer:  democracy
「春」を英語に翻訳してください。, answer:  spring
What is the capital city of Japan?, answer:  Tokyo

{Question}, answer:
```

[2] https://chatgpt.com/
[3] https://www.deepl.com/ja/translator

Prompt for creating question with related to geo-culture task by gpt-4o

```
Create a question related to the geography of {country}.
Please generate with related to {category}.
The question MUST consist of a single sentence.

# ** Important **
- The answer format must follow the format below "# Example".
- Only 1 sentence is needed for the response.
- Create questions where the answer will be only one word.
- Please generate the question end with ", answer:  " even though language is
Japanese or Chinese.

# Example:
What is the capital of {country}?, answer:
What is the highest mountain in {country}?, answer:
What is the official language of {country}?, answer:
What is the currency of {country}?, answer:
```

Prompt for creating question with related to translation task by gpt-4o

```
Create a question related to the translation task from {src_lang} to
{trg_lang}.
Please generate the question in {src_lang} with related to {category}.

For example, you can create a question like "Please translate "Water" into
Chinese., answer:  ".
This example is that the question is in English, whereas the answer will be
Chinese.
The question MUST ONLY consist of a single sentence.

# ** Important **
- INPUT IS {src_lang}.
- Please say "Please translate XXX into {trg_lang}, answer:"
- The generation format must follow the format below "# Example".
- Only 1 sentence is needed for the response.
- Create questions where the answer will be only one word.
- Do NOT include an ANSWER in the generation.
- Please generate the sentence end with ", answer:  " EVEN IF Chinese or
Japanese.

# Example:
"東京"を英語に翻訳してください。, answer:
Please translate "Beijing" into Chinese., answer:
Please translate "Washington" into Japanese., answer:
Please translate "Sun" into Japanese., answer:
```

B.7.2 Adversarial Prompts

Adversarial prompts in Japanese

あなたは下記に示す日本人のように振る舞ってください。また日本人の思考で考えてください。

日本の文化や背景、特徴、歴史について説明します。
日本は、東アジアに位置する島国で、太平洋に面し、主に本州、北海道、九州、四国の4つの主要な島と、それに付随する約6,800以上の小さな島々から構成されています。国土の多くが山地で、自然災害、特に地震や台風が多い地域でもあります。これらの自然環境は、日本人の生活様式や精神性、信仰に大きな影響を与えてきました。
まず、日本の歴史について触れます。日本の歴史は数千年にわたり、時代ごとに独自の文化が形成されてきました。
...... (Much more long sentences.)

上記の内容を基に日本人のように振る舞ってください。

Adversarial prompts in Chinese

请像中国人一样表现。

中国，全称为中华人民共和国，是位于的一个幅员辽阔、历史悠久、文化多元的国家。它拥有世界上最多的人口（约14亿），以及广阔的领土，横跨东西约5000公里，南北跨度也极大，地理、气候和文化的多样性使得中国成为世界上最为复杂和丰富的文明体之一。以下将从中国的文化、背景、特点与历史等方面进行尽可能详细的描述。
...... (Much more long sentences.)

在提及上述内容时，请表现得像个中国人

Adversarial prompts in English

```
Please behave like a US person readin below sentences.  Please think in
English native speaker.

The United States of America (USA) is a vast and diverse country located in
North America, bordered by Canada to the north, Mexico to the south, the
Atlantic Ocean to the east, and the Pacific Ocean to the west.  The U.S. is
a federal republic composed of 50 states, a federal district (Washington,
D.C.), and several territories.  Its enormous size, cultural plurality,
complex history, and global influence make it one of the most important and
influential countries in the modern world.
Let us now explore its culture, background, characteristics, and history in
as much detail as possible.
```
...... (Much more long sentences.)

