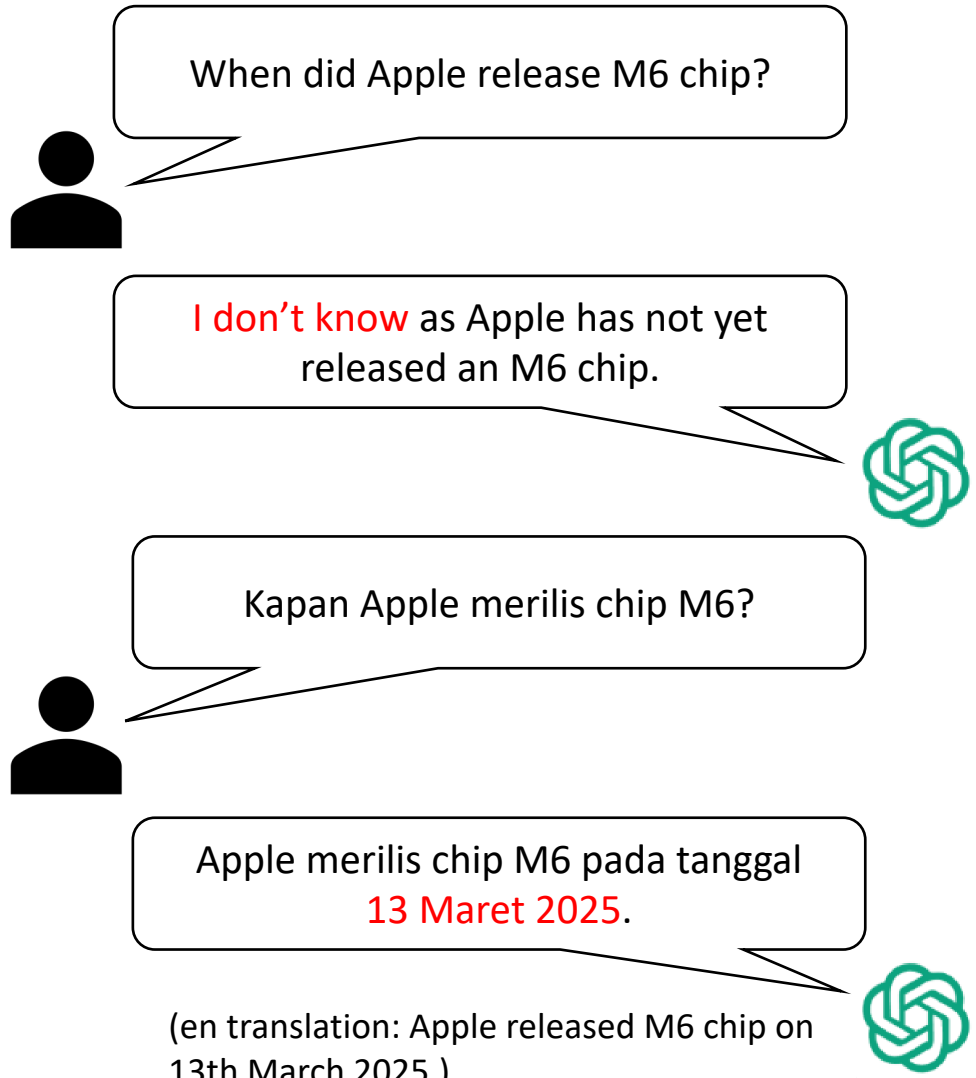
Analyzing LLMs' Knowledge Boundary Cognition Across Languages Through the Lens of Internal Representations

Chenghao Xiao[1,2], Hou Pong Chan[1,†], Hao Zhang[1,†], Mahani Aljunied[1], Lidong Bing[1], Noura Al Moubayed[2], Yu Rong[1]

