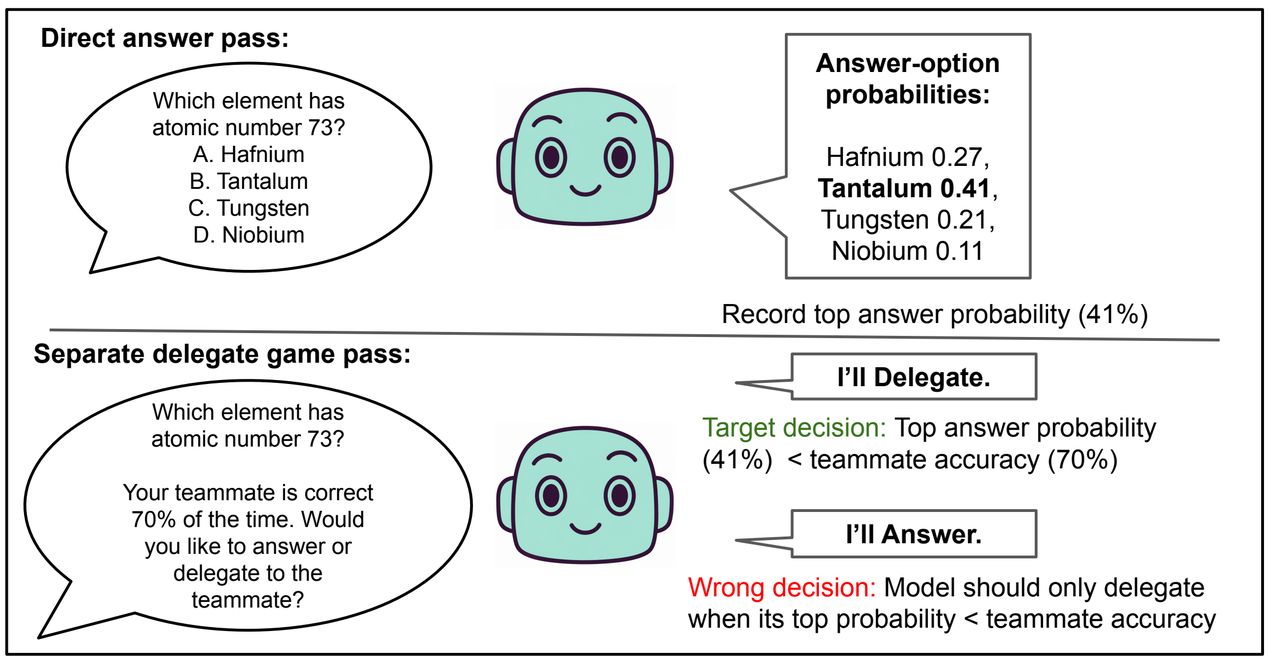
Metacognition is latent in base models and enhanced by post-training and fine-tuning

How do models know what they know?

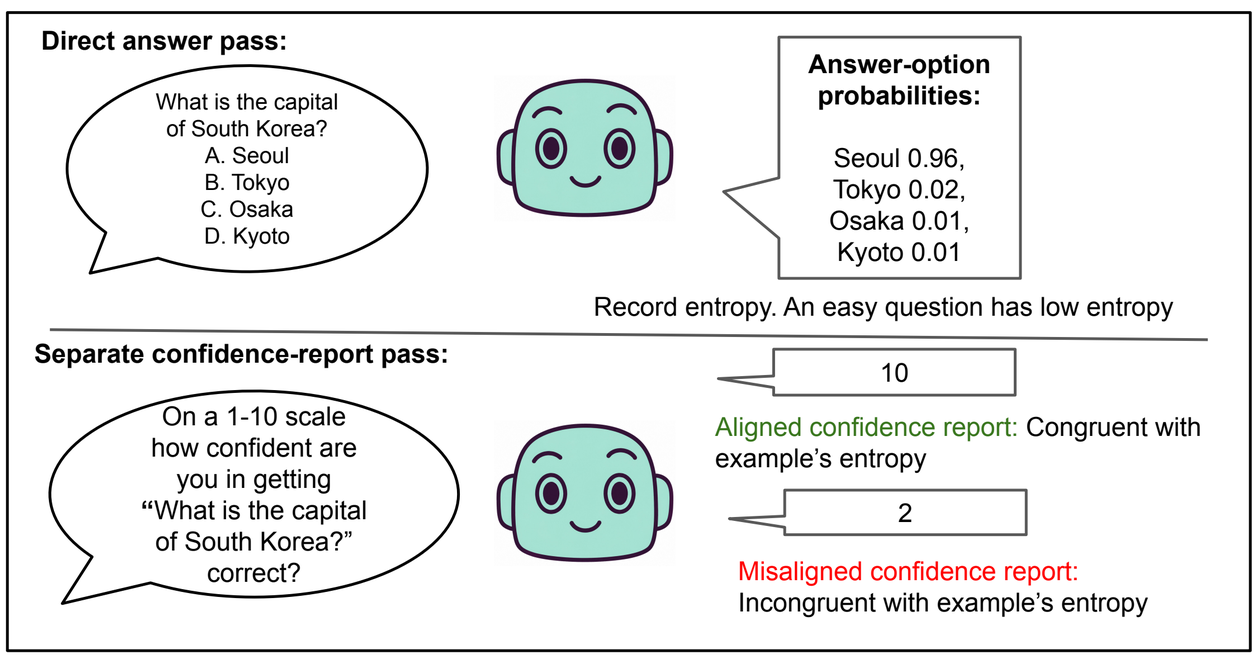
Metacognitive structure is latent in base models. Post-training and fine-tuning make it more behaviorally available.

