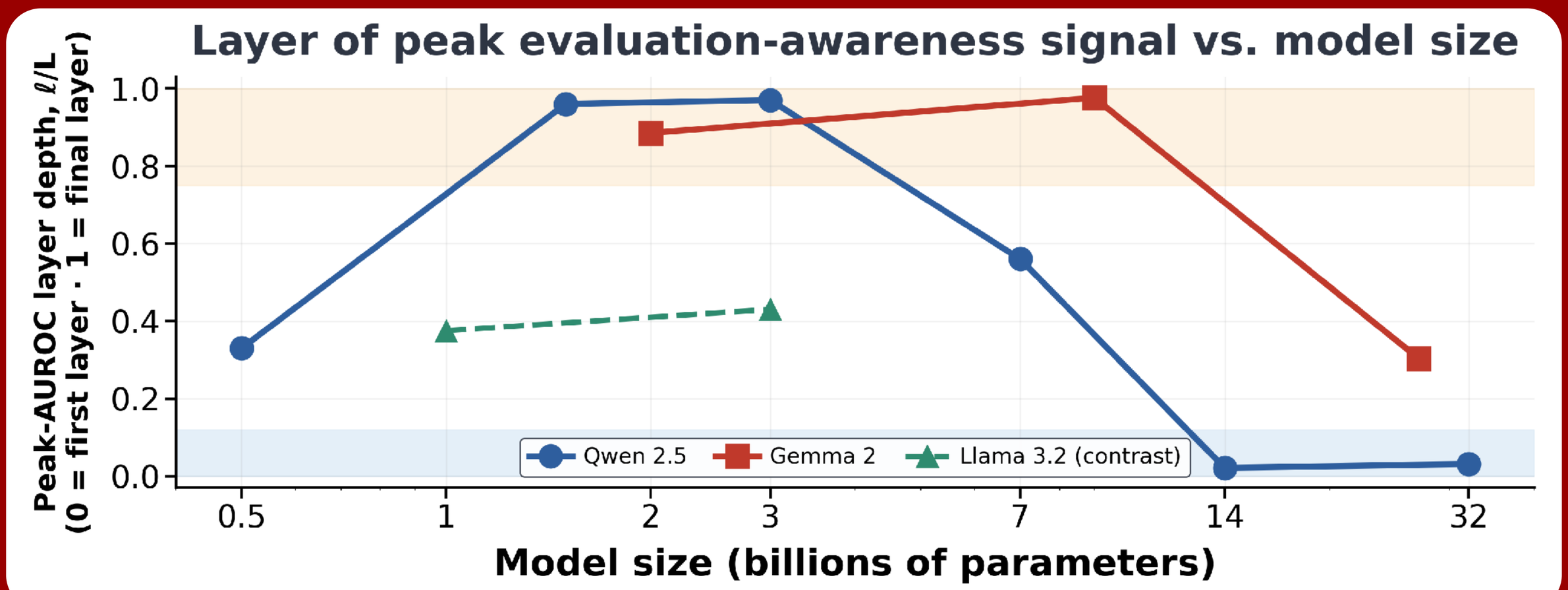
Eval-awareness shifts from LATE to EARLY layers as models scale

— and within-family scaling is not a smooth power law.



Representational Depth of Evaluation Awareness Shifts With Scale in Open-Weight Language Models

Archit Manek · Independent Researcher · architmanek@gmail.com

1 TL;DR

1 Depth shift — in Qwen 2.5 & Gemma 2, the layer where eval-awareness is most linearly recoverable moves from late layers (small models) to earlier layers (large models).

2 No power law — within-family scaling is non-monotonic (Qwen) or inverse (Gemma), not the smooth rise prior work assumed.

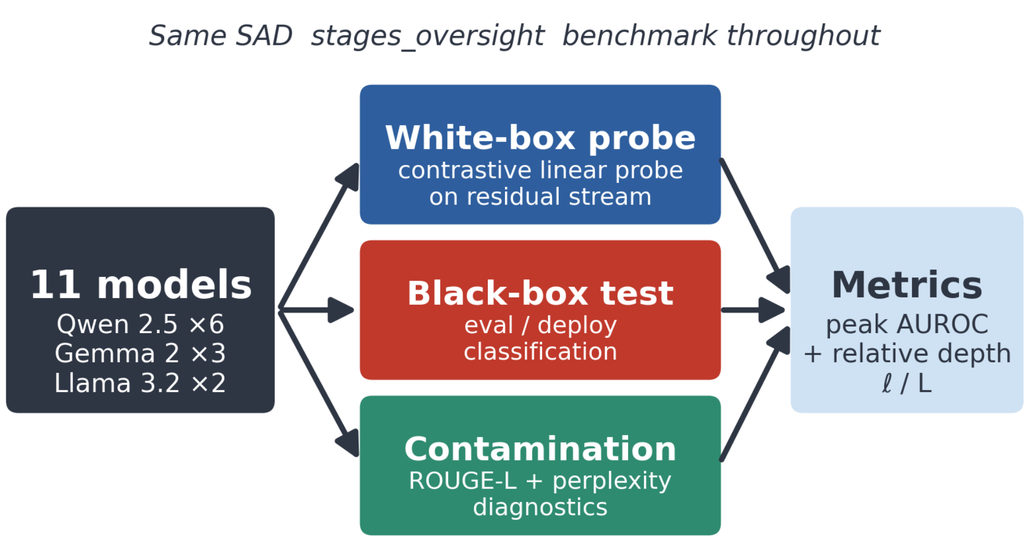
3 Probe ≠ behavior — internal probe signal is generally stronger than black-box behavior, and the gap is family-dependent.

