Reusable Uncertainty Representations Support Metacognitive Behavior in Llama-3.3-70B

Kirubeswaran O R · Chris Ackerman

TU Darmstadt · MATS

kirubeswaran.obula@tu-darmstadt.de

1 PROBLEM

Knowing what you don't know is a hallmark of intelligent behavior. For LLMs, this means monitoring uncertainty well enough to report confidence, withhold unreliable answers, or delegate. But these behaviors can also be shaped by prompts and surface cues.

Do QA-derived uncertainty directions get reused across metacognitive tasks?

2 SETUP

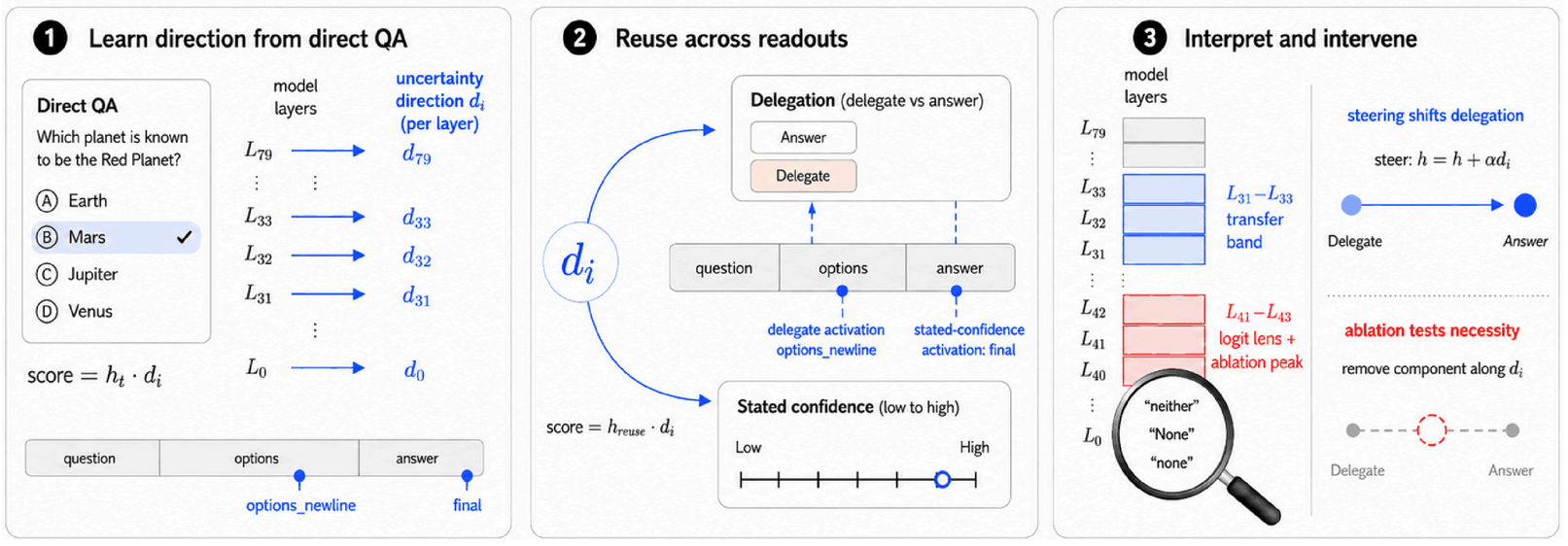
Model
Llama-3.3-70B-Instruct

Datasets
TriviaMC & PopMC
500 MCQs each

Source task
Direct A/B/C/D question answering

Meta-readouts
Stated confidence & answer-vs-delegate

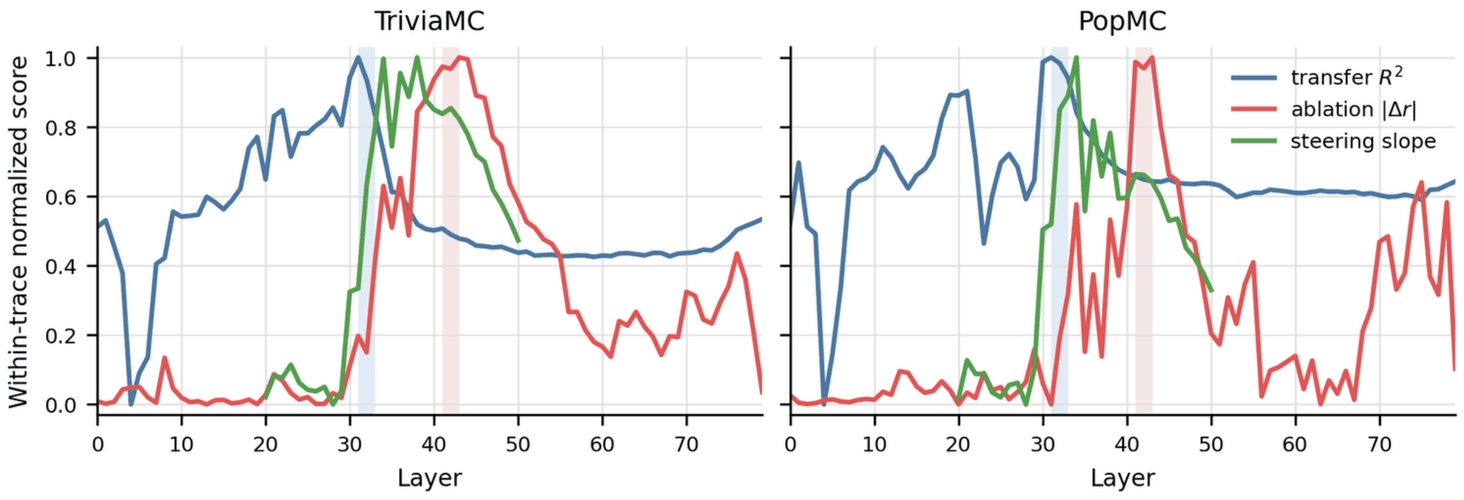
3 METHOD



Learn a layerwise uncertainty direction from **direct QA**; apply the **same frozen direction** to confidence and delegation prompts with no refitting; then interpret and intervene.