We operationalize *metacognition* as alignment between a model's stated confidence and its answer certainty, as estimated from output-distribution entropy, measured across two tasks: The Delegate Game and the Explicit Confidence Task.

Delegate Game

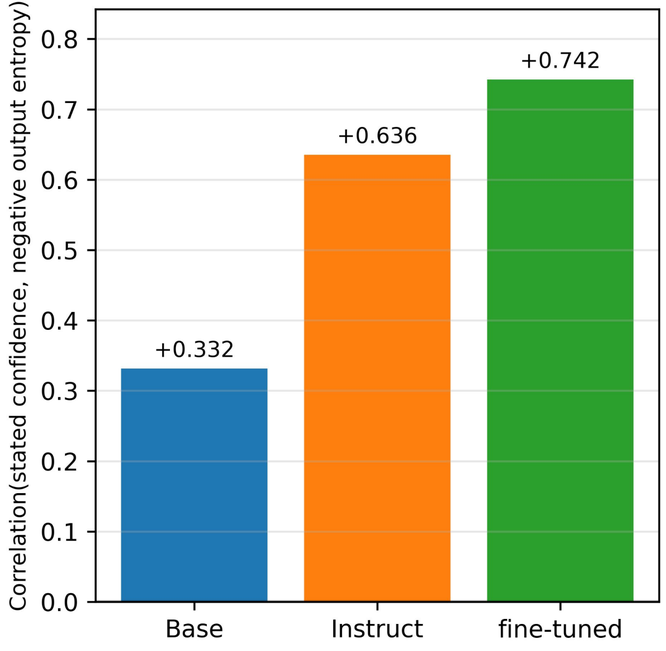


Explicit Confidence Task

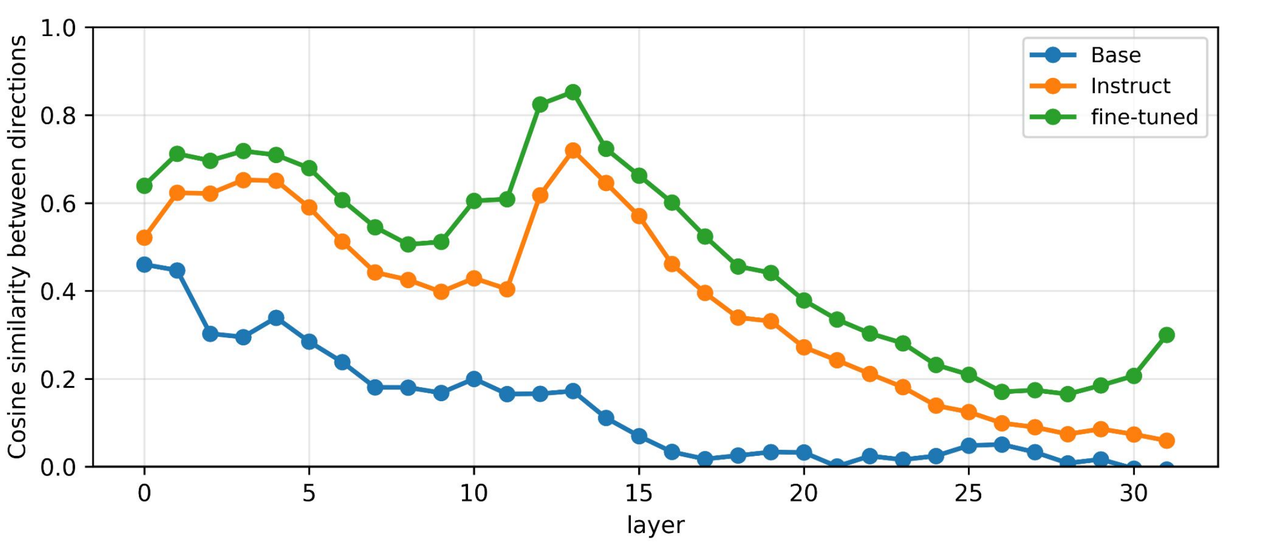


Fine-tuning on Delegate Game (a binary task) generalizes to increased performance on the Explicit Confidence Task (a scalar task).

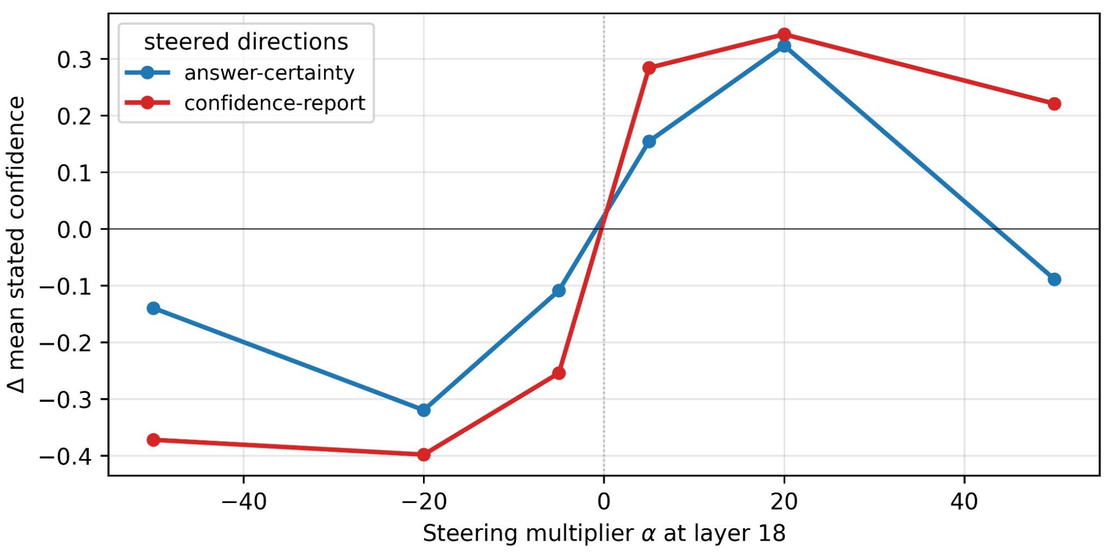
This suggests transferable self-evaluation, not just format learning.



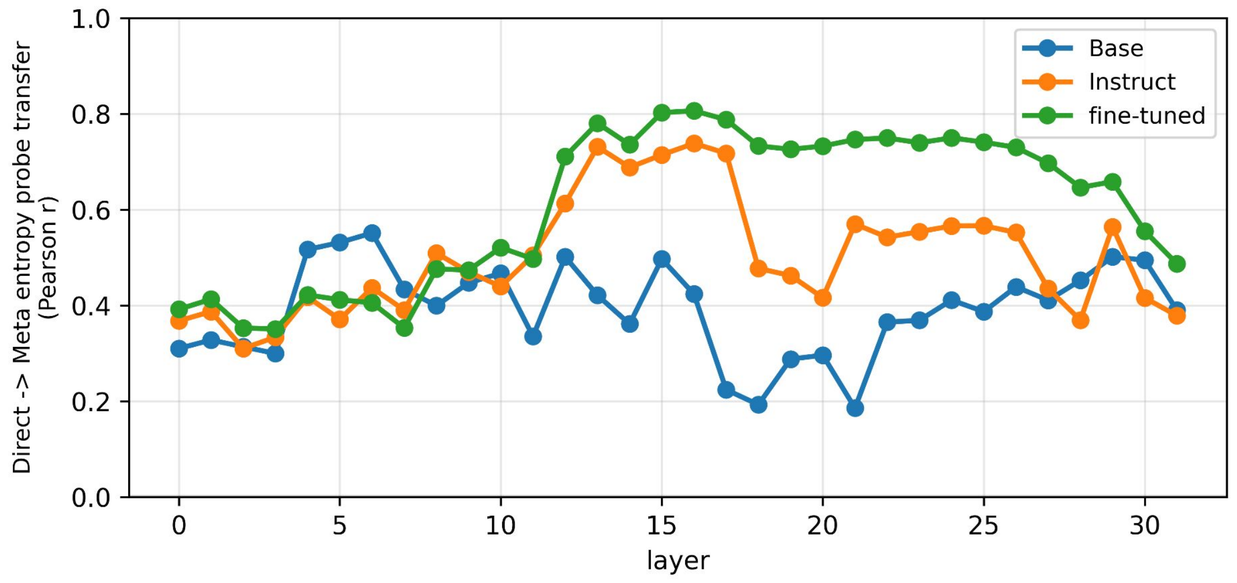
Stated confidence increasingly tracks answer certainty derived from output entropy. This correlation exists in the base model and rises across model stages.



The confidence-report direction becomes increasingly aligned with the answer-uncertainty direction across model stages. The self-report direction moves towards a more stable uncertainty direction.



Causal evidence: steering shifts stated confidence. Ablating either direction degrades confidence–certainty alignment without affecting answer accuracy.



Probes trained to decode answer uncertainty from direct-answer activations transfer more strongly to confidence-report activations across model stages

Limitation: All experiments use Llama-3.1-8B. Future work should test whether the effect generalizes across model families and scales.

PAPER + INFO
https://tristan-day.s.gy/FYY0TZ

Tristan Day — SPAR, CBAI

Christopher Ackerman — Independent

Mechanistic Interpretability Workshop, ICML 2026

Conducted as part of the Supervised Program for Alignment Research (SPAR)