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]

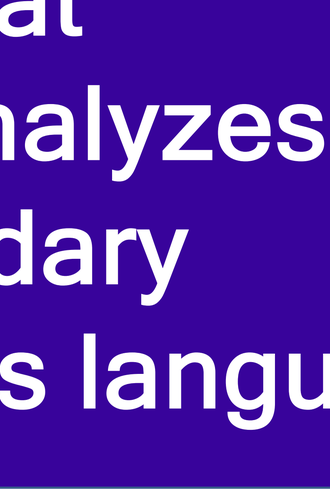
How LLMs encode knowledge boundary across layers

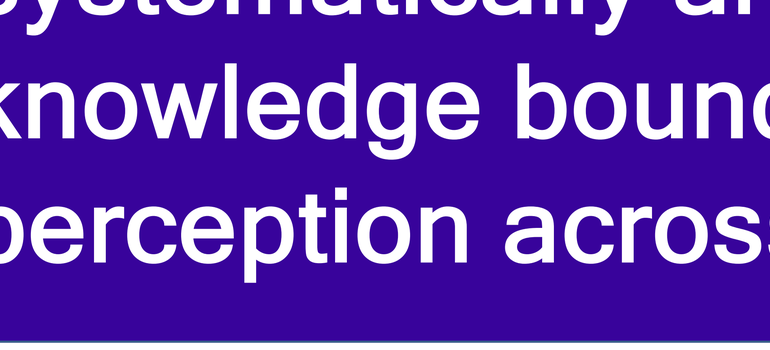
Main method

- Probe internal representations of LLMs faced with knowledge boundary data (e.g., answerable and unanswerable questions)
- Evaluate probe performance on in-distribution and OOD languages.

Locating knowledge boundary representations in LLMs

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

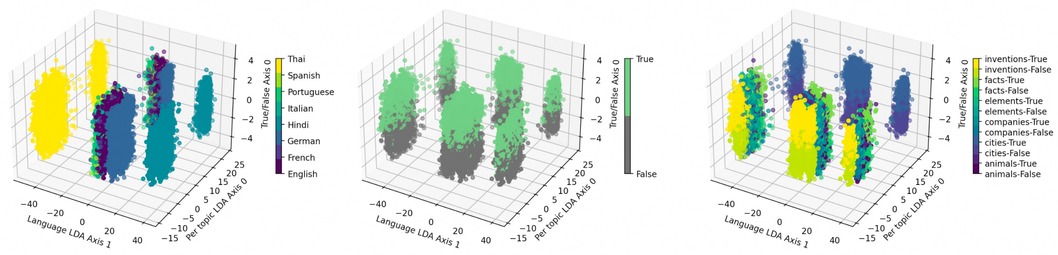




Linear Geometry of knowledge boundary representations

Linear Structure

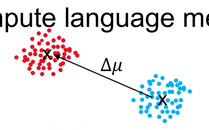
- Project knowledge representations of different languages into sub-spaces of language, correctness, and topic, knowledge representations appear linearly separable



Training-free alignment methods

- Mean-shifting.

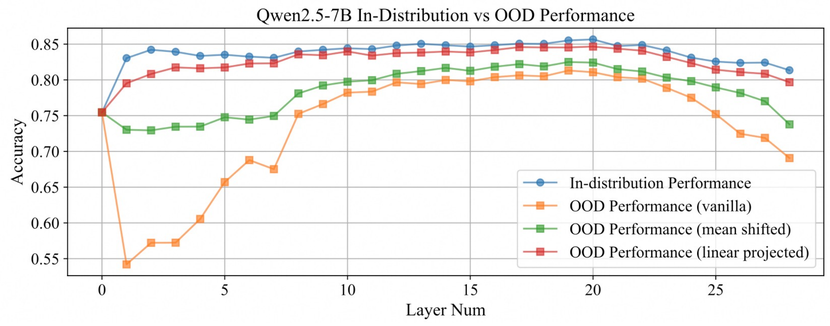
Compute language mean difference.



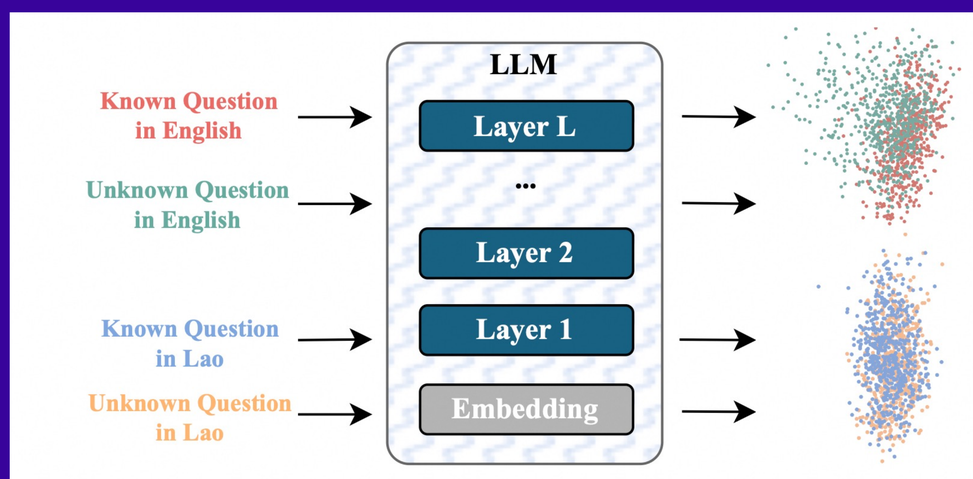
- Linear Projection

Learns a transformation matrix $\mathbf{W}$

$$\mathbf{X}^{\text{test}}_{\text{shifted}} = \mathbf{X}^{\text{test}}_{\text{ood}}\mathbf{W}$$



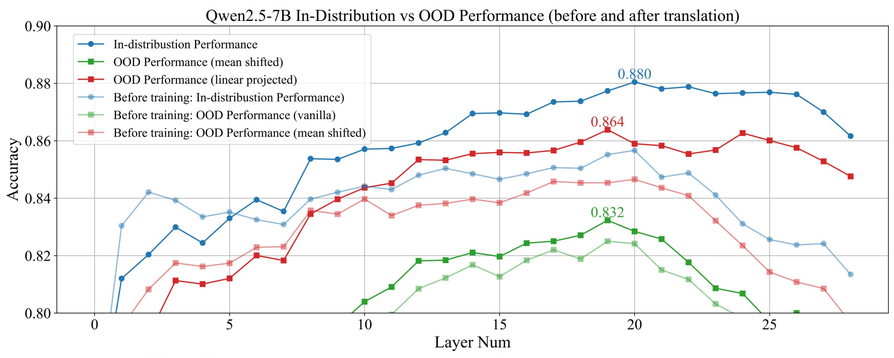
The first study that systematically analyzes LLMs' knowledge boundary perception across languages



Can we further enhance knowledge boundary across languages?

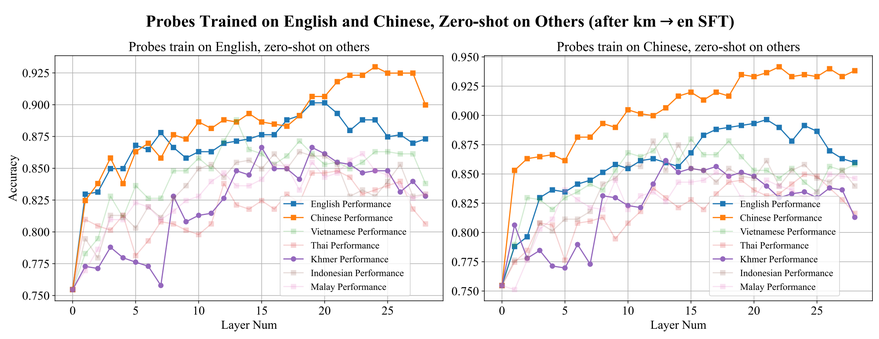
Bilingual Question-only Finetuning

- It turns out that fine-tuning on bilingual question pairs (Lao -> English) can improve knowledge boundary representations across all languages.



A surprising ``defense'' mechanism of primary language.

e.g., Chinese representations attain a surprising enhancement after Khmer->English fine-tuning.



A multilingual knowledge boundary evaluation suite

Multilingual Evaluation Suite

- Given the absence of standard testbeds for multilingual 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
TrueFalseMultilingual	General True/False Statements	en, es, de, it, pt, fr, id, th

- FreshQA-Parallel

We augment FreshQA by flipping each question's premise, and expand to 8 languages.

- SeaRefuse

Questions with Existing and non-existent entities, in southeast Asian languages.

- TrueFalse-Multilingual

We expand the commonly use TrueFalse Statements dataset into 8 languages.

Code and Data: