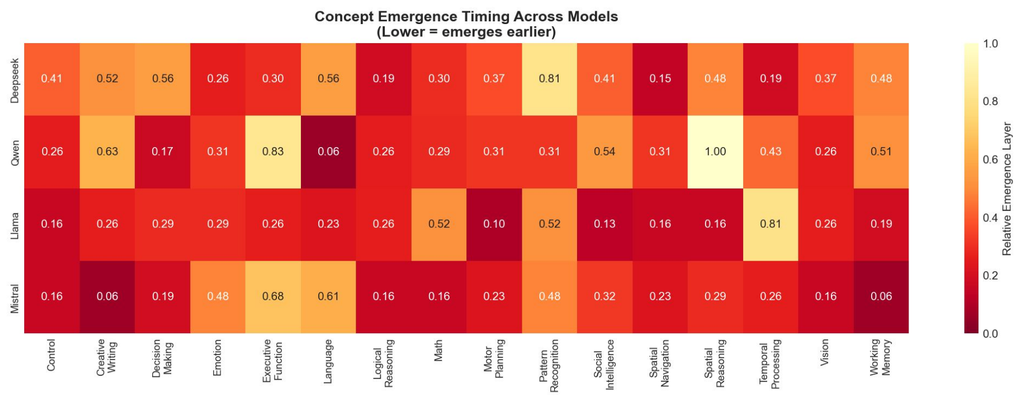
LLMs process cognitive categories in a shared early-to-late order

Layer-Wise Category Structure in Large Language Models: A Cross-Architecture Analysis of Feature Decodability

Background: We wanted to know if different LLM architectures organize task capabilities (like logic vs. vision) similarly across their layers. We probed 4 architectures across 16 task categories.

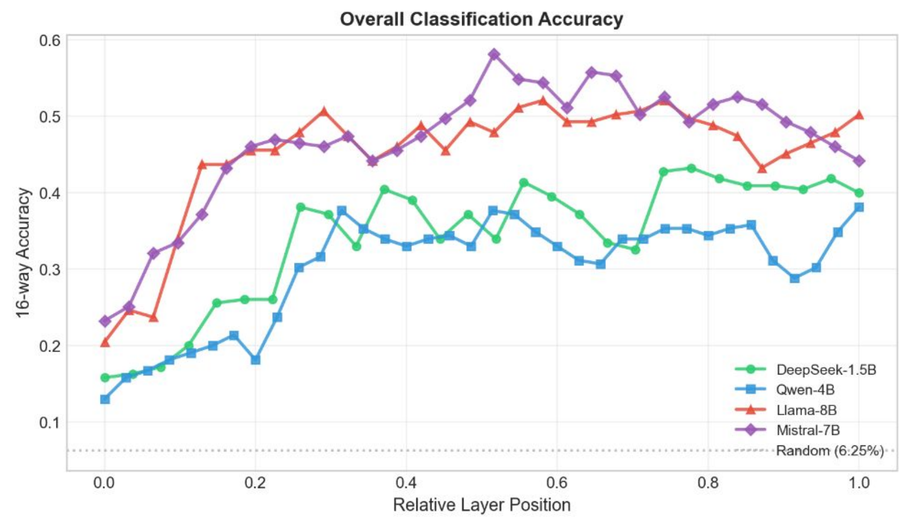
Result 1: Early vs. Late Categories

Spatial Navigation and Logical Reasoning consistently decode early across models, while Executive Function and Pattern Recognition emerge later. Overall, each model is early dominant.

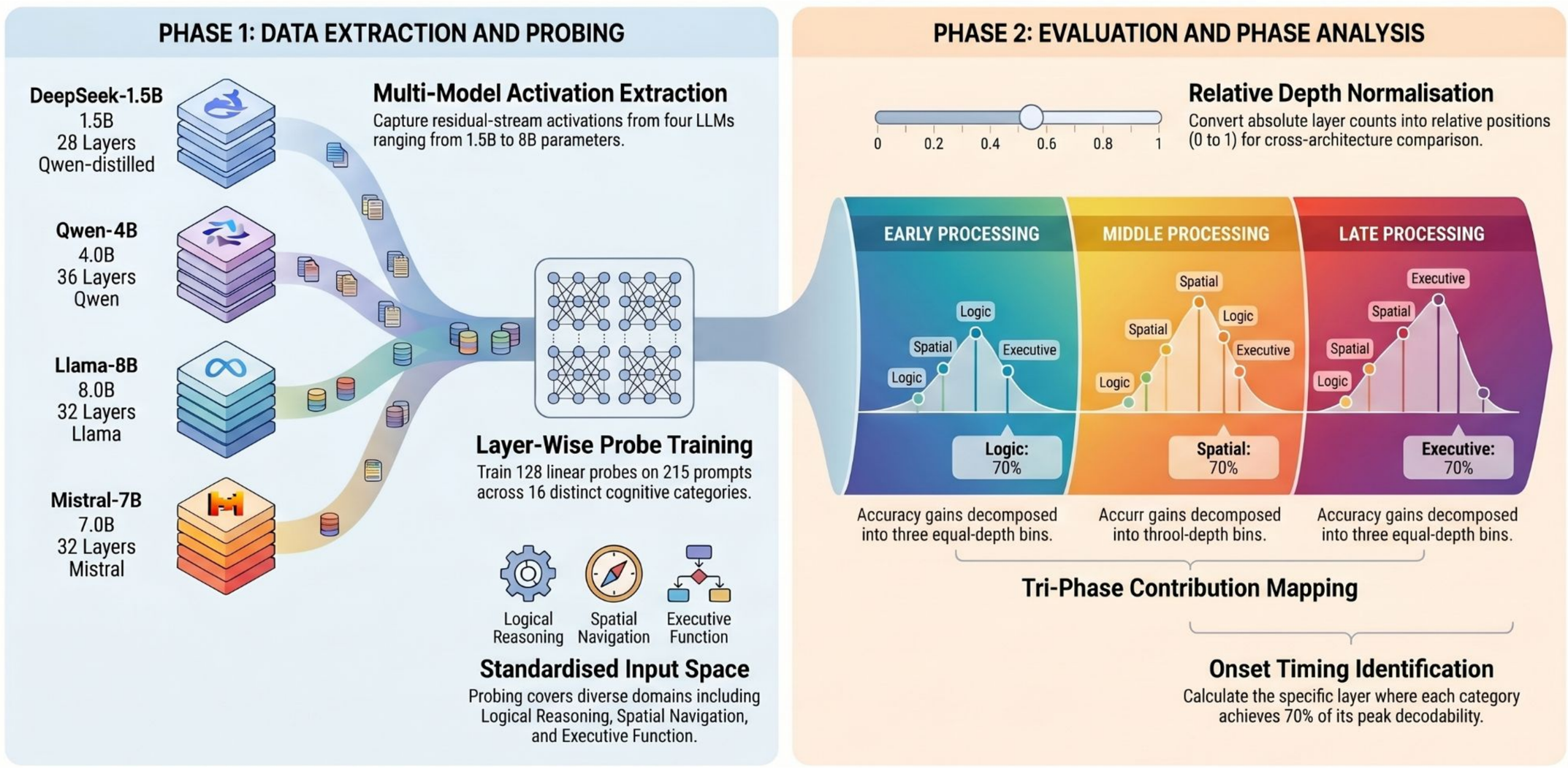


Result 2: Late-Layer Architectural Shifts

While the early processing order is shared, models diverge at the end. Mistral-7B shows late-layer accuracy degradation, suggesting its final layers optimize for output behavior over separable features.



Methods



Limitation: Our 215-prompt design is underpowered for fine-grained ranking claims, and the observed orderings are sensitive to preprocessing choices and prompt phrasing.

PAPER

Guus Bouwens

Mechanistic Interpretability Workshop, ICML, Seoul, South Korea, 2026

ICML International Conference On Machine Learning