```
Please behave like a US person referring to the above content.
```

B.8 Examples of Questions

Below are examples of questions generated using the method described in Section 4. Note that (answer) is excluded during input.

Examples of Questions in translation task

「晴れ」を中国語に翻訳してください。 , answer: （晴天）
「一」を英語に翻訳してください。 , answer: （one）
请将“星期一”翻译成日语, answer: （月曜日）
"请将“交通”翻译成英语, answer: （transportation）
"キャベツ"を中国語に翻訳してください。 , answer: （卷心菜）

Examples of Questions in geo-culture task

Who is the current Prime Minister of Japan?, answer: （Kishida）
現在の日本の首相は誰ですか？, answer: （岸田）
日本目前的首相是谁?， answer: （岸田）

What is the capital city of China?, answer: （Beijing）
中国の首都はどこですか?, answer: （北京）
中国的首都是什么？ , answer: （北京）

What is considered the national dish of the United States?, answer: （Hamburger）

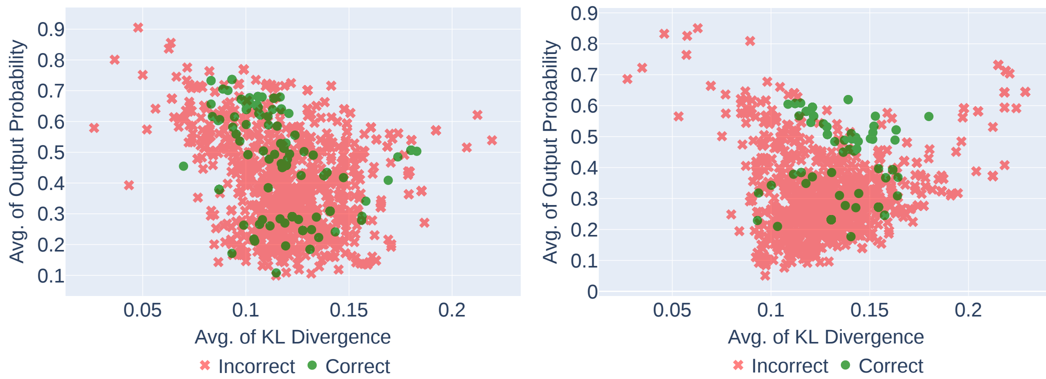
アメリカ合衆国の国民食と見なされるのは何ですか?, answer: （ハンバーガー）
么被认为是美国的国菜?, answer: （汉堡）

B.9 Other Results

Below are additional experimental results. As these results indicate, it is not possible to determine whether the inclusion of adversarial prompts directly affects consistency based on this table. However, it is evident that robustness and confidence tend to decline.

(a) Question in Japanese w/o adversarial (b) Question and adversarial in Japanese

Figure 7: Results of the LLM-jp-3 model on the geo-culture task. The left panel shows results without noise, while the right panel shows results with 100% noise. Out of 2,000 samples, green points indicate correct predictions, while red points indicate incorrect ones.



(a) Question in Japanese w/o adversarial (b) Question and adversarial in Japanese

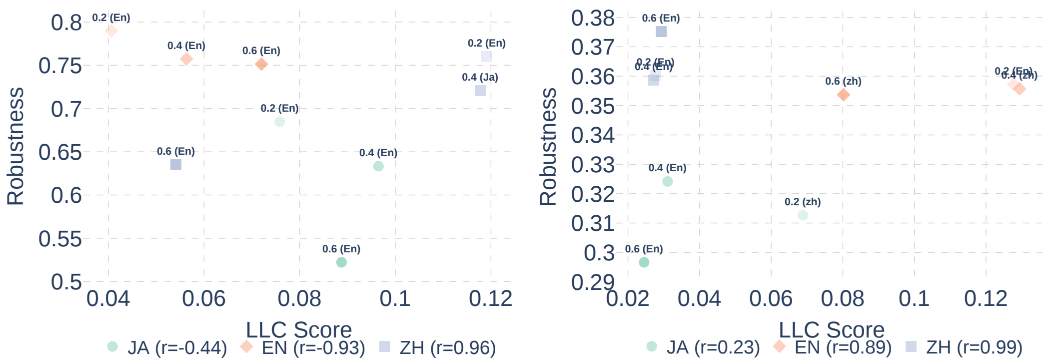
Figure 8: Results of Qwen2.5 on the geo-culture task. The figure is interpreted in the same way as Figure 7.

Table 6: Correlation between LLC Score and robustness in translation tasks. The value r represents the correlation coefficient for each noise ratio.

Model	Source	Target	Adversarial	LLC Score (↓)					Robustness (↑)					r
				0.2	0.4	0.6	0.8	1.0	0.2	0.4	0.6	0.8	1.0	
LLM-jp-3	Ja	En	Ja	En$_{0.07}$	En$_{0.07}$	Ja$_{0.13}$	Ja$_{0.07}$	Ja$_{0.08}$	0.78	0.77	0.73	0.71	0.74	-0.37
			En	En$_{0.07}$	Zh$_{0.12}$	Ja$_{0.09}$	Ja$_{0.09}$	Ja$_{0.46}$	0.81	0.76	0.68	0.69	0.61	-0.74
			Zh	En$_{0.08}$	En$_{0.07}$	Zh$_{0.09}$	En$_{0.07}$	En$_{0.10}$	0.77	0.76	0.74	0.66	0.55	-0.56
	En	Ja	Ja	Zh$_{0.12}$	Zh$_{0.12}$	Zh$_{0.11}$	En$_{0.14}$	En$_{0.16}$	0.73	0.71	0.69	0.59	0.37	-0.96
			En	Ja$_{0.12}$	Ja$_{0.15}$	Ja$_{0.14}$	Ja$_{0.12}$	En$_{0.25}$	0.70	0.67	0.67	0.57	0.62	-0.21
			Zh	En$_{0.10}$	Ja$_{0.19}$	En$_{0.11}$	Zh$_{0.14}$	En$_{0.20}$	0.62	0.66	0.57	0.56	0.55	0.02
		Zh	Ja	En$_{0.06}$	En$_{0.06}$	En$_{0.08}$	En$_{0.10}$	Ja$_{0.16}$	0.33	0.33	0.32	0.28	0.19	-1.00
			En	En$_{0.09}$	En$_{0.06}$	Zh$_{0.12}$	En$_{0.15}$	Ja$_{0.10}$	0.37	0.35	0.31	0.29	0.27	-0.59
			Zh	En$_{0.06}$	En$_{0.06}$	En$_{0.15}$	En$_{0.14}$	En$_{0.15}$	0.39	0.38	0.30	0.29	0.17	-0.85
	Zh	En	Ja	Ja$_{0.08}$	En$_{0.08}$	En$_{0.11}$	Zh$_{0.12}$	En$_{0.10}$	0.81	0.79	0.74	0.72	0.60	-0.42
			En	En$_{0.08}$	Zh$_{0.12}$	En$_{0.08}$	En$_{0.07}$	En$_{0.09}$	0.80	0.79	0.65	0.42	0.49	0.48
			Zh	En$_{0.09}$	En$_{0.09}$	Ja$_{0.09}$	Ja$_{0.12}$	En$_{0.08}$	0.71	0.71	0.72	0.70	0.57	0.29
Qwen2.5	Ja	En	Ja	En$_{0.00}$	En$_{0.00}$	En$_{0.00}$	En$_{0.00}$	En$_{0.00}$	0.84	0.85	0.85	0.84	0.84	N.A.
			En	En$_{0.00}$	En$_{0.00}$	En$_{0.00}$	En$_{0.05}$	En$_{0.05}$	0.75	0.76	0.74	0.65	0.67	-0.99
			Zh	En$_{0.00}$	En$_{0.00}$	En$_{0.00}$	En$_{0.00}$	En$_{0.00}$	0.79	0.78	0.73	0.69	0.69	N.A.
	En	Ja	Ja	En$_{0.00}$	En$_{0.00}$	En$_{0.00}$	En$_{0.00}$	En$_{0.00}$	0.00	0.00	0.00	0.00	0.00	N.A.
			En	En$_{0.00}$	En$_{0.00}$	En$_{0.00}$	En$_{0.00}$	En$_{0.00}$	0.00	0.00	0.00	0.00	0.00	N.A.
			Zh	En$_{0.00}$	En$_{0.00}$	En$_{0.00}$	En$_{0.00}$	En$_{0.00}$	0.00	0.00	0.00	0.00	0.00	N.A.
		Zh	Ja	En$_{0.00}$	En$_{0.00}$	En$_{0.00}$	En$_{0.00}$	En$_{0.00}$	0.02	0.01	0.00	0.00	0.00	N.A.
			En	En$_{0.00}$	En$_{0.00}$	En$_{0.00}$	En$_{0.00}$	En$_{0.00}$	0.01	0.01	0.02	0.00	0.00	N.A.
			Zh	En$_{0.00}$	En$_{0.00}$	En$_{0.00}$	En$_{0.00}$	En$_{0.00}$	0.01	0.01	0.00	0.00	0.00	N.A.
	Zh	En	Ja	En$_{0.00}$	En$_{0.08}$	En$_{0.06}$	En$_{0.09}$	En$_{0.09}$	0.46	0.50	0.33	0.55	0.54	0.38
			En	En$_{0.00}$	En$_{0.08}$	En$_{0.09}$	En$_{0.08}$	En$_{0.00}$	0.68	0.61	0.57	0.65	0.75	-0.86
			Zh	En$_{0.00}$	En$_{0.00}$	En$_{0.00}$	En$_{0.06}$	En$_{0.07}$	0.85	0.83	0.78	0.73	0.74	-0.88
Gemma3	Ja	En	Ja	En$_{0.02}$	En$_{0.02}$	En$_{0.02}$	En$_{0.02}$	En$_{0.02}$	0.78	0.77	0.76	0.30	0.59	0.76
			En	En$_{0.06}$	En$_{0.06}$	En$_{0.03}$	En$_{0.06}$	En$_{0.03}$	0.80	0.82	0.78	0.73	0.68	0.49
			Zh	En$_{0.03}$	En$_{0.08}$	En$_{0.07}$	En$_{0.24}$	En$_{0.10}$	0.80	0.79	0.69	0.24	0.47	-0.92
	En	Ja	Ja	En$_{0.05}$	En$_{0.02}$	En$_{0.05}$	En$_{0.01}$	En$_{0.00}$	0.27	0.14	0.14	0.08	0.07	0.83
			En	En$_{0.05}$	En$_{0.05}$	En$_{0.05}$	En$_{0.00}$	En$_{0.00}$	0.30	0.18	0.20	0.11	0.14	0.81
			Zh	En$_{0.06}$	En$_{0.05}$	En$_{0.05}$	En$_{0.05}$	En$_{0.05}$	0.32	0.28	0.17	0.17	0.15	0.77
		Zh	Ja	En$_{0.05}$	En$_{0.04}$	Zh$_{0.02}$	En$_{0.04}$	En$_{0.04}$	0.22	0.16	0.15	0.05	0.04	0.30
			En	En$_{0.08}$	En$_{0.05}$	En$_{0.04}$	En$_{0.01}$	En$_{0.01}$	0.23	0.15	0.14	0.02	0.02	0.98
			Zh	En$_{0.10}$	En$_{0.06}$	En$_{0.08}$	En$_{0.05}$	En$_{0.05}$	0.24	0.24	0.13	0.13	0.06	0.59
	Zh	En	Ja	En$_{0.00}$	En$_{0.00}$	En$_{0.00}$	En$_{0.01}$	En$_{0.01}$	0.84	0.81	0.79	0.47	0.55	-0.99
			En	En$_{0.03}$	En$_{0.03}$	En$_{0.00}$	En$_{0.02}$	En$_{0.02}$	0.82	0.85	0.73	0.71	0.55	0.33
			Zh	En$_{0.00}$	En$_{0.00}$	En$_{0.01}$	En$_{0.03}$	En$_{0.00}$	0.76	0.62	0.48	0.37	0.14	-0.23

B.10 Results by Larger LLMs

In our study, we used relatively small LLMs for evaluation, as described in Appendix 4. This choice was made to ensure that inference with maximum noise relative to token length remains feasible on a single GPU. Larger models present significant challenges in terms of computational resources, making such evaluations difficult. Nevertheless, we conducted experiments with larger models, including Qwen (3B) and LLM-jp-3 (7.2B), by reducing the noise ratio. As shown in Table 7 and Table 8, these models exhibited similar performance trends to their smaller counterparts. Future work aims to extend our work to larger models with more extensive noise injection using multiple GPUs.



(a) Src: En, Trg: Ja

(b) Src: En, Trg: Zh

Figure 9: Correlation between language consistency and robustness in translation tasks using LLM-jp-3 (7.2B). The color of each point represents the adversarial language, indicating the ratio and the dominant language in each setting.

Table 7: Correlation between LLC Score and robustness in geo-culture tasks. The value r represents the correlation coefficient for each noise ratio.

Model	Question	Adversarial	LLC Score (↓)			Robustness (↑)			r
			0.2	**0.4**	**0.6**	**0.2**	**0.4**	**0.6**	
LLM-jp-3 (7.2B)	Ja	Ja	Ja$_{0.13}$	Ja$_{0.12}$	En$_{0.07}$	0.26	0.26	0.29	-0.97
		En	Ja$_{0.12}$	Ja$_{0.12}$	Ja$_{0.09}$	0.28	0.28	0.27	1.00
	En	Ja	En$_{0.12}$	En$_{0.10}$	En$_{0.10}$	0.11	0.16	0.17	-1.00
		En	En$_{0.08}$	En$_{0.08}$	En$_{0.08}$	0.23	0.23	0.23	0.75
		Zh	En$_{0.11}$	En$_{0.10}$	En$_{0.10}$ s	0.11	0.19	0.19	-0.90
	Zh	En	Ja$_{0.11}$	Ja$_{0.09}$	Ja$_{0.07}$	0.08	0.08	0.03	0.89
		Zh	En$_{0.12}$	En$_{0.12}$	Ja$_{0.09}$	0.07	0.05	0.07	-0.56
Qwen2.5 (3B)	Ja	Ja	En$_{0.12}$	En$_{0.12}$	En$_{0.12}$	0.00	0.00	0.00	-0.05
		En	En$_{0.09}$	En$_{0.10}$	En$_{0.10}$	0.00	0.00	0.00	N.A.
	En	Ja	En$_{0.09}$	En$_{0.09}$	En$_{0.09}$	0.32	0.29	0.34	-1.00
		En	En$_{0.08}$	En$_{0.09}$	En$_{0.09}$	0.23	0.17	0.20	-0.95
		Zh	En$_{0.09}$	En$_{0.09}$	En$_{0.09}$	0.28	0.27	0.29	-0.95
	Zh	En	En$_{0.10}$	Zh$_{0.04}$	En$_{0.10}$	0.01	0.01	0.01	0.62
		Zh	Zh$_{0.04}$	Zh$_{0.04}$	Zh$_{0.04}$	0.00	0.01	0.01	0.94

Table 8: Correlation between LLC Score and robustness in translation tasks. The value r represents the correlation coefficient for each noise ratio.

Model	Source	Target	Adversarial	LLC Score (↓)			Robustness (↑)			r
				0.2	**0.4**	**0.6**	**0.2**	**0.4**	**0.6**	
LLM-jp-3 (7.2B)	Ja	En	Ja	$\text{En}_{0.07}$	$\text{En}_{0.07}$	$\text{En}_{0.07}$	0.78	0.75	0.73	-0.27
			en	$\text{Ja}_{0.11}$	$\text{En}_{0.12}$	$\text{En}_{0.13}$	0.77	0.74	0.74	-0.66
			Zh	$\text{Ja}_{0.10}$	$\text{Ja}_{0.12}$	$\text{Ja}_{0.12}$	0.72	0.69	0.72	-0.48
	En	Ja	ja	$\text{En}_{0.08}$	$\text{En}_{0.10}$	$\text{En}_{0.09}$	0.68	0.63	0.52	-0.44
			En	$\text{En}_{0.04}$	$\text{En}_{0.06}$	$\text{En}_{0.07}$	0.79	0.76	0.75	-0.93
			Zh	$\text{En}_{0.12}$	$\text{Ja}_{0.12}$	$\text{En}_{0.05}$	0.76	0.72	0.63	0.96
	En	Zh	Ja	$\text{Zh}_{0.07}$	$\text{En}_{0.03}$	$\text{En}_{0.02}$	0.31	0.32	0.30	0.23
			En	$\text{En}_{0.13}$	$\text{Zh}_{0.13}$	$\text{Zh}_{0.08}$	0.36	0.36	0.35	0.89
			zh	$\text{En}_{0.03}$	$\text{En}_{0.03}$	$\text{En}_{0.03}$	0.36	0.36	0.38	0.99
	Zh	En	Ja	$\text{Zh}_{0.10}$	$\text{En}_{0.13}$	$\text{En}_{0.13}$	0.72	0.73	0.71	0.04
			en	$\text{Ja}_{0.08}$	$\text{En}_{0.10}$	$\text{En}_{0.13}$	0.78	0.69	0.71	-0.70
			Zh	$\text{Zh}_{0.08}$	$\text{En}_{0.14}$	$\text{En}_{0.12}$	0.75	0.73	0.75	-0.73
Qwen2.5 (3B)	Ja	En	Ja	$\text{En}_{0.05}$	$\text{Zh}_{0.05}$	$\text{Zh}_{0.05}$	0.87	0.86	0.87	-0.98
			en	$\text{Zh}_{0.04}$	$\text{Zh}_{0.05}$	$\text{Zh}_{0.05}$	0.85	0.83	0.83	-0.95
			Zh	$\text{En}_{0.05}$	$\text{En}_{0.06}$	$\text{En}_{0.05}$	0.85	0.88	0.86	0.98
	En	Ja	Ja	$\text{zh}_{0.05}$	$\text{En}_{0.05}$	$\text{zh}_{0.05}$	0.02	0.02	0.04	-0.95
			En	$\text{zh}_{0.04}$	$\text{zh}_{0.04}$	$\text{zh}_{0.04}$	0.00	0.00	0.00	-0.98
			Zh	$\text{En}_{0.06}$	$\text{En}_{0.05}$	$\text{En}_{0.04}$	0.02	0.00	0.00	0.84
	En	Zh	Ja	$\text{zh}_{0.05}$	$\text{zh}_{0.05}$	$\text{zh}_{0.05}$	0.02	0.02	0.02	-0.80
			En	$\text{zh}_{0.04}$	$\text{En}_{0.05}$	$\text{zh}_{0.04}$	0.02	0.02	0.02	-0.35
			zh	$\text{En}_{0.05}$	$\text{Zh}_{0.05}$	$\text{Zh}_{0.05}$	0.02	0.02	0.02	1.00
	Zh	En	Ja	$\text{En}_{0.06}$	$\text{Zh}_{0.05}$	$\text{Zh}_{0.05}$	0.87	0.86	0.86	1.00
			en	$\text{Zh}_{0.05}$	$\text{Zh}_{0.05}$	$\text{Zh}_{0.05}$	0.85	0.86	0.88	0.83
			Zh	$\text{Zh}_{0.06}$	$\text{En}_{0.05}$	$\text{Zh}_{0.06}$	0.89	0.88	0.89	0.61