

Do multilingual LLMs have a single latent language? It depends on how you measure it!

Multilingual LLMs transfer information across languages. How is that information internally organized?

Hypothesis 1: Language-specific but aligned representations

Turkish space French space English space  
kedi chat cat

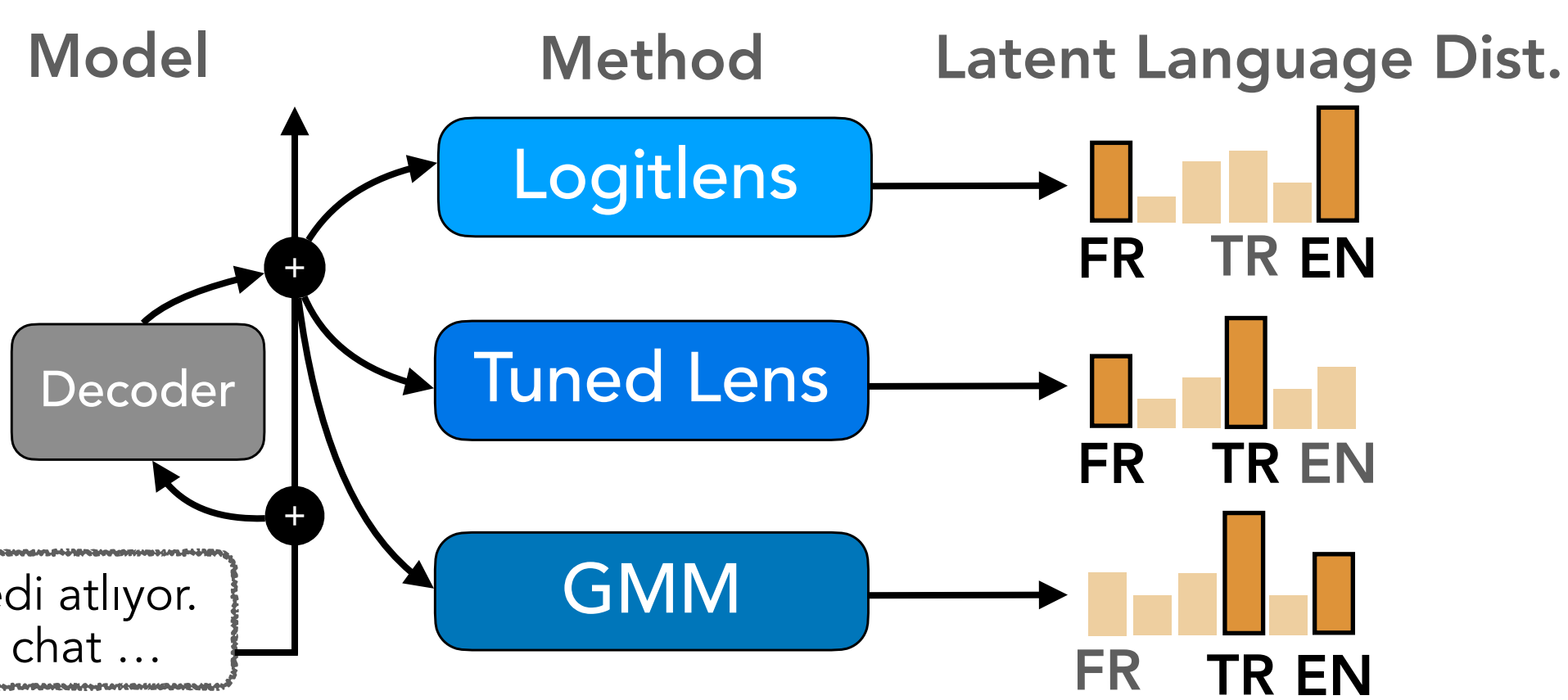
Hypothesis 2: Language-neutral shared concept space

kedi chat cat → [CAT concept] → output lang

Hypothesis 3: Language-mediated sharing via a pivot

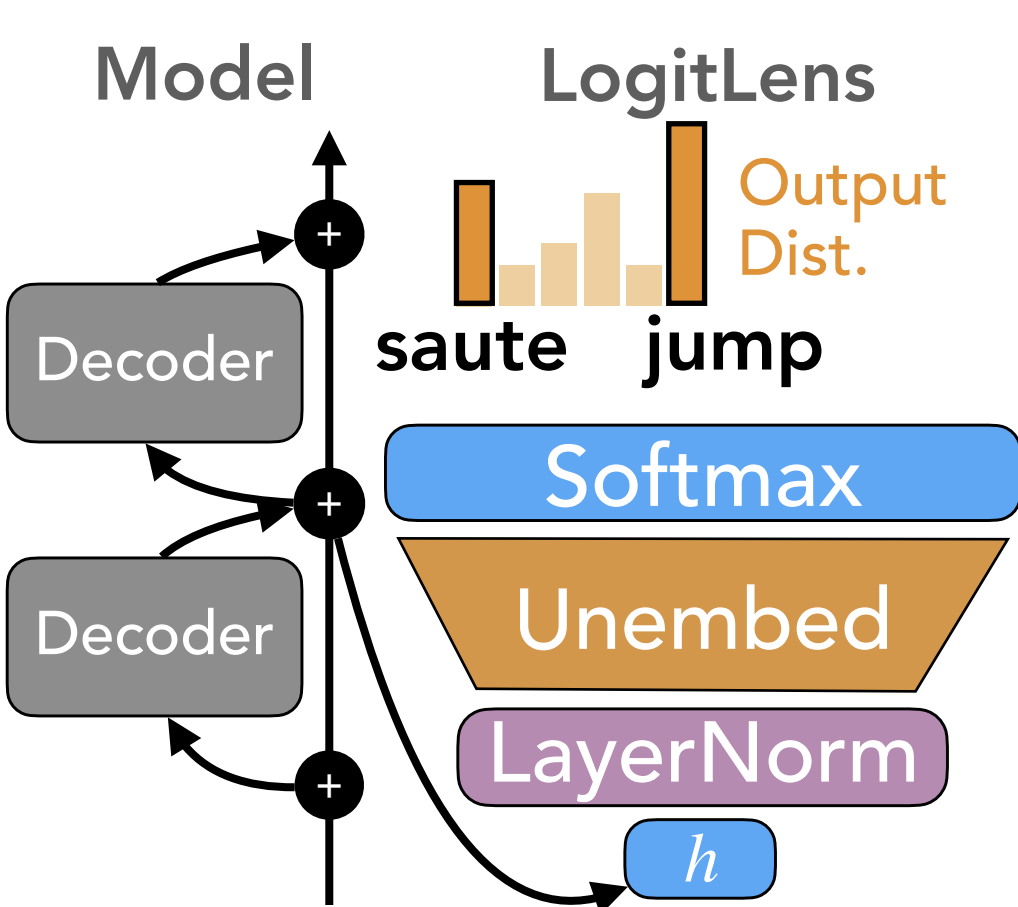
TR input → EN-like internal state → FR output

Latent Language Identification

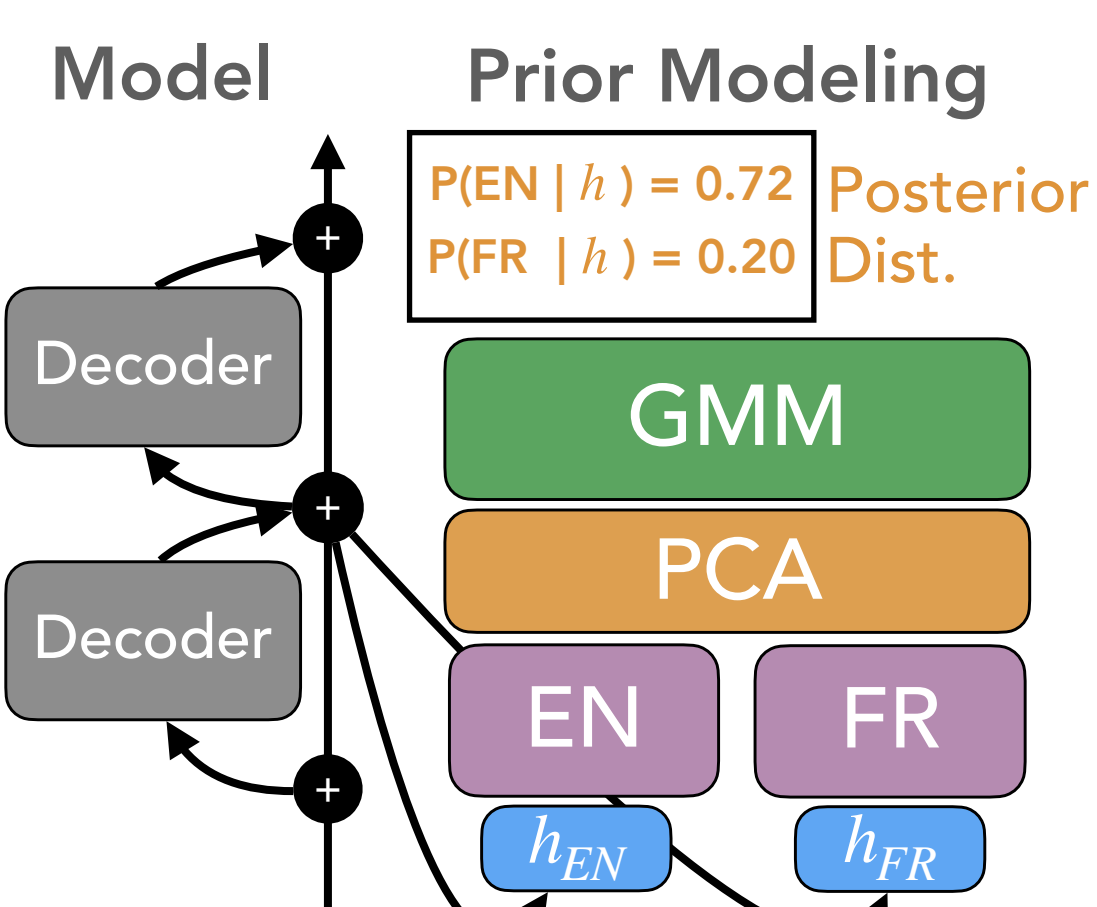


Existing work infers latent language using different evidence.

Evidence 1: Decoding-based



Evidence 2: Representation-based



Do these probes ever recover the same latent language or different?

Evaluation Setup

Tasks

Controlled

- Copy
- Cloze
- Translation

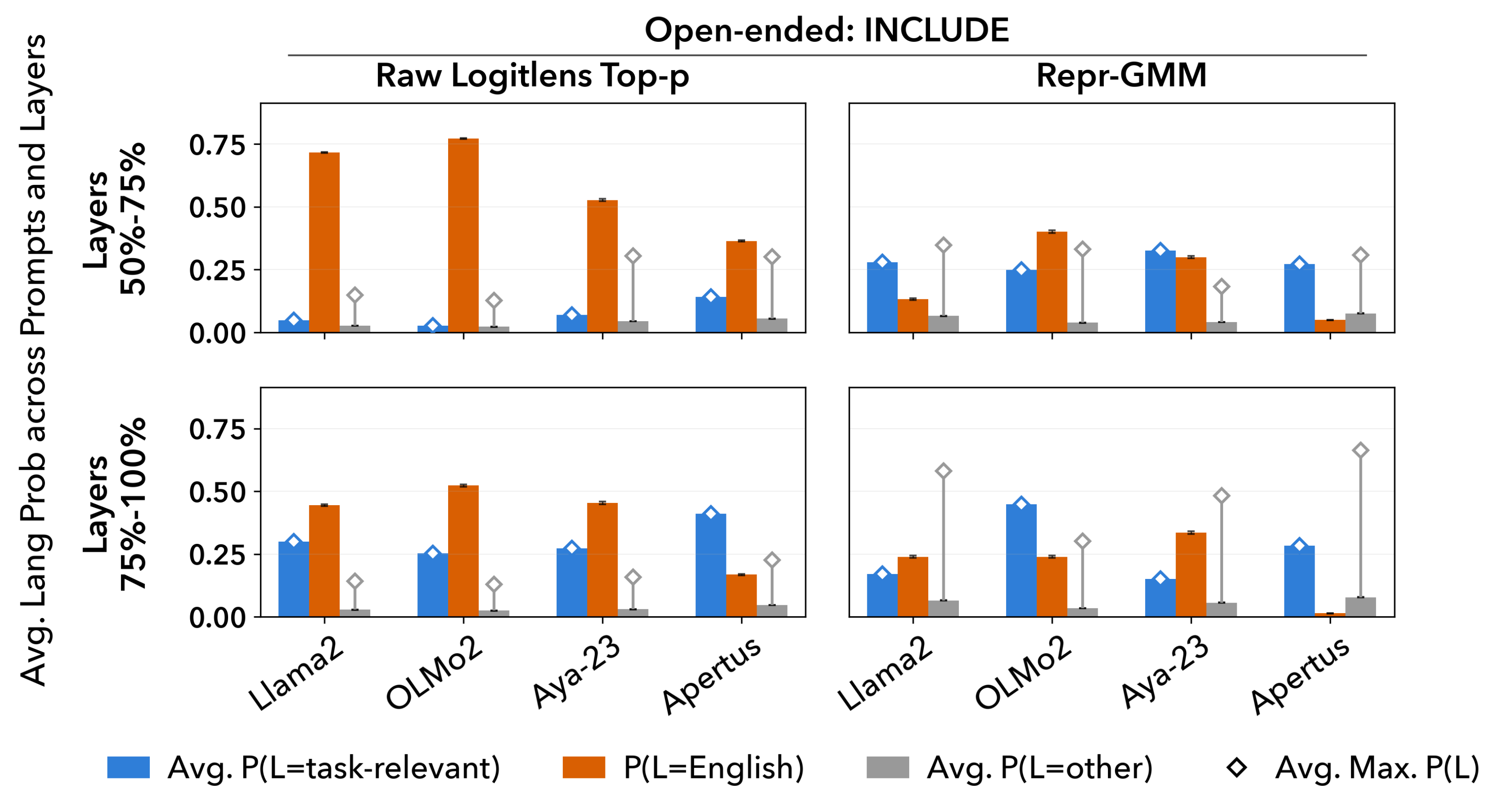
Open-Ended

- PUD
- INCLUDE

27 Languages

12 Models

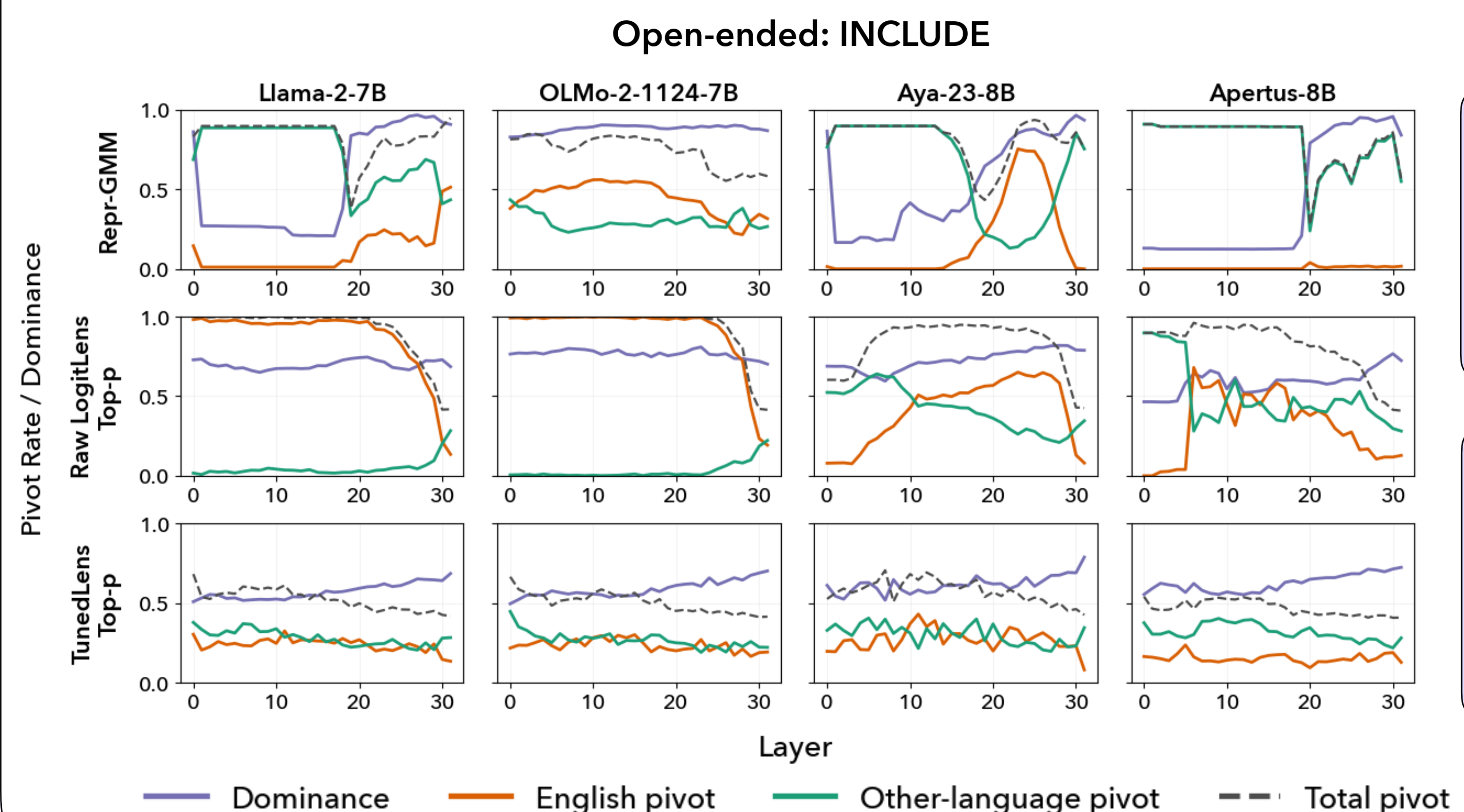
Probes systematically disagree.



Representation probes become task-language aligned earlier.

Decoding probes retain a stronger English signal.

Even choices within a probe family change the answer.



Raw logitlens appears more English dominant.

Tuned lens appears more task-language aligned.

What drives disagreement?

SIGNIFICANT

- Probe family
- Model multilinguality
- Training stage

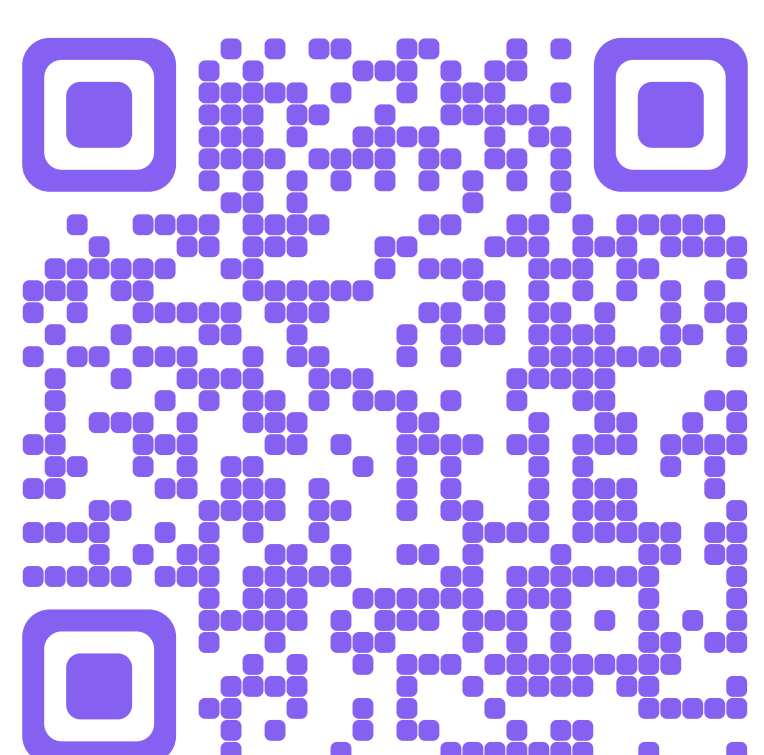
MODEST

- Instruction tuning

NEGLIGIBLE

- Domain / Topic (e.g., Math vs. Social Sciences)
- Dataset language coverage

Scan for Paper & Code



There is no probe independent latent language trajectory.

The disagreement has an interpretable structure:  
GMM-based probes mix earlier;  
Logitlens-based preserves stronger English signals.

Treat LLID as diagnostic evidence, not ground truth.