Analyzing LLMs' Knowledge Boundary Cognition Across Languages Through the Lens of Internal Representations

Chenghao Xiao[1,2,†] **Hou Pong Chan**[1,‡] **Hao Zhang**[1,‡] **Mahani Aljunied**[1]
Lidong Bing[1] **Noura Al Moubayed**[2] **Yu Rong**[1,3]
[1] DAMO Academy, Alibaba Group
[2] Department of Computer Science, Durham University
[3] Hupan Lab, Hangzhou, China
{xiaochenghao.xch,houpong.chan,hz.hhea2e}@alibaba-inc.com

Abstract

While understanding the knowledge boundaries of LLMs is crucial to prevent hallucination, research on the knowledge boundaries of LLMs has predominantly focused on English. In this work, we present the first study to analyze how LLMs recognize knowledge boundaries across different languages by probing their internal representations when processing known and unknown questions in multiple languages. Our empirical studies reveal three key findings: 1) LLMs' perceptions of knowledge boundaries are encoded in the middle to middle-upper layers across different languages. 2) Language differences in knowledge boundary perception follow a linear structure, which motivates our proposal of a training-free alignment method that effectively transfers knowledge boundary perception ability across languages, thereby helping reduce hallucination risk in low-resource languages; 3) Fine-tuning on bilingual question pair translation further enhances LLMs' recognition of knowledge boundaries across languages. Given the absence of standard testbeds for cross-lingual knowledge boundary analysis, we construct a multilingual evaluation suite comprising three representative types of knowledge boundary data. Our code and datasets are publicly available at https://github.com/DAMO-NLP-SG/LLM-Multilingual-Knowledge-Boundaries.

1 Introduction

The rapid advancement of large language models (LLMs) has revolutionized their capacity to store and utilize knowledge across languages (Dubey et al., 2024; Qwen et al., 2025). Understanding the knowledge boundaries of Large Language Models (LLMs) is critical, as LLMs tend to hallucinate when attempting to answer questions beyond their knowledge. Yet, research on knowledge boundaries of LLMs has predominantly focused on English (Azaria and Mitchell, 2023; Marks and Tegmark, 2024; Li et al., 2024; Cheang et al., 2023). Misaligned knowledge boundaries between languages can lead to inconsistent and unsafe outputs in cross-lingual applications. Therefore, it is crucial to determine whether the knowledge boundary perceptions of LLMs observed in English can be similarly identified in other languages or transferred across them.

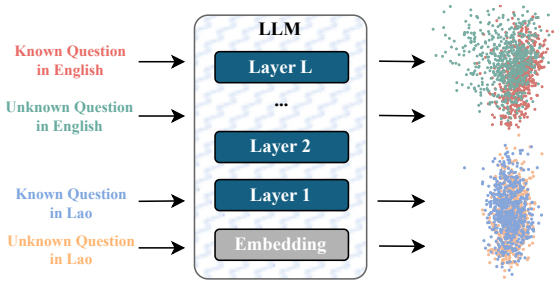

Figure 1: Our goal is to analyze LLMs' cognition of knowledge boundaries across different languages by inspecting their representations of knowledge boundaries in multiple languages. The right-hand side of the figure illustrates the representations when an LLM encodes known and unknown questions in English and Lao.

To fill this gap, we are pioneering the investigation into *how LLMs perceive and encode knowledge boundaries across languages*, as illustrated in Figure 1. Through probing the representations of LLMs, our work reveals novel structural geometry, training-free transfer methods, and training methods to jointly enhance cross-lingual awareness.

We begin by applying layer-wise probing on parallel multilingual questions with true/false premises to examine **how LLMs perceive knowledge boundary across languages and layers**. Initial experiments (§4) show that 1) the cognition of knowledge boundaries is encoded in the middle to mid-upper layers of LLMs. 2) The representation subspaces of different languages converge into a uni-

[†]Work done during internship at Alibaba DAMO Academy.
[‡]Corresponding authors.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 24099–24115
July 27 - August 1, 2025

fied knowledge space, in which probes exhibit the best inter-language transferability. 3) Low-resource language representations provide high zero-shot transferability to high-resource language representations, but not vice versa.

Motivated by the above observations, we further explore **whether a specific structure exists in the geometry of multilingual knowledge boundary representations, allowing training-free alignment methods to transfer abilities between languages**. Experiments (§5) show that language difference for knowledge boundary is encoded in a linear structure, and training-free alignment methods, such as mean-shifting and linear projection, can effectively transfer the knowledge boundary perception across languages. Notably, linear projection largely closes the performance gap between in-distribution (ID) and out-of-distribution (OOD) language representations, demonstrating improved cross-lingual transferability. We also find that projecting high-resource language representations onto low-resource subspaces removes extraneous noise encoded by the inherent high dimensionality of high-resource languages. As a result, compared to the low-resource language representation themselves, probes trained on low-resource languages achieve better performance on these projected high-resource language representations than on low-resource language representation themselves, revealing a "weak-to-strong generalization" pattern.

Building on the success of training-free alignment and the distinctive "*weak-to-strong generalization*" pattern, we are intrigued by **if fine-tuning on certain languages can further refine their knowledge boundary perception ability, and generalize this improvement to other languages?** As prior work (Zhang et al., 2024b) demonstrates that fine-tuning LLMs solely on question translation data can improve their cross-lingual performance in various downstream tasks like sentiment analysis, we investigate whether such fine-tuning can enhance LLMs' cognition of knowledge boundaries across different languages. Our experiments show that SFT using only question translation data effectively improves LLMs' perception of their knowledge boundaries across languages. We further observe a latent safeguard mechanism where the primary language representation is enhanced when the model is fine-tuned on non-dominant language pairs consisting of unanswerable questions.

To address the current lack of standard testbeds for evaluating the generalization of knowledge boundary cognition across languages, we constructed three types of multilingual knowledge boundary datasets, including: 1) questions with true/false premises; 2) entity-centric answerable/unanswerable questions; and 3) true/false statements. For true/false-premise questions, we first contribute an augmented version of FRESHQA (Vu et al., 2024) by flipping the premise of each question, and creating its corresponding multilingual version. For entity-centric answerable/unanswerable questions, we create a multilingual QA dataset using an entity-swapping framework, where the content is written in authentic language for each target language, based on a diverse array of real and fictional entities. We also translate the widely used true/false statement dataset in (Azaria and Mitchell, 2023) into 8 languages.

To our knowledge, we are the first to comprehensively study the generalization of LLM knowledge boundary awareness across languages. In summary, our contributions are three-fold: 1) We present the first comprehensive analysis of how LLMs' cognition of knowledge boundaries transfers across languages. 2) We develop a multilingual evaluation suite comprising three representative types of knowledge boundary data, providing valuable resources for both evaluating and enhancing LLMs' perception of knowledge boundaries across languages. 3) Our training-free transfer methods significantly reduce the performance gap in probing across languages, demonstrating effectiveness as an additional signal in LLM generation that substantially reduces the risk of hallucination.

2 Related Work

2.1 Knowledge Boundaries of LLMs

While knowledge boundaries are crucial for mitigating hallucinations and unsafe generation, the concept remains underdefined (Li et al., 2024). From a generative perspective, Yin et al. (2024) classify LLM's knowledge into prompt-agnostic knowledge, prompt-sensitive knowledge, and unknown knowledge. Complementarily, representations-based studies focus on whether LLMs' knowledge boundary perception is reflected in their internal states, through probing LLM's representations with true/false statements and analyzing their geometric patterns (Azaria and Mitchell, 2023; Marks and Tegmark, 2024; Bürger et al., 2024). However, most work in this line is limited in true/false statements, and no work has systematically investigated

Dataset	Languages	# train samples	# test samples
FRESHQAPARALLEL	en, zh, vi, th, id, ms, km, lo	-	9,600 (1,200 per lang.)
SEAREFUSE	en, zh, id, th, vi	64,075 (∼10-15k per lang.)	6,000 (1,200 per lang.)
TRUEFALSEMULTILANG	en, es, de, it, pt, fr, id, th	-	48,680 (6,085 per lang.)

Table 1: Statistics of the three datasets in our constructed evaluation suite.

multilingual knowledge boundaries, except Bürger et al. (2024) briefly showed that probes trained on English show certain generalization to German statements. In this work, we show that most inter-language generalization gaps can be closed with training-free subspace alignment.

2.2 Multilingualism in LLMs

Multilingual large language models increasingly support diverse languages through larger-scale training (Qwen et al., 2025; Dubey et al., 2024) and targeted efforts towards multilinguality (BigScience Workshop, 2023; Üstün et al., 2024; Zhang et al., 2024c). Studies identify language-specific neurons in multilingual large language models and steer LLMs' behaviors by perturbing such neurons (Tang et al., 2024; Zhao et al., 2024), revealing sparse language-specific neurons and "anchor language" processing in middle layers (Zhao et al., 2024). Mu et al. (2024) link parallel multilingual inputs during inference to inhibit neurons and precise neuron activation. Cross-lingual improvement patterns emerge during fine-tuning: Zhang et al. (2024b) find that question translation fine-tuning boosts multilingual performance, even for languages that are not directly fine-tuned on. In this work, we provide more concrete evidence to this method from a representation perspective.

3 Datasets

We construct a multilingual evaluation suite to analyze how LLMs generalize their knowledge boundary cognition across languages, comprising three datasets: FRESHQAPARALLEL, SEAREFUSE, and TRUEFALSEMULTILANG. These cover three types of knowledge boundary data: questions with true or false premises, entity-centric answerable/unanswerable questions, and true/false statements. Table 1 summarizes the dataset statistics.

FRESHQAPARALLEL Our FRESHQAPARALLEL dataset extends FRESHQA (Vu et al., 2024) by creating parallel true/false premise questions. Human annotators are asked to manually invert the premise type of each question, *e.g.,* converting "When did Google release ChatGPT?" (false-premise) to "When did OpenAI release ChatGPT?" (true-premise). We highlight the difficulty of converting true-premise questions to false-premise due to larger search space, and provide more details of the annotation in Appendix E. We further apply GPT-4o to translate the dataset into 7 languages, which are then quality-checked by linguists in the team.

SEAREFUSE We propose the SEAREFUSE Benchmark, which is composed of unanswerable questions about non-existing entities and answerable questions in English, Chinese, Indonesian, Thai, and Vietnamese. We devise an entity-swapping approach and a generation-based approach to construct questions about fake entities. The entity swapping approach constructs unanswerable questions by swapping the named entities in questions from open-source QA datasets into non-existent entities. The questions form SEAREFUSE-H training/test set. Each question in the test set is verified by linguists. The generation-based approach relies on GPT-4o to synthesize unanswerable questions, resulting in the SEAREFUSE-G test set. More details are in Appendix F.

TRUEFALSEMULTILANG We apply GPT-4o to translate the TRUEFALSE dataset (Azaria and Mitchell, 2023), a dataset of true and false statements on six different topics, into 7 languages (es, de, it, pt, fr, id, th) that form the intersection that SOTA LLMs (Qwen et al., 2025; Dubey et al., 2024) claim to support. The quality of the translation is verified by linguists.

4 Knowledge Boundary Probing

We first systematically analyze our primary question: *can we probe the representations of language models to detect knowledge boundaries, and what patterns emerge across languages and layers?*

Experiment Settings. We utilize `Qwen2.5-7B` and `Llama-3.1-8B` in our main experiments and further study the scaling with `Qwen2.5-14B` and

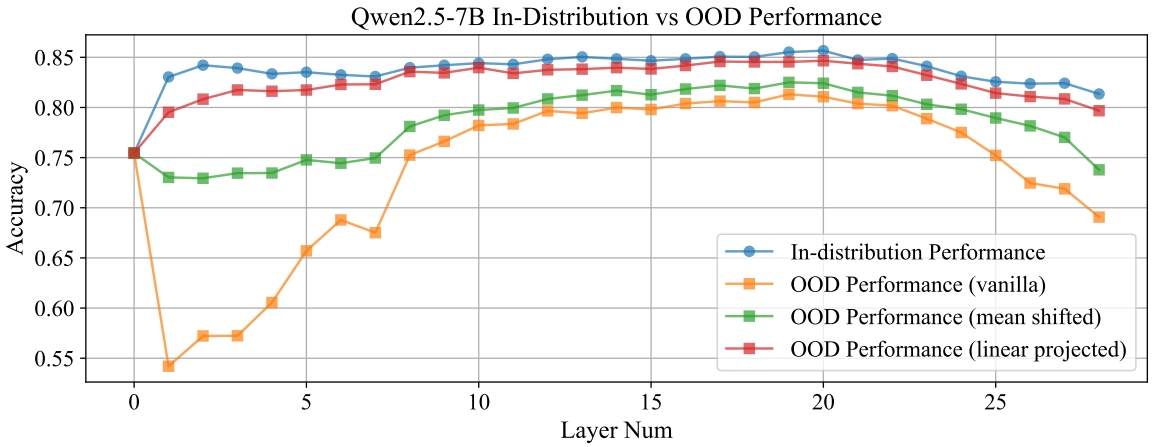

Figure 2: In-distribution and OOD performance of layer-wise probes trained on `Qwen2.5-7b` representations, across different methods. All scores are averaged across all languages.

`Qwen2.5-72B`. For each model, we train $k \times m$ in-distribution linear classifiers $f : \mathbb{R}^d \to \mathcal{C}$, using the last token representations of questions $\mathbf{E} \in \mathbb{R}^{n \times d}$ for each language, where k is the number of layers, m is the number of languages, n is the number of questions and d the embedding dimensionality. Each layer- and language-specific classifier (trained on in-distribution data) is evaluated zero-shot on all other languages. We use **accuracy** as the main metric for evaluating probing models. We average per-language probe performance in each layer (*e.g.,* an 8-language setup yields 64 scores per layer).

RQ1.1: ***Do probes transfer cross-lingually, and what's the pattern across layers, languages, and models?***

Layers. Figure 3 visualizes the transferability matrices of probing model on `Qwen2.5-7B`'s 3rd and 19th layers, which exhibit the worst and best average transferability across languages, respectively. Figure 2 reveals a significant gap between in-distribution (blue line) and OOD (orange line) performance, which is most pronounced in the bottom layers. In these layers, language models begin incorporating language-specific static embeddings and processing within corresponding language subspaces before converging to a unified knowledge representation space shared across languages in the middle and mid-upper layers.

Languages. High and low-resource languages exhibit distinct performance and transferability patterns (Figure 3). 1) *High-resource languages achieve superior ID accuracy but transfer poorly to low-resource languages.* For instance, all languages provide the lowest transferability to Khmer, *e.g.,* English→Khmer drops from 89% to 79% even in the most transferable layer. 2) *Mid-resource languages offer modest transferability to both high- and similar-resource languages*, such as Thai→Chinese and Thai→Indonesian in the 19th layer. 3) *Low-resource languages, like Khmer, show the best relative transferability*, with consistent performance across all layers and languages. These patterns suggest that discriminative features for low-resource languages exist within high-resource representations but not vice versa. More evidences are provided in the *weak-to-strong generalization* section of §5.3.

Models. The above findings hold across model families (`Llama`, `Qwen`), sizes (7B to 72B), and variants (base/instruct). We refer to the Appendix for full results across models. Representations show subpar cross-lingual transferability in the bottom layers. The transferability peaks in middle (`Llama`) and mid-upper layers (`Qwen`), then diverges again in final layers.

Our results offer more concrete and direct evidence for the 3-phase hypothesis in Zhao et al. (2024), where language models first process input in the source language, then "think" in English or other anchor languages (such as English and Chinese for `Qwen`), and finally switch back to the source language for generation.

RQ1.2: ***Do language models fail to express their knowledge boundary multilingually, even when the boundary is stored in their representations?*** The promising probing performance demonstrates that LLMs internally encode knowledge boundary awareness, such as identifying false-premise questions. But have they fully utilized the aware-

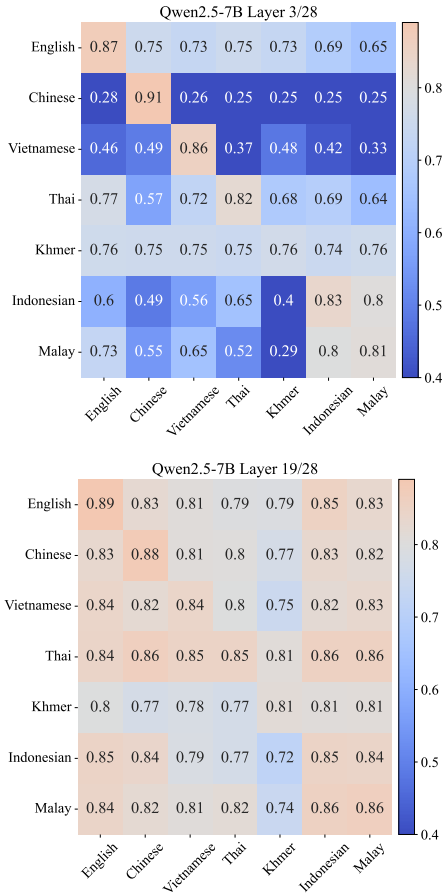

Figure 3: The performance of probing models on the 3rd and 19th layers of `Qwen2.5-7B` with different training and testing language combinations. The y-axis and x-axis represent the languages used for training and inference of the probing models, respectively.

ness in generation? We compare models' performance on FreshQA's false-premise subset with and without hints suggesting the question may be false-premise. As shown in Table 2, providing premise hints significantly improves the performance of probes across languages, with the largest gain of 24.49% for `Qwen2.5-7B-Instruct` on Vietnamese and 17.69% for `Qwen2.5-72B-Instruct` on Malay. These results indicate that while models possess discriminative representations of knowledge boundaries, they may underutilize the information during generation. This highlights the value of training multilingual knowledge boundary probes in practical applications, like detecting false-premise on the fly and prompting re-generation with corrected hints.

5 Training-free Transfer

Given the distinct patterns across languages, a natural question is *whether knowledge boundary perception capabilities can be transferred between languages?* In this section, we investigate the geometric structure in the knowledge boundary representations and explore training-free alignment methods to exploit such structures.

	en	zh	vi	th	km	id	ms	lo
Qwen2.5-7B-Instruct								
Baseline	30.61	36.05	19.73	19.73	8.16	22.45	19.05	0.68
FP-Hinted	41.50	45.58	44.22	32.65	11.56	38.10	37.41	2.04
Qwen2.5-72B-Instruct								
Baseline	58.50	60.54	61.90	55.10	33.33	59.18	55.78	31.29
FP-Hinted	72.11	70.75	68.03	67.35	44.90	72.79	73.47	38.10

Table 2: Performance comparison between models without and with hints suggesting the question may be false-premise (Baseline vs. FP-Hinted).

5.1 Subspace Geometry

RQ2.1: ***Does a linear structure exist in the geometry of knowledge boundary representations across languages?***

We investigate whether an LLM's perception towards knowledge boundaries is encoded in language-neutral ways. Using embeddings from layers that yield the best cross-language transferability (*e.g.,* the 19th layer of `Qwen2.5-7B`), three LDA classifiers are trained on encodings from the concatenated TRUEFALSEMULTILANG dataset across all languages. The classifiers are trained using three different label sets: 1) language labels, 2) the Cartesian product of the domain set and truth/falseness, and 3) binary truth/falseness labels.

Figure 4 visualizes embeddings projected onto axes taken from the three LDA subspaces. Annotated with the three label sets, the results reveal: 1) the model represents questions in a language-neutral way, presenting a parallel structure in the projected axes (Figure 4, left); 2) truth/falsity is encoded within the model (Figure 4, middle); and 3) true/false statements can be separated in a topic-agnostic way by a near-horizontal hyperplane (Figure 4, right). Similar patterns emerge for entity-aware binary answerability in the SEAREFUSE dataset (Appendix A, Figure 7).

5.2 Subspace Projections

RQ2.2: ***Can we transfer the abilities to perceive knowledge boundaries between languages by linearly aligning their representation space without training?***

Motivated by the language neutrality in representing true/falsity, we further explore the effectiveness of training-free embedding subspace alignment methods, *i.e.,* **Mean Shifting** and **Linear**

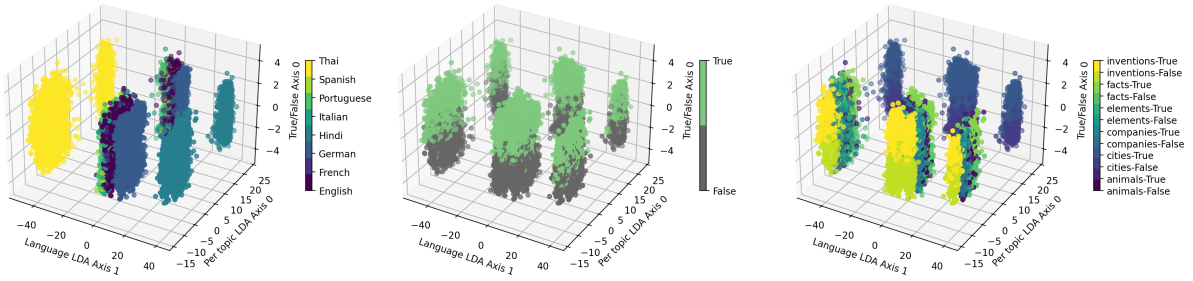

Figure 4: We project the 19th layer of `Qwen2.5-7B` onto a linear subspace where the x-axis encodes languages, the y-axis encodes topic-aware true/falsity, and the z-axis encodes binary truth/falsity. We visualize the same scatter plot with three different ground-truth label sets annotated respectively.

Projection, to exploit linear structures and align language subspaces.

Mean Shifting. We denote the in-distribution (source language) and out-of-distribution (target language) training sets as $\mathbf{X}_{\text{in}}^{\text{train}} \in \mathbb{R}^{n \times d}$ and $\mathbf{X}_{\text{ood}}{}^{\text{train}} \in \mathbb{R}^{n \times d}$, respectively, where n is the number of parallel samples and d is the feature space dimensionality. Prior work (Chang et al., 2022; Xu et al., 2023) reveals that language differences are mainly encoded by language subspace representation means in multilingual encoder models. We assess whether mean-shifting can serve as an effective baseline for LLMs. This method aims to adjust target language embeddings so their mean aligns with the source language embeddings. The mean features for both distributions can be computed as: $\boldsymbol{\mu}_{\text{in}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_{\text{in},i}$, $\boldsymbol{\mu}_{\text{ood}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_{\text{ood},i}$, where $\mathbf{x}_{\text{in},i}$ and $\mathbf{x}_{\text{ood},i}$ are the embeddings of i-th sample in $\mathbf{X}_{\text{in}}^{\text{train}}$ and $\mathbf{X}_{\text{ood}}^{\text{train}}$, respectively, from which the difference between the two mean vectors can be attained: $\Delta\boldsymbol{\mu} = \boldsymbol{\mu}_{\text{in}} - \boldsymbol{\mu}_{\text{ood}}$.

This difference vector $\Delta\boldsymbol{\mu}$ represents the translation needed to align the mean of the OOD subspace with that of the in-distribution subspace, which is then applied to the OOD test set $\mathbf{X}_{\text{ood}}^{\text{test}} \in \mathbb{R}^{p \times d}$ to obtain the shifted test set $\mathbf{X}_{\text{shifted}}^{\text{test}}$: $\mathbf{X}_{\text{shifted}}^{\text{test}} = \mathbf{X}_{\text{ood}}^{\text{test}} + \Delta\boldsymbol{\mu}$. Appendix I shows that language means can also be approximated with non-parallel corpus.

Linear Projection Using the same notation of $\mathbf{X}_{\text{in}}^{\text{train}} \in \mathbb{R}^{n \times d}$ and $\mathbf{X}_{\text{ood}}^{\text{train}} \in \mathbb{R}^{n \times d}$, the linear transformation $\mathbf{W} \in \mathbb{R}^{d \times d}$ that projects vectors from the target language subspace to the source language subspace is estimated by solving the least squares problem: $\mathbf{W} = \arg\min_{\mathbf{W}} \|\mathbf{X}_{\text{in}} - \mathbf{X}_{\text{ood}}^{\text{train}}\mathbf{W}\|_F^2$, where $\|\cdot\|_F$ denotes the Frobenius norm. The solution for $\mathbf{W}$ is computed using the Moore-Penrose pseudo-inverse: $\mathbf{W} = \mathbf{X}_{\text{ood}}^{\text{train}+}\mathbf{X}_{\text{in}}$, where $\mathbf{X}_{\text{ood}}^{\text{train}+} = \mathbf{V}\boldsymbol{\Sigma}^{+}\mathbf{U}^{\top}$ is derived from the Singular Value Decomposition (SVD) $\mathbf{X}_{\text{ood}}^{\text{train}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\top}$. Here, $\boldsymbol{\Sigma}^{+}$ replaces non-zero singular values σ_i with $1/\sigma_i$. For full-rank $\mathbf{X}_{\text{ood}}^{\text{train}}$, it reduces to $\mathbf{W} = (\mathbf{X}_{\text{ood}}^{\text{train}\top}\mathbf{X}_{\text{ood}}^{\text{train}})^{-1}\mathbf{X}_{\text{ood}}^{\text{train}\top}\mathbf{X}_{\text{in}}$. Given that our representation matrices mostly don't have full column rank (*e.g.,* the number of questions in FreshQA is far smaller than embedding dimensionality), SVD is used to preserve numerical stability.

Once $\mathbf{W}$ is estimated, it can be used to project the out-of-distribution test set $\mathbf{X}_{\text{ood}}^{\text{test}} \in \mathbb{R}^{p \times d}$ into the in-distribution subspace, resulting in the shifted test set $\mathbf{X}_{\text{shifted}}^{\text{test}}$: $\mathbf{X}_{\text{shifted}}^{\text{test}} = \mathbf{X}_{\text{ood}}^{\text{test}}\mathbf{W}$. For scalability, inter-language projection matrices can be precomputed by sampling $\mathbf{X}_{\text{ood}}$ and $\mathbf{X}_{\text{in}}$ from larger parallel corpora $\mathbb{C}_{\text{ood}}$ and $\mathbb{C}_{\text{in}}$.

5.3 Findings

Mean Shifting largely improves transferability. As shown in Figure 2, mean shifting (green line) substantially enhances transferability across layers compared to vanilla representations, indicating that language difference is largely encoded in subspace mean. However, the remaining gap compared to in-distribution performance suggests that additional factors, such as orientation and dimensional magnitudes (*i.e.,* anisotropic stretching in each dimension), also affect the subspace differences.

Linear Projection nearly matches in-distribution performance. In Figure 2, linear projection (red line) nearly closes the gap with in-distribution accuracy. The linear transformation, learned from language-pair training sets, robustly captures geometric relationships rather than overfitting to example-specific features.

The gap between mean-shifted and linearly-projected representations highlights the additional complex mapping needed to align language subspaces. Meanwhile, the promising cross-lingual

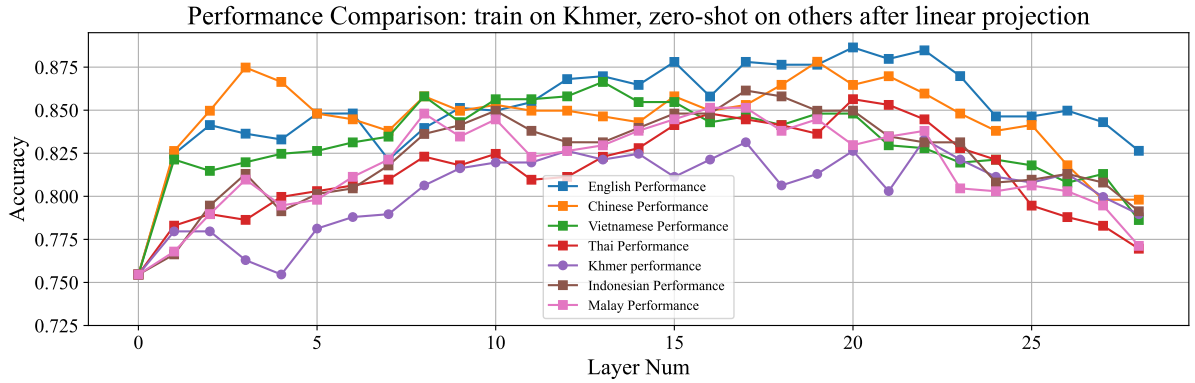

Figure 5: Performance of layer-wise probes trained on Khmer for different languages post-projection. We use linear projection to project each language's representations onto the subspace of Khmer. It is observed that Khmer probes perform better on OOD languages than on the Khmer language itself.

transferability of probes on projected representations highlights their applicability for generation. In practice, one can pretrain probes on all languages and precompute linear projection for every language pair. In inference time, a lightweight language detector can be applied to identify the inferred language, and use the linear projection $\mathbf{W}_{\text{j}\to\text{k}}$ to transform language j's representations into language k's subspace (the language that transfers best to j) to use the optimal probes.

Using `Qwen2.5-14B` as an example, the best performance for Thai (88.31%) and Malay (88.64%) is achieved by applying the 30$^{\text{th}}$-layer Chinese probe after linearly projecting onto the Chinese subspace, rather than using their probes. Similarly, Khmer (88.15%) and Lao (86.64%) perform best with the 29$^{\text{th}}$-layer Malay and 32$^{\text{nd}}$-layer Thai probes, respectively. This high→mid, mid→low transferability trend showcases the relative dominance of language representations due to levels of language resources in the pretraining of LLMs.

A weak-to-strong generalization pattern. We observe an intriguing pattern: probes trained on low-resource languages (*e.g.,* Khmer, Lao) perform better on other languages after post-projection than on their own (*e.g.,* the Khmer example in Figure 5).

We hypothesize that projecting high-resource language representations onto low-resource subspaces acts as a denoising/regularization step, removing language-specific nuances such as syntactic specifics. Table 3 provides empirical evidence using FRESHQAPARALLEL. By dividing the examples into 80%-20% train-test splits and ensuring each original question with its flipped-premise counterpart in the same split, we compute the SVD of the centered test set representations $\mathbf{X} \in \mathbb{R}^{n\times d}$, yielding singular values $\{\sigma_i\}_{i=1}^{n}$, and approximate effective dimensionality by $\min\left\{k \in \mathbb{N} \mid \sum_{i=1}^{k}\sigma_i^2 \geq 0.95 \cdot \sum_{i=1}^{n}\sigma_i^2\right\}$, reflecting the compactness of task-relevant information in the representation space.[1] We also compute the participation ratio (PR) as $\frac{\left(\sum_{i=1}^{n}\sigma_i^2\right)^2}{\sum_{i=1}^{n}\sigma_i^4}$.[2] Next, we project languages onto other subspaces via $\mathbf{W}$, computed using $\mathbf{X}_{\text{lang1}}^{\text{train}} \in \mathbb{R}^{4n\times d}$ and $\mathbf{X}_{\text{lang2}}^{\text{train}} \in \mathbb{R}^{4n\times d}$, and recalculate these metrics.

	Effective Dimensionality		Participation Ratio	
	Original	Projected	Original	Projected
Khmer	103	97	15.93	18.07
English	116	87	26.26	19.29

Table 3: Effective Dimensionality and Participation Ratios: original subspace and after English and Khmer projected onto each other's subspace.

Table 3 compares the results in English (high-resource) and Khmer (low-resource). English's higher intrinsic effective dimensionality suggests it encodes more features, potentially language-specific and task-irrelevant details. In contrast, Khmer's subspace is more compact and task-focused, likely due to limited training data that forces LLMs to prioritize generalizable features during pretraining. However, projecting English onto Khmer's subspace reduces its effective dimensionality (116 → 87), stripping away noise or redundant features. English's higher PR indicates diffuse variance across many directions, whereas

[1]The effective dimensionality is approximated by the minimum number of principal components required to explain 95% of the total variance

[2]Participation ratio (PR) denotes the effective number of dimensions contributing significantly to the total variance.

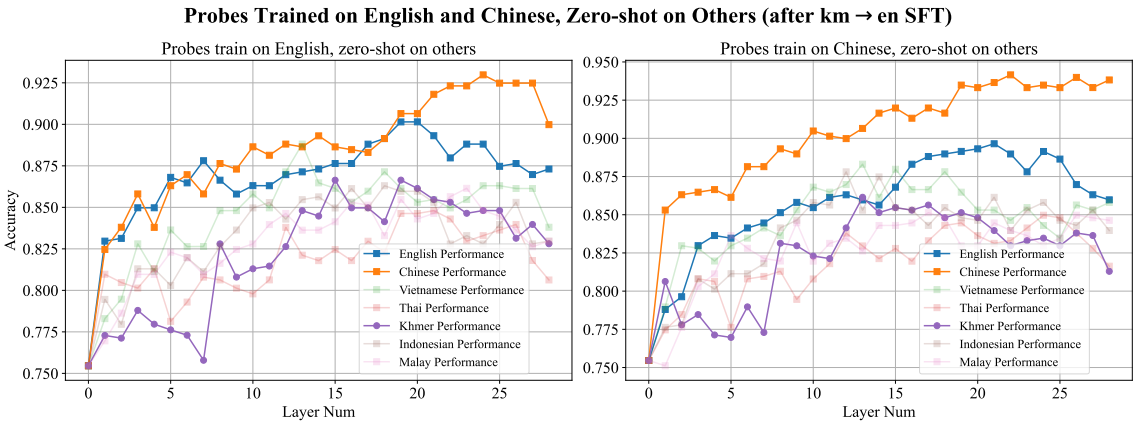

Figure 6: After SFT'ed on Khmer-English translation pairs, `Qwen2.5-7B` exhibits unexpected enhancement in Chinese representations, which otherwise would achieve ~ 0.90 or lower accuracy across probes.

Khmer's lower PR reflects tighter feature alignment with the tasks. English → Khmer subspace reduces PR (26.26 → 19.29), aligning closer to Khmer's compactness. Khmer → English subspace increases PR (15.93 → 18.07), indicating a loss of focus. Thus, low-resource subspaces act as inductive biases for cross-lingual tasks by filtering noise and preserving core semantics, while high-resource subspaces risk carrying superfluous details that degrade transfer.

6 Training-based Enhancement

Having shown the promising performance of aligning representations across language subspaces and the weak-to-strong generalization in language pairs, we further investigate *whether training on specific languages can enhance performance that generalizes to others.* Prior research (Zhang et al., 2024b) shows that fine-tuning question translation data, even without ground-truth answers, enhances LLM performance across languages in various downstream tasks. Inspired by it, we further ask:

RQ3: Can finetuning on question translation data enhance the knowledge boundary cognition ability of LLMs across languages?

Experiment Settings. We fine-tune `Llama-3.1` and `Qwen2.5` (from 7B to 72B) via SFT on question translation pairs (*e.g.,* English-Lao), which are constructed using the English subset of SEAREFUSE, translated by GPT-4o and verified by linguists. We measure the performance post-SFT using the same protocols in §4 and §5.

We conduct experiments on all model sizes with both Full-parameter fine-tuning and LoRA. For LoRA, we use a rank 32 and alpha 64. We use a learning rate of $1e-5$ for full fine-tuning and $1e-4$ for LoRA. A global batch size of 512 is maintained for all settings through different gradient accumulation setups. Qwen2.5-72B is trained on 4 nodes of 32 A800 80G GPUs in total with DeepSpeed ZeRO-3 for full-parameter training.

SFT on the bilingual translation pairs consistently enhances cross-lingual knowledge boundary cognition. Probes trained on `Qwen2.5-7B` and `Qwen2.5-72B` show consistent gains across languages after Khmer→English translation fine-tuning (see Figures 8, 9 in Appendix C), with `Qwen2.5-7B` achieving a $+2.3\%$ average gain in the best layer, yielding a 88% average accuracy in classifying true/false premises. This improvement holds across language families and bilingual pair selections: fine-tuning `Llama-3.1-8B` on Khmer→Thai or Lao→English pairs enhances upper-layer performance by $> 10\%$ (Figures 11, 12), indicating the upper-layer "expression spaces" across languages has been better aligned.

A *self-defense* mechanism. Fine-tuning on low-resource→English translation pairs unexpectedly enhances the model's strongest inherent language. For instance, `Qwen2.5` fine-tuned on Khmer-English pairs exhibits emergent alignment favoring Chinese representations (Figure 6), significantly improving performance with probes trained in other languages (full results in Figure 14). We hypothesize that processing unanswerable questions in low-resource languages (*e.g.,* Khmer) activates latent safety mechanisms linked to the model's primary languages, akin to a "self-defense" reflex where boundary awareness in weaker languages reinforces safeguards in stronger ones.

7 Conclusion

We present the first comprehensive study on how LLMs perceive knowledge boundaries across languages through a systematic analysis of their internal representations. Our findings reveal that knowledge boundary cognition is encoded in middle to middle-upper layers across languages, following a linear structural pattern that enables effective cross-lingual transfer. Motivated by this, we propose a training-free alignment method and targeted fine-tuning with question translation data to transfer knowledge boundary perception ability across languages. Our multilingual evaluation suite provides a valuable resource for future research. These insights not only advance our understanding of LLMs' cross-lingual knowledge boundary perception but also offer practical strategies for improving transfer to low-resource languages.

Limitations

While we extensively study probing representations of LLMs' perception of questions, it is of interest to investigate how these representations evolve in the generation process. For instance, given the paradigm shifts to explicit long chain-of-thoughts in the generation process, it is possible that LLMs reach the "aha moment" in the reasoning process when they become aware that the question reaches its knowledge boundaries, where this representation becomes more linearly separable to the probes that classify this perception. However, we note that it is very difficult to locate the information when the transformation takes place and extract the corresponding representations. We leave this interesting line of exploration to future work.

References

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

BigScience Workshop. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Lennart Bürger, Fred A. Hamprecht, and Boaz Nadler. 2024. Truth is universal: Robust detection of lies in llms. *Preprint*, arXiv:2407.12831.

Tyler Chang, Zhuowen Tu, and Benjamin Bergen. 2022. The geometry of multilingual language model representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 119–136, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chi Seng Cheang, Hou Pong Chan, Derek F. Wong, Xuebo Liu, Zhaocong Li, Yanming Sun, Shudong Liu, and Lidia S. Chao. 2023. Can lms generalize to future data? an empirical analysis on text summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 16205–16217. Association for Computational Linguistics.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Nan Duan. 2016. Overview of the nlpcc-iccpol 2016 shared task: Open domain chinese question answering. In *Natural Language Understanding and Intelligent Applications*, pages 942–948. Springer International Publishing.

Nan Duan and Duyu Tang. 2018. Overview of the nlpcc 2017 shared task: Open domain chinese question answering. In *Natural Language Processing and Chinese Computing*, pages 954–961. Springer International Publishing.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and

et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard H. Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Trans. Assoc. Comput. Linguistics*, 9:1012–1031.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018. Tempquestions: A benchmark for temporal question answering. In *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, pages 1057–1062. ACM.

Moxin Li, Yong Zhao, Yang Deng, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See-Kiong Ng, and Tat-Seng Chua. 2024. Knowledge boundary of large language models: A survey. *Preprint*, arXiv:2412.12472.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *Preprint*, arXiv:2310.06824.

Yongyu Mu, Peinan Feng, Zhiquan Cao, Yuzhang Wu, Bei Li, Chenglong Wang, Tong Xiao, Kai Song, Tongran Liu, Chunliang Zhang, and JingBo Zhu. 2024. Revealing the parallel multilingual learning within large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6976–6997, Miami, Florida, USA. Association for Computational Linguistics.

Kiet Van Nguyen, Duc-Vu Nguyen, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2020. A vietnamese dataset for evaluating machine reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 2595–2605. International Committee on Computational Linguistics.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 784–789. Association for Computational Linguistics.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720, Bangkok, Thailand. Association for Computational Linguistics.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 2020. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857, Suzhou, China. Association for Computational Linguistics.

Shaoyang Xu, Junzhuo Li, and Deyi Xiong. 2023. Language representation projection: Can we transfer factual knowledge across languages in multilingual language models? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3692–3702, Singapore. Association for Computational Linguistics.

Xunjian Yin, Xu Zhang, Jie Ruan, and Xiaojun Wan. 2024. Benchmarking knowledge boundary for large language models: A different perspective on model evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2270–2286, Bangkok, Thailand. Association for Computational Linguistics.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024a. R-tuning: Instructing large language models to say 'I don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies (Volume 1: Long Papers), pages 7113–7139, Mexico City, Mexico. Association for Computational Linguistics.

Shimao Zhang, Changjiang Gao, Wenhao Zhu, Jiajun Chen, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2024b. Getting more from less: Large language models are good spontaneous multilingual learners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8037–8051, Miami, Florida, USA. Association for Computational Linguistics.

Wenxuan Zhang, Hou Pong Chan, Yiran Zhao, Mahani Aljunied, Jianyu Wang, Chaoqun Liu, Yue Deng, Zhiqiang Hu, Weiwen Xu, Yew Ken Chia, Xin Li, and Lidong Bing. 2024c. Seallms 3: Open foundation and chat multilingual large language models for southeast asian languages. *CoRR*, abs/2407.19672.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? In *Advances in Neural Information Processing Systems*, volume 37, pages 15296–15319. Curran Associates, Inc.

A Language-neutrality of Entity-aware binary Answerability

Having shown the linear structure of true/false statement representations across languages, we visualize the SEAREFUSE benchmark dataset we construct in this paper here. We concatenate the training set of SEAREFUSE of all 5 languages, yielding a representation set of over 120k for training LDA classifiers.

Due to the lack of topic and domain information as in the true/false statement datasets, we train 2 LDA classifiers with the concatenated SEAREFUSE representations of all languages, with 1) language labels. 2) binary existence labels of entities (exist/non-exist). In Figure 7, we project the representations onto a subspace where the horizontal axes (x,y) are from the language LDA subspace, and the vertical axis (z) is from the exist/non-existing (thus answerable/unanswerable) subspace.

B Transferability across Layers Full Results

We include full results of transferability across layers brought by our training-free methods, across model families and sizes. *e.g.,* Figure 10 presents results of Llama-3.1-8B.

C Transferability Pattern Change after SFT'ed on Question Pairs

Figure 11, and 12 present Llama-3.1-8B performance on FreshQA, after the model has been fine-tuned on Lao-English, Khmer-Thai SeaRefuse question translation pairs without answerability indications, respectively. As seen, the last layers show a drastic improvement over Figure 10 (before training), showing an internal alignment brought to across languages, beyond the bilingual pairs involved in training.

D Pattern 2 Full Visualization

Figure 14 visualizes the self-defense pattern of Qwen2.5 across all language probes.

E Details of Data Construction of FRESHQAPARALLEL

Although flipping a false-premise question to true-premise could be simpler by referring to search engines or reasoning based on the groundtruth answers about the original questions in FreshQA, flipping a true-premise question to false-premise is more difficult, given the larger search space (many ways to make it false-premise). We note that, the flipped question must be talking about the same topic with the original question; without getting too creative (e.g., GPT-4o would frequently fail to do this task as it generates too fictional entities such as transforming QS World Ranking to Hogwarts Ranking); and not grounded in the futuristic property relative to model's training cutoff time (e.g., Who is the US president in 2024 is not deemed a false-premise question to a model trained in 2023, although it's unanswerable without retrieval-augmented generation; However "Who was the US president in 2035" is false-premise, which assumes a future date is in the past).

Below are some more difficult examples we communicate to the human annotators and our linguist team:

Example 1.a. "How many children does Leonardo DiCaprio have?" Questions like these are not false-premise questions. Even though Leonardo DiCaprio doesn't have children, this question doesn't assume that he has, as the answer to this question can be 0, and thus answerable.

Example 1.b. "How old is Leonardo DiCaprio's second kid?" This is a valid false-premise question because it has assumed a false fact (that the entity has 2 or more kids).

Similarly, Example 2.a. "How many championships has Chris Paul won?" (not a valid false-premise question because it doesn't assume anything; the answer can be 0). Example 2.b. "When did Chris Paul win the second championship?" (a valid false-premise question because it has assumed a false fact.)

For question (id 124) How many humans have landed on Mars? This is not a false-premise question because it's answerable. Questions like "How many humans have landed on Venus?" are not valid false-premise either. In this case, we can write something like "How many humans have landed on Mars as part of the first manned mission in 2020?" that meets the requirements (making the question not answerable because non-factual).

F Details of Data Construction of SEAREFUSE

We first prompt GPT-4o to generate fake entity candidates of the types of person, location, organization, event, and work of art, respectively. The entity swapping approach first applies the spacy[3] NER tool to detect named entities from human-written answerable factoid questions. Then, we randomly select a named entity from each source question and then swap the selected entity with a random entity from the non-existing entity candidates of the same type. As the detected named entities may not be accurate, we further ask linguists to verify the quality of the constructed unanswerable questions and rectify the disfluent questions, resulting in the SEAREFUSE-H test set. This set contains 100 answerable and 100 unanswerable questions for each language. The generation-based approach first prompts GPT-4o to write a 120-word story for a randomly selected fake entity candidate. Then we prompt GPT-4o to write four factoid questions about the selected fake entity based on the generated story, resulting in the SeaRefuse-G test set. This set includes 500 answerable and 500 unanswerable questions for each language, except for Vietnamese, which has 483 of each.

The answerable questions are collected from open source factoid QA training sets, including ParaRel (Zhang et al., 2024a; Elazar et al., 2021), SQUAD (Rajpurkar et al., 2018), Web Questions QA[4], TempQ (Jia et al., 2018), NLPCC-KBQA (Duan, 2016; Duan and Tang, 2018), iapp-wiki-qa[5], vi-quad (Nguyen et al., 2020), facqa-qa-factoid-itb (Wilie et al., 2020), and TyDi QA (Clark et al., 2020).

[3]https://spacy.io

[4]https://github.com/brmson/dataset-factoid-webquestions

	en	id	th	vi	zh
before	97.25	95.83	93.25	95.45	93.58
after	97.67	95.92	93.33	95.20	95.92
gain	0.42	0.09	0.08	-0.25	2.34

Table 4: SEAREFUSE test set performance after training on km-en translation on the SEAREFUSE English training subset.

	layer 4	layer 5	layer 17	layer 18
vanilla	53.10	66.25	85.25	86.03
parallel	**75.57**	**78.90**	**86.53**	86.48
non-parallel	71.98	77.85	86.00	**86.56**

Table 5: Conducting mean-shifting with parallel examples and non-parallel examples.

G Additional Results on SEAREFUSE

Table 4 presents the per-language probing results of on the SEAREFUSE-H and SEAREFUSE-G concatenated test set, before and after `Qwen2.5-7B` going through Khmer→English training on the SEAREFUSE English training subset. Note that translating from Khmer to English only covers entities that would appear in English contexts. The model's generalization to the SEAREFUSE test set, which comprises entities grounded in each language, shows the effectiveness of the method. As seen in Table 4, the Chinese representations also achieve a large gain even without being directly trained on, in line with the safeguard mechanism of primary language we find in the main paper.

H Additional Results on FRESHQAPARALLEL

Here, we provide additional `Qwen2.5-14B` results on our augmented FRESHQAPARALLEL in Figure 13. While our augmented FRESHQAPARALLEL makes FreshQA much more difficult compared to the original version, bilingual question translation training still brings consistent gain to our mean-shifting and linear projection method, consistently outperforming vanilla zero-shot transfer.

[5]https://github.com/iapp-technology/iapp-wiki-qa-dataset

	ID avg.	OOD avg.
mean pooling	86.12	79.52
last-token pooling	**86.77**	**83.95**

Table 6: In-distribution and OOD performance across different pooling methods.

I Non-parallel Approximation for Mean-shifting

Here, we discuss whether our methods can be applied on non-parallel examples of large scales.

For mean-shifting, larger non-parallel corpus can work as well. We conduct the following experiments with English, Thai and Chinese: To construct non-parallel samples to compute language means, we sample ∼7k examples per language from our SeaRefuse training set, which are **non-parallel** as all questions are grounded in different language's contexts collected from different seed datasets. We run the same FreshQA experiment in the main paper with the three languages, where the parallel setting uses means computed on the fly with FreshQA training examples in each run (∼480 pairs). As shown in Table 5, means computed by non-parallel samples also effectively approximate mean difference across languages, performing on-par with the parallel setting or even outperforming it in some cases (e.g., in layer 18).

In contrast, linear projection naturally requires aligned data. In this work, we show that a few hundreds of parallel data samples suffice to learn a robust projection matrix.

J Pooling Methods

In Table 6, we ablate the effect of pooling methods, validating the choice of last-token pooling for knowledge boundary probing.

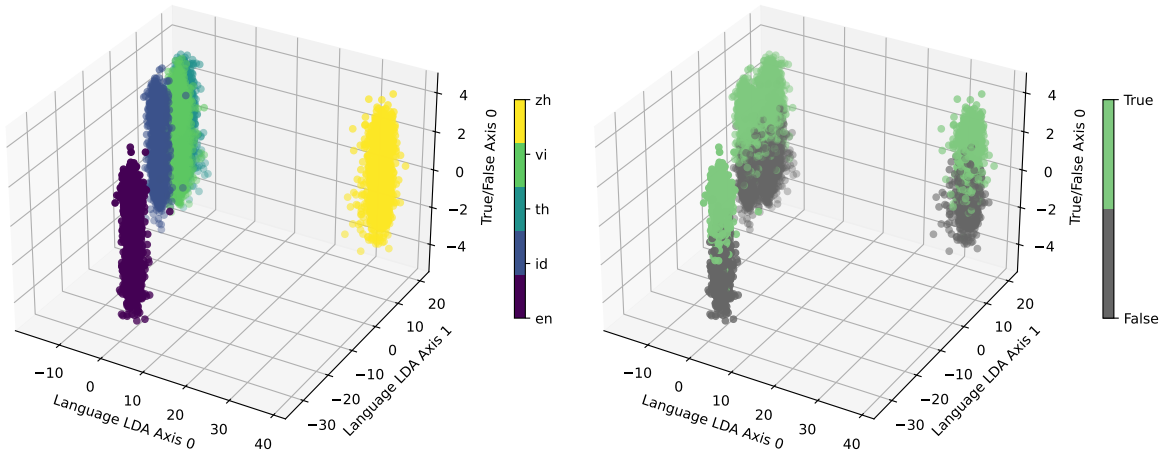

Figure 7: Projecting Qwen-2.5-7B's 19th layer representations onto two languages axes (x,y), and one answerability axis (z) - determined by whether the entities in the question exist in real world.

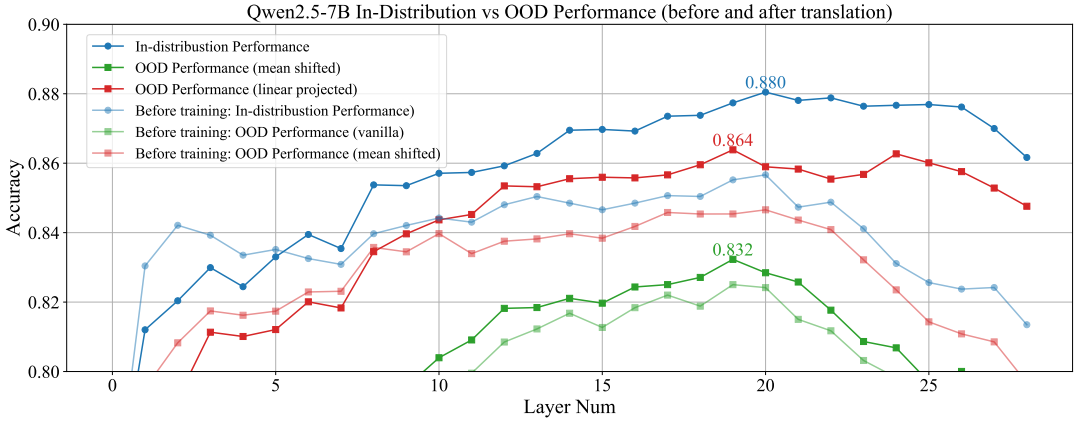

Figure 8: 7B model performance on FreshQA before and after bilingual translation training on km-en question pairs. The results are averaged across all languages, aligning with the setting throughout the paper.

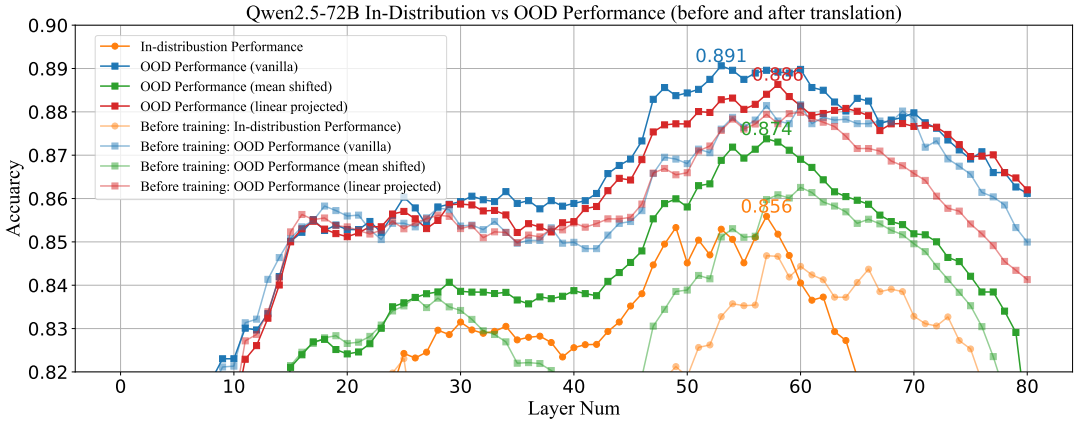

Figure 9: 72B model performance on FreshQA before and after bilingual translation training on km-en question pairs. The results are averaged across all languages, aligning with the setting throughout the paper.

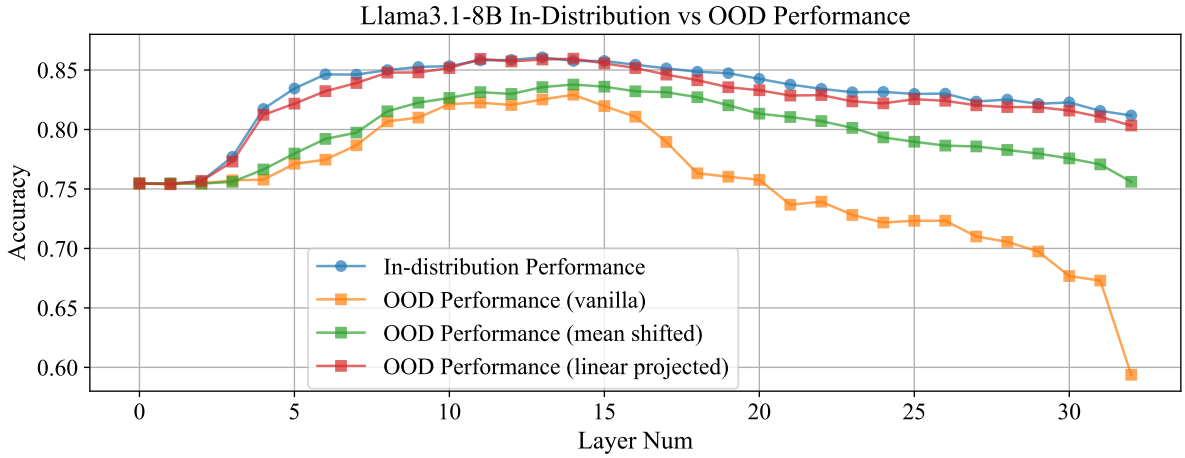

Figure 10: Llama-3.1-8B training-free alignment methods performance.

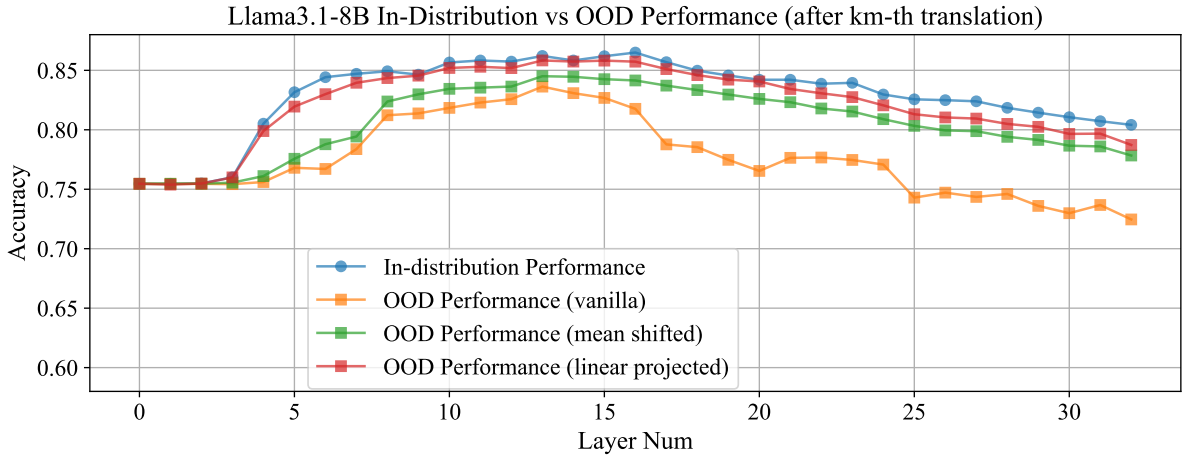

Figure 11: Llama-3.1-8B training-free alignment methods performance, after the model has been SFT'ed on Khmer-Thai translation.

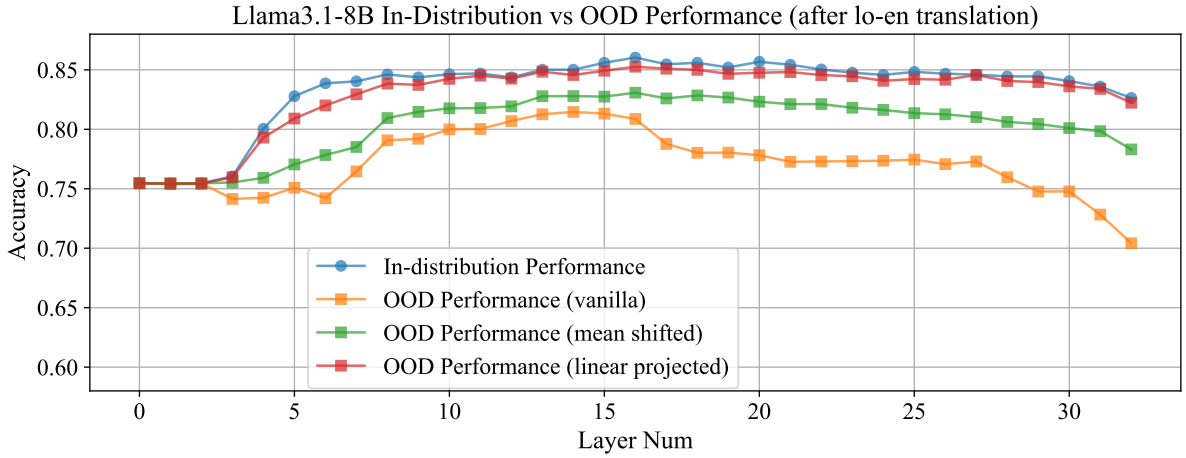

Figure 12: Llama-3.1-8B training-free alignment methods performance, after the model has been SFT'ed on Lao-English translation.

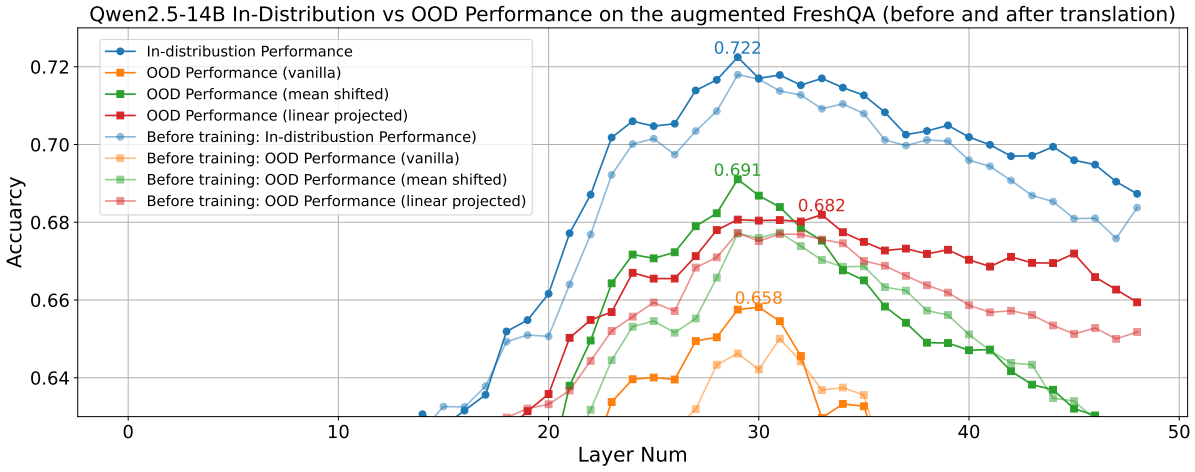

Figure 13: 14B model performance on our augmented FRESHQAPARALLEL before and after bilingual translation training on km-en question pairs.

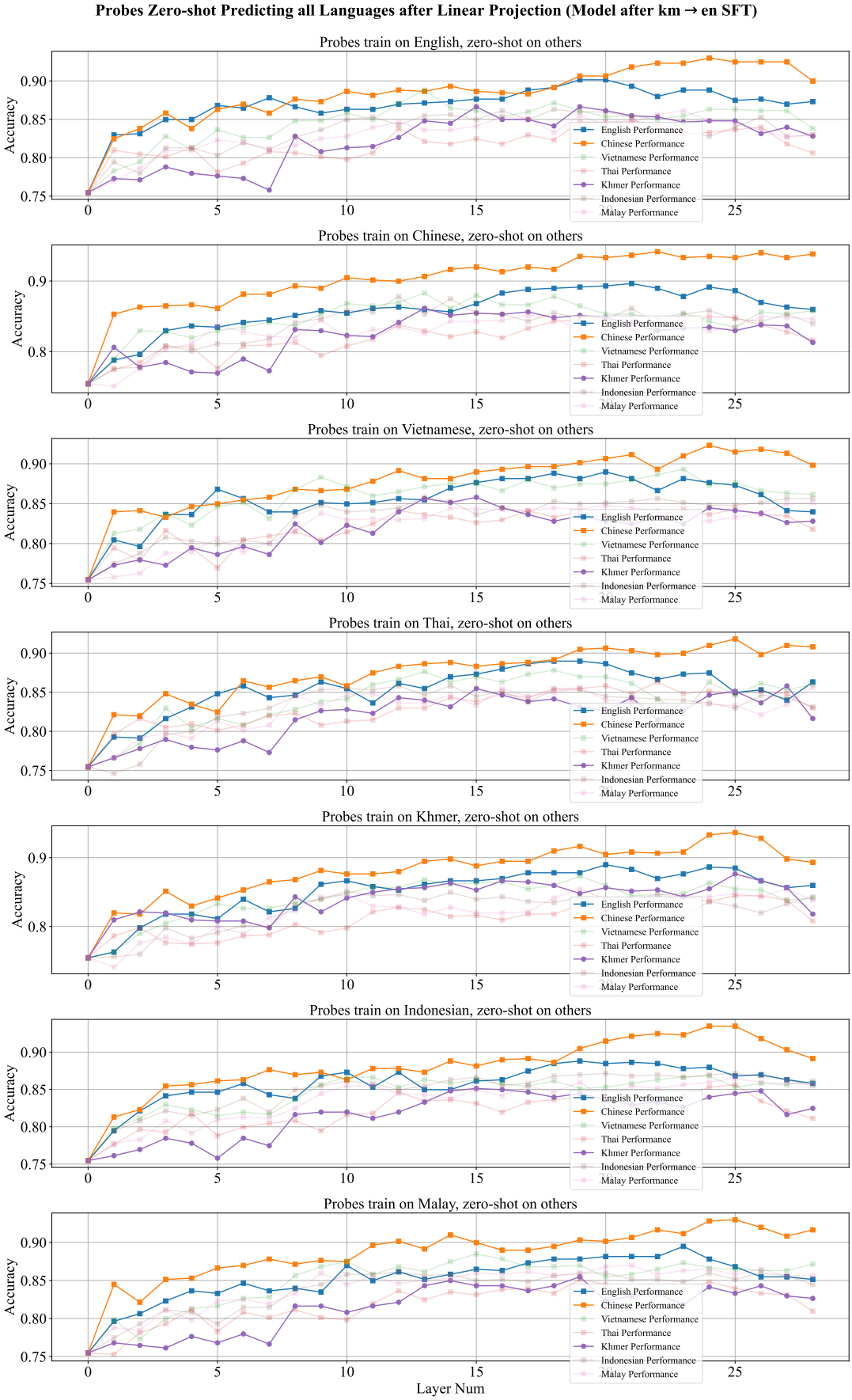

Figure 14: Full transferability results for probes trained on all languages, after `Qwen2.5-7B` has been fine-tuned on Khmer-English translation.