4 MAIN RESULTS

Layer localization of effects



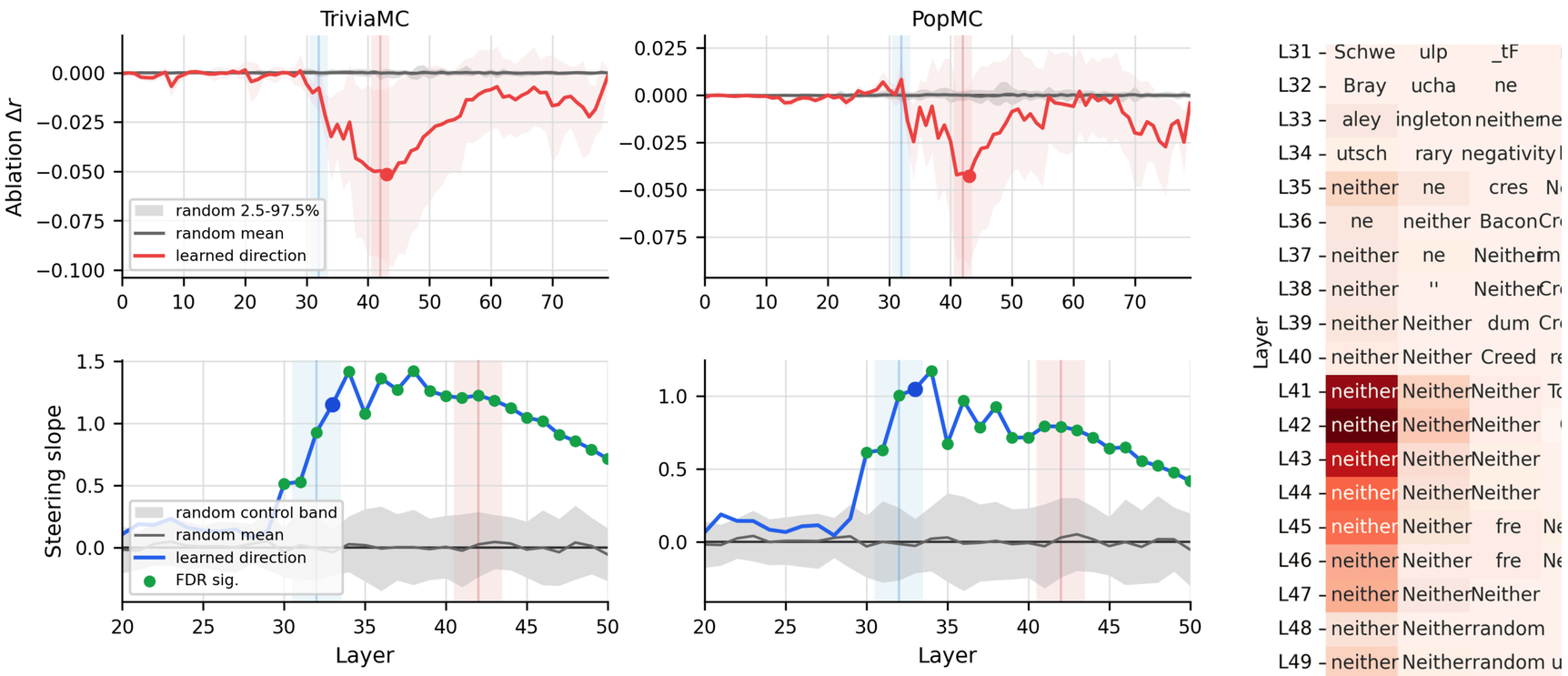
Traces are normalized within each measure: read layer location, not height. **Transfer** (blue) peaks around L31 to L33; **steering** (green) becomes strong in the same mid-layer range; **ablation** (red) peaks later around L41 to L43.

Key numbers, R^2 unless noted (top-logit direction)

Dataset	Delegation transfer R^2	Confidence transfer R^2	Confidence coupling r
TriviaMC	0.405	0.577	0.364
PopMC	0.327	0.373	0.544

Key result: QA-derived directions transfer across readouts. **Confidence tracks the direction more cleanly;** delegation is weaker because it adds a policy decision: whether uncertainty should lead to deferral. Directions also align across datasets, suggesting the effect is not a single-dataset artifact.

5 INTERVENTION



Top: ablation Δr against random controls. **Bottom:** steering slope across L20 to L50; green dots mark FDR-significant layers. **Right:** logit lens shows abstention-like tokens peaking around L41 to L43.

Steering the QA-derived direction shifts delegation, supporting **sufficiency**. **Ablation is weaker:** removing the direction only modestly reduces the uncertainty→delegation link, so the direction is **not uniquely necessary**. Logit lens maps the same late-layer band to **abstention-like tokens** such as none and neither.

TAKE-HOME

QA-derived uncertainty directions transfer to confidence and delegation prompts without refitting, align across datasets, and can causally shift delegation behavior. **Main lesson: separate the internal representation from the behavior.** The uncertainty signal can be present even when delegation reads it out weakly.

6 LIMITATIONS

- One model family (Llama-3.3-70B-Instruct).
- Two factual multiple-choice datasets.
- Steering shows sufficiency; ablation does not establish unique necessity.

Paper under review · draft available on request