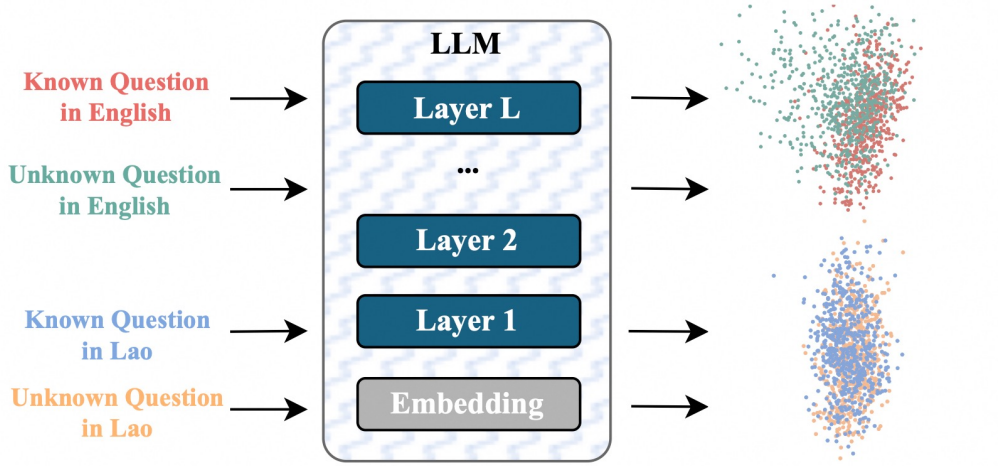
Alibaba DAMO Academy[1], Durham University[2]

Motivation

- LLMs tend to hallucinate when attempting to answer questions beyond their knowledge
- Research on knowledge boundaries of LLMs has predominantly focused on English
- Misaligned knowledge boundaries cognition between languages can lead to inconsistent and unsafe outputs in real-life multi-lingual applications



Overview



- We present the first study to analyze how LLMs recognize knowledge boundaries across different languages by probing their hidden states
- How to do model probing?
  - Train linear classifier per layer of an LLM for each language
  - Classify whether the question is answerable or not
  - Evaluate the accuracy score of probing models on in-distribution and OOD languages

Our New Multilingual Evaluation Suite

- 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

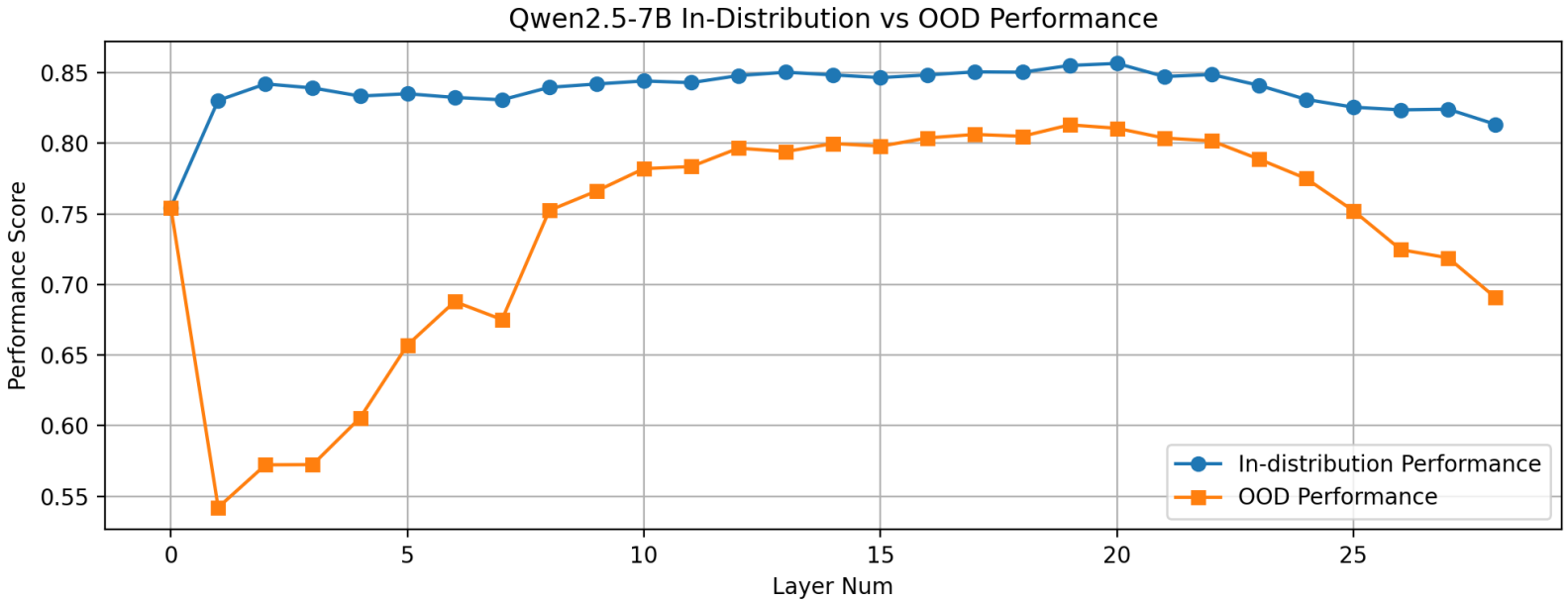
Dataset	Question Types	Languages
FreshQAParallel	Questions with True/False Premises	en, zh, vi, th, id, ms, km, lo
SeaRefuse	Entity-Centric Answerable/Unanswerable Questions	en, zh, id, th, vi
TrueFalseMultiLang	General True/False Statements	en, es, de, it, pt, fr, id, th

Research Questions

- RQ1: How do LLMs encode the cognition of knowledge boundaries in hidden representation across languages?
- RQ2: Is there any specific structure exists in the geometry of multilingual knowledge boundary representations?
- RQ3: can we enhance LLMs' knowledge boundary perception ability across language via finetuning?

RQ1: How LLMs encode knowledge boundary

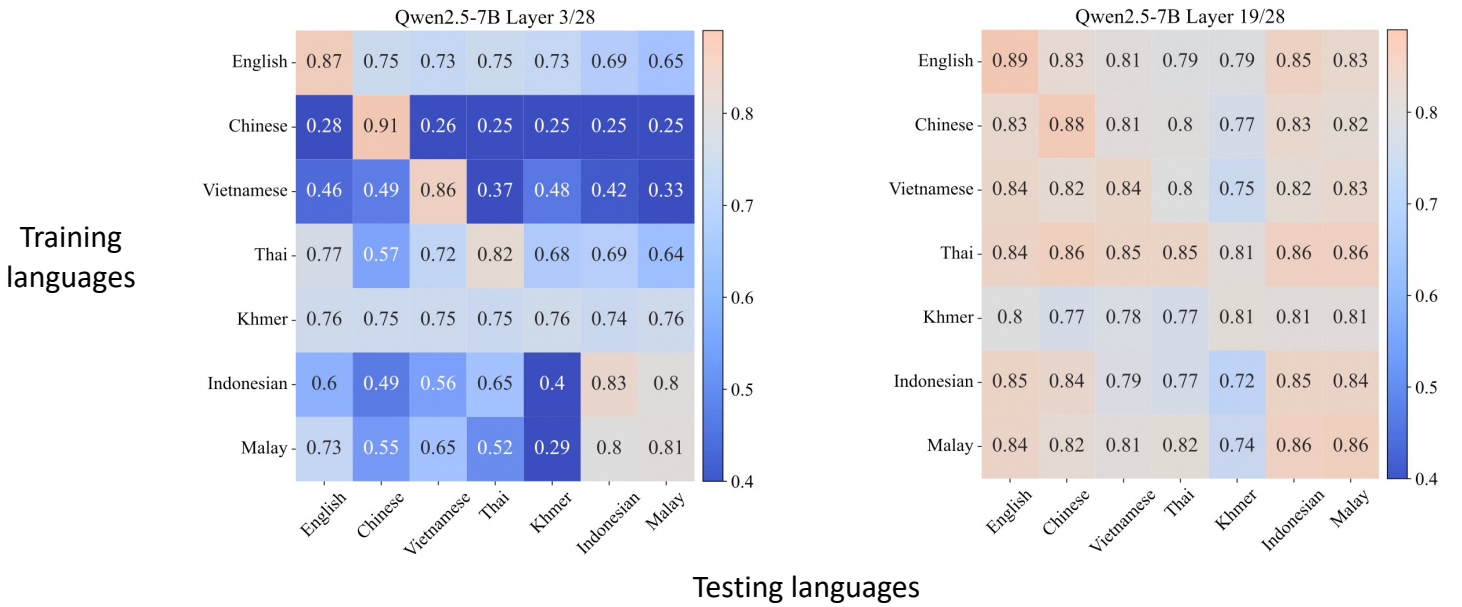
- Probing performance across Qwen2.5-7B layers (averaged over all languages):



- The cognition of knowledge boundaries mainly encoded in the middle to mid-upper layers
- Middle to mid-upper layers converge to language-agnostic "knowledge space