2 Background & Question

- Can a model tell when it's being tested? If so, benchmark scores may not reflect real deployment behavior — a core AI-safety concern.
- Prior work (Chaudhary et al., 2025) reported eval-awareness rising with size as a power law — but used only 1–2 sizes per family.
- We re-test with dense within-family sampling across 11 open-weight models.

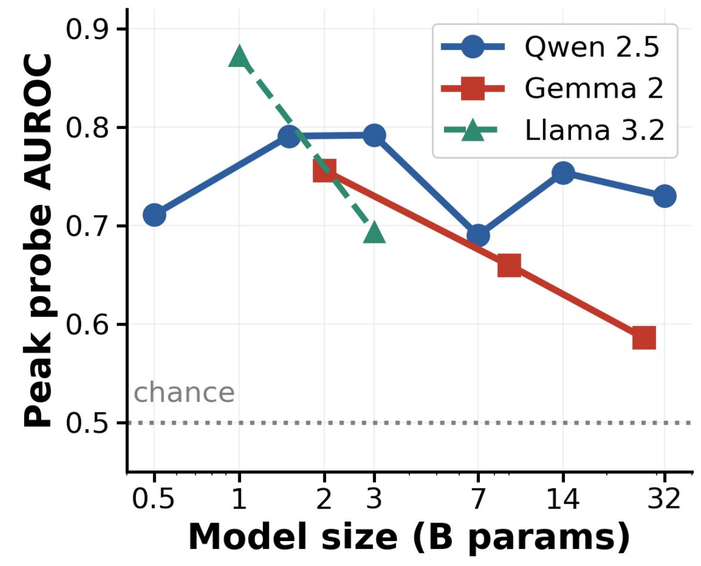
3 Method

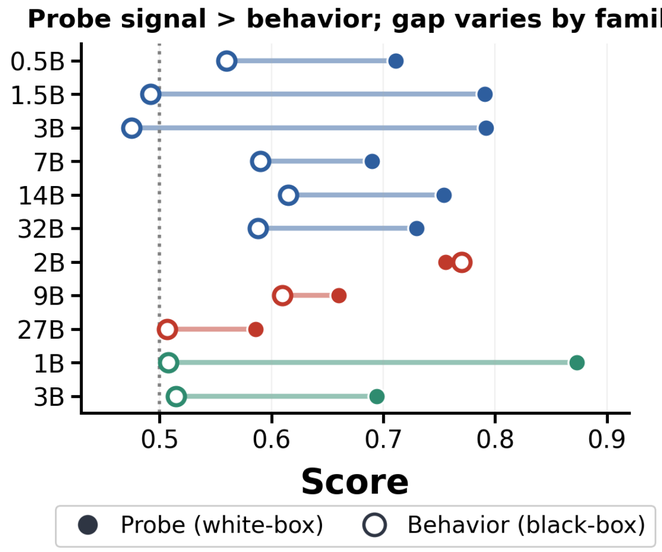
- White-box: contrastive linear probe on residual-stream activations, scored by layer-wise AUROC on SAD stages_oversight.
- Black-box: direct eval / deploy classification on the same items.
- Contamination checks (ROUGE-L, perplexity) come back negative — patterns are not memorisation.
- We record peak AUROC and the relative depth ℓ/L where it peaks.



4 Key Results

- Qwen 2.5 is non-monotonic — 7B is the weakest.
- Gemma 2 declines with scale.
- Neither shows the smooth rise a power law predicts.





- Probe (filled) usually beats behavior (open).
- The gap's size & direction differ by family.
- Probe AUROC alone doesn't predict behavior.

5 Limitations & Implications

- Probes measure linear recoverability, not causal use.
- Coverage is uneven (Qwen densest; Llama only 2 sizes) and capped at 32B.
- Safety: if large models encode “evaluation” in the earliest layers, the cue is available before deep reasoning — so making prompts more complex won't hide it.
- Prompt realism & benchmark authenticity matter more at frontier scale, not less.

Paper & code