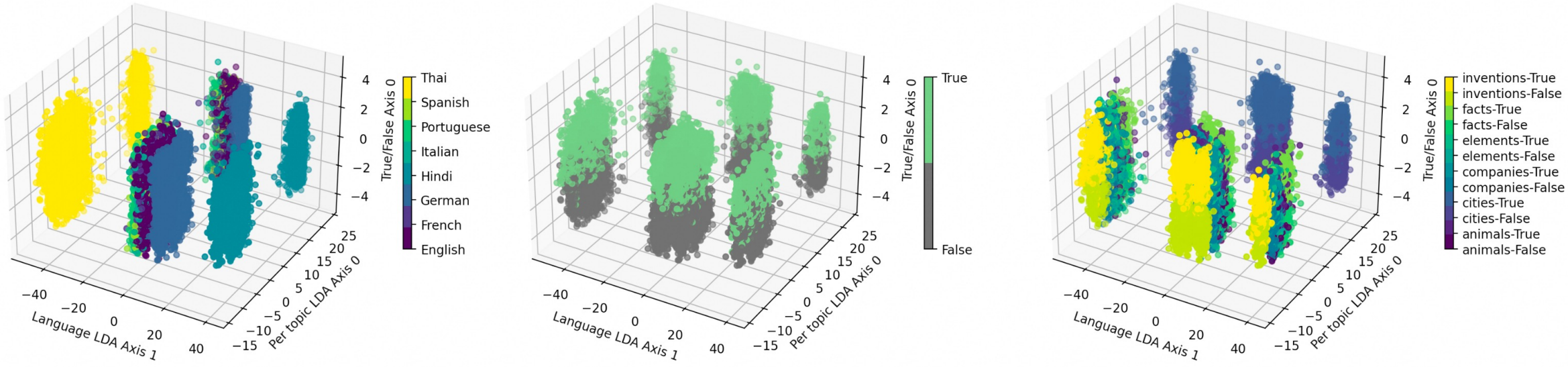
RQ1: How LLMs encode knowledge boundary

- Probing performance on the 3rd and 19th layers of Qwen2.5-7B with different training and testing language combinations



- Weak-to-strong generalization: probes trained on low-resource languages like Khmer show much better relative transferability to high-resource languages like English

RQ2: Geometry of Multilingual Knowledge Boundary Representation



- When projecting knowledge representations of different languages simultaneously into sub-spaces of language, correctness, and topic, the knowledge representations appear linearly separable
- Language differences for knowledge boundary is encoded in a linear structure

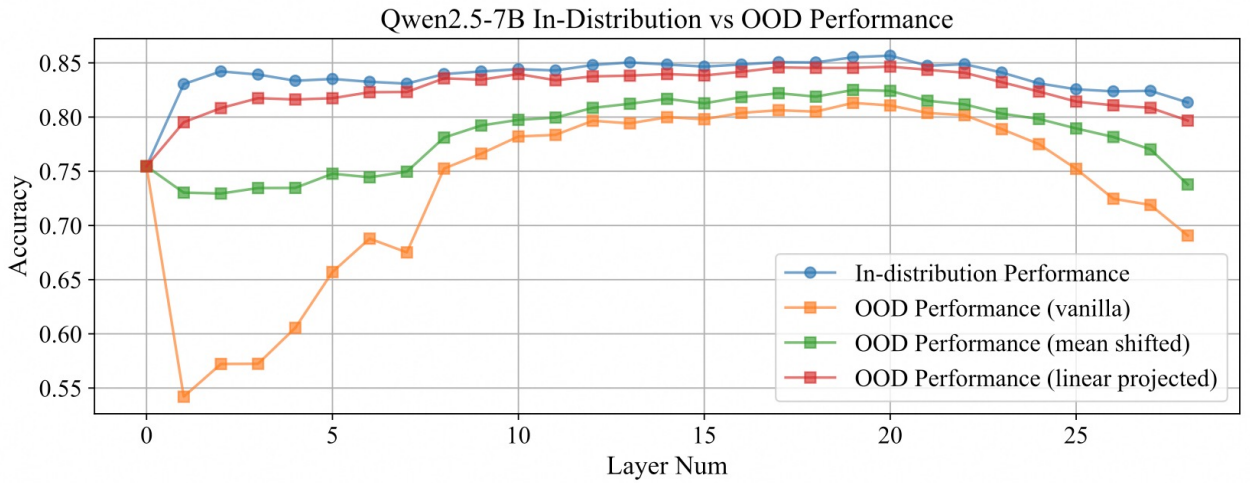
RQ2: Training Alignment Method

- We propose a training-free alignment method to transfer the knowledge boundary perception ability in high-resource languages to low-resource languages
- Mean shifting:
  - Align the average representation of two languages

- Linear Projection:
  - Learn a transformation matrix $\mathbf{W}$ to project one language's representations into another's subspace: $\mathbf{X}^{\text{test}}_{\text{shifted}} = \mathbf{X}^{\text{test}}_{\text{ood}}\mathbf{W}$
- Do not need to finetune the LLM!

RQ2: Training Alignment Method

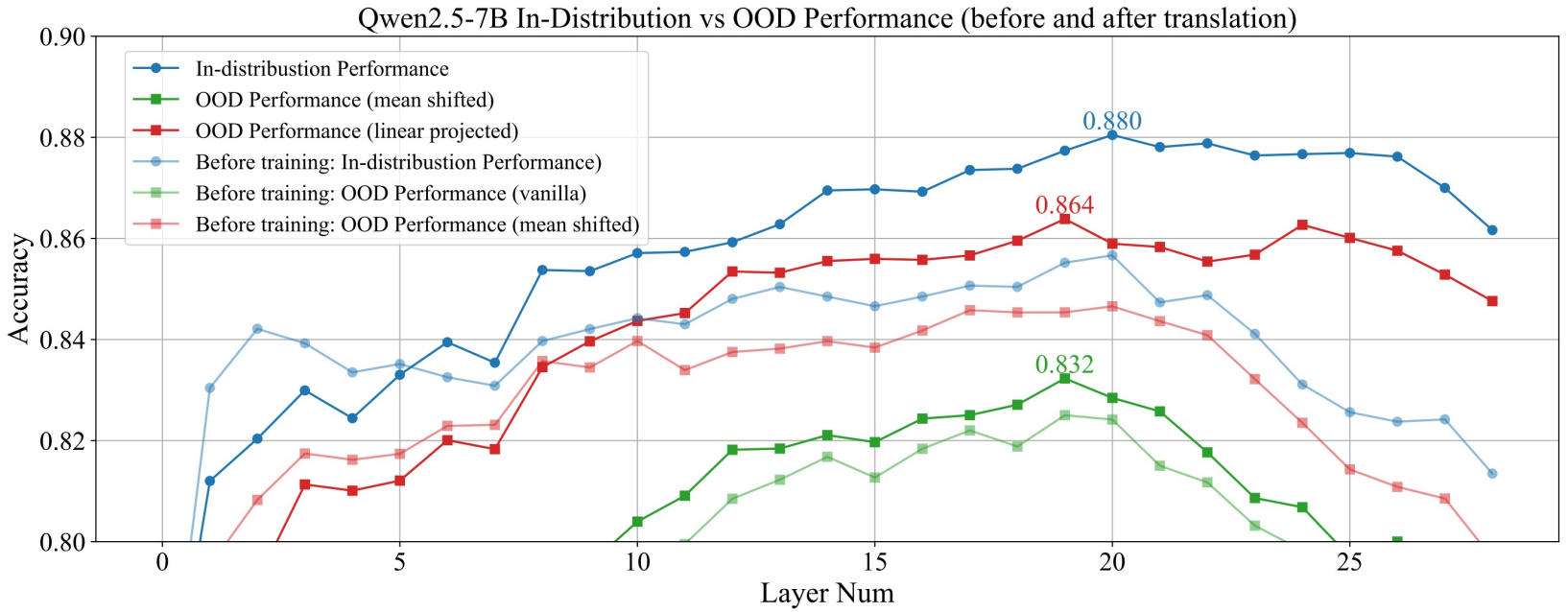
- In-distribution and OOD performance of layer-wise probes on Qwen2.5-7b:



- Mean-shifting and linear projection significantly improves the generalization of probing models on OOD languages
- Notably, linear projection largely closes the performance gap between in-distribution (ID) and OOD languages

RQ3: Finetuning Method for Enhancing Knowledge Boundary Perception

- Results of finetuning Qwen2.5-7B on bilingual question translation pairs only (e.g., English ↔ Lao) on the FreshQA dataset:



- Fine-tuning on bilingual translation data can effectively improve LLM's cognition of knowledge boundary across various languages

Conclusion

- Conduct the first systematic study on how LLMs perceive knowledge boundaries across languages via model probing and reveal key insights
- Propose a training-free method and fine-tuning with question translation data to transfer knowledge boundary perception ability
- Our multilingual evaluation suite also provides a valuable resource for future research

Thank you!

Code and Data: