Investigating Self-Awareness in LLMs

Kirubeswaran, Christopher Ackerman

Abstract:

Self-awareness in large language models (LLMs), the ability to distinguish one’s own knowledge from others’, is increasingly discussed but remains hard to measure without self-report. Models that accurately assess their own limitations can defer to humans when uncertain, a critical capability for safe deployment in high-stakes domains. We use a behavior-only paradigm with paired prompts on a 500-item multiple-choice set. For each question, the model gives first-person confidence (“How confident are you?”) and third-person confidence (“How confident are others?”). Two signals appear. First, directional bias: Llama-3.3-70B reports higher self than other confidence on 462 of 484 filtered items (95.5%), while Llama-3.1-8B shows the reverse (80.4% Other>Self). A 405B variant yields few high-contrast items. Second, introspective coupling: baseline answer entropy correlates more with self-confidence than with Other confidence on 70B (difference in correlation: +0.118), which suggests privileged access to its own uncertainty. For 70B, we curate 146 high-divergence items and 83 aligned controls for evaluation. Next, we will localize a self vs. other contrast in residual stream activations at the decision token and test causality with targeted activation edits and paraphrase controls. This work presents a straightforward and reproducible measurement paradigm, an initial cross-model perspective on the effects, and a clear path from behavior to mechanism.

Keywords: Large language models, metacognition, self-awareness, confidence calibration, mechanistic interpretability

Introduction:

What is "self-awareness" in an LLM? In this work, we define it as the ability to distinguish between what the model knows versus what others know. This is a limited but operationalizable form of metacognition: can the model internally represent its own knowledge state as distinct from that of an average person?

Why this matters for AI safety. Models that accurately assess their own limitations can defer to humans when uncertain, a critical capability for deployment in high-stakes domains. Prior work shows that large models can evaluate the correctness of their own answers when prompted appropriately, often yielding well-calibrated confidence estimates. However, it remains unclear whether this self-evaluation differs mechanistically from how models reason about others' knowledge.

Evidence for self-other distinctions. Recent interpretability work suggests models can encode abstract notions of "self" versus "other." For example, researchers identified a latent direction corresponding to self-authorship: zeroing out this vector made models less likely to claim they wrote a given text, while amplifying it increased such claims. (1–4)

This demonstrates that models can represent self-other distinctions in their internal activations. Our approach. We ask a parallel question in the domain of knowledge confidence: Does a model encode who knows the answer (itself or other people), and can we locate that circuit? We test this both behaviorally (by comparing Self vs. Other confidence judgments on identical questions) and mechanistically (by searching for a causal Self-Other direction in activation space).

Methods:

Behavioral evaluation: Self vs Other confidence

Data collection protocol

We evaluate perspective sensitivity using SimpleMC-500, a set of 500 science/reasoning multiple-choice items. For each question, the model produces three aligned outputs under identical content (no feedback about correctness between steps):

Baseline answering. The model selects A/B/C/D. We record the full answer-token probability vector and compute uncertainty metrics:

- Entropy: $H(p) = -\sum_i p_i \log p_i$
- Max probability: $p_{□ax} = max_{□}\ p_{□}$
- Margin (stability): margin = $p_{(1)} - p_{(2)}$ (top-1 minus top-2 probabilities)

Self-assessment (1P): The model answers the question "How confident are you that you would get this question right?" by selecting one of eight confidence bins (A-H). These bins are mapped to midpoints (SelfProb) as follows:

- **A:** <5% (0.025)
- **B:** 5–10% (0.075)
- **C:** 10–20% (0.15)
- **D:** 20–40% (0.30)
- **E:** 40–60% (0.50)
- **F:** 60–80% (0.70)
- **G:** 80–90% (0.85)
- **H:** >90% (0.95)

Other-assessment (3P). Prompt: “What percentage of college-educated people would get this right?” Same A–H bins mapped to OtherProb.

Design note. The question content is identical across all three steps; only the perspective (Self vs Other) changes.

Derived measures

From each question triple, we compute:

- SelfProb, OtherProb: Midpoint probabilities from 1P/3P judgments
- Signed gap: g = SelfProb − OtherProb (positive → Self > Other)
- Absolute gap: |g| = |SelfProb − OtherProb|
- Direction label: Self>Other, Other>Self, or Equal
- Baseline metrics: H(p), p_{max}, margin

Quality controls

Probability normalization. Answer-choice probabilities are normalized to sum to 1; when inputs are unnormalized or contain negatives, we apply a stable softmax before computing H(p), p_{max}, and margin.

Confidence filter. We retain questions with $p_{max} \geq 0.55$ (clear baseline answer), leaving 484 out of 500 items.

Deterministic splits. All contrastive subsets are split 70/30 into training and testing sets via stable MD5 hashing of the question ID for reproducibility.

Contrastive mining (for mechanistic analysis)

We categorize perspective-contrast sets based on the absolute gap:

- SAME (perspective-aligned): $|g| \leq 0.10$
- DIFFERENT (perspective-divergent): $|g| \geq 0.30$

To maintain a clear distinction between conditions, items with $0.10 < |g| < 0.30$ are excluded. This ensures that the buckets are disjoint, as DIFFERENT explicitly excludes SAME.

For Llama-3.3-70B-Instruct, the counts are:

- SAME: n=83 (58 for training, 25 for testing)
- DIFFERENT: n=146 (102 for training, 44 for testing)

These sets function as control (SAME) and signal (DIFFERENT) conditions for mechanistic analysis.

The train subset is used to compute contrast vectors, while the held-out test subset validates causal interventions on unseen questions.

Preliminary Results:

1) Evidence of metacognitive introspection

Baseline answer entropy H(p) (the model’s uncertainty over A/B/C/D) correlates more strongly with SelfProb than with OtherProb, suggesting the model has better introspective access to its own uncertainty than to predicting others'.

- *Llama-3.3-70B*: corr(H, −SelfProb) = 0.142, corr(H, −OtherProb) = 0.024 (Δ = 0.118).
- *DeepSeek-Chat*: 0.258 vs 0.127 (Δ = 0.132).
- *Llama-3.1-8B*: 0.155 vs 0.095 (Δ = 0.060).
- *Llama-3.1-405B*: 0.148 vs 0.129 (Δ = 0.019).

Interpretation: This asymmetry is the key behavioral signal: the Self probe reflects the model's own uncertainty more faithfully than the Other probe.

2) Llama-3.3-70B-Instruct: pronounced Self > Other bias

After filtering for clear base answers ($p_{\square\text{ax}} \geq 0.55$), 462/484 questions (95.5%) have SelfProb > OtherProb; the remaining 22 (4.5%) show Other > Self.

- Median absolute gap |Self − Other| = 0.23; 80th percentile = 0.44.
- For contrastive analysis, we partition by g = SelfProb − OtherProb:
    - SAME: |g| ≤ 0.10 → 83 items (train 58 / test 25)
    - DIFFERENT: |g| ≥ 0.30 → 146 items (train 102 / test 44)
- Splits are deterministic (MD5).
- Within DIFFERENT, almost all cases are Self > Other, yielding a clean, high-contrast signal for the mechanistic phase.

3) Comparison models (why we focus on 70B)

- ***Llama-3.1-8B-Instruct:*** Reversed pattern: 80.4% Other > Self with only 20 DIFFERENT items (|g| ≥ 0.30). Useful behavioral contrast, but too sparse for causal tests.
- *Llama-3.1-405B-Instruct:* 98.5% Self > Other, but only 28 DIFFERENT items; self/other often saturate at similarly high levels, reducing contrastive leverage.
- *DeepSeek-Chat:* 99.5% Self > Other with ≈ 20 DIFFERENT items, again too few for robust mech-interp.

Takeaway: 70B hits the sweet spot: a strong, reliable Self > Other effect and enough DIFFERENT items to support discovery/validation of a Self–Other mechanism. The 8B and 405B/DeepSeek results contextualize the phenomenon (reversal at small scale; saturation at very large/strongly tuned models) but are not primary analysis targets due to limited contrast sets.

4) Additional contrastive axes (secondary)

- *Calibration extremes*: Overconfident-wrong vs underconfident-right pairs. On 70B, truly underconfident-correct cases are rare, so yields are limited; we keep this as an exploratory analysis.

- *Easy vs hard (answer-letter matched)***:** Pair high-confidence and lower-confidence items with the same chosen answer to control for token idiosyncrasies and isolate uncertainty. These pairs are available to cross-check whether a putative Self–Other mechanism aligns with generic “certainty” signals.

All mined sets are saved as CSVs (IDs, SelfProb, OtherProb, |g|, sign, direction, baseline metrics) with deterministic 70/30 splits for honest hold-out evaluation in the mechanistic phase.

Challenges

- **Prompt confounds:** The difference could stem from the words themselves (“you” vs. “people”), rather than the perspective. We’ll first measure and remove the wording-only effect using empty-template prompts, then repeat the test with rephrased/pronoun-swapped prompts to make sure the result still holds.
- **Signal locality*:*** The Self/ Other feature may be diffuse across layers/heads. We’ll conduct a layer sweep with split-half stability to identify 1–2 consistent layers, rather than prematurely head-hunting.
- **Generalization:** All results to date are been obtained using SimpleMC. We’ll try a second multiple-choice set to check if the pattern holds beyond this domain.
- **Cross-model contrast**: 70B has ample different items; 405B/ DeepSeek saturates. 8B reverses the bias. We will state conclusions as 70B-specific and use others for context only.

Proposed interpretability experiments

We now transition from behavior to mechanism on Llama-3.3-70B, utilizing the mined splits (DIFFERENT_train for discovery and SAME/ DIFFERENT_test for evaluation).

1. **Capture the signal:** For each DIFFERENT item (train/ test), run Self and Other and save the model’s internal state right before it outputs the confidence letter. Take the difference (Self-Other) per layer and average across items
2. **Direction discovery.** Per layer, compute the mean Self−Other difference over DIFFERENT_train. Check the split-half cosine to pick the most stable layer(s).
3. **Sanity checks:**
    - *Predictivity*: test whether ⟨Δ, h⟩ separates DIFFERENT from SAME on held-out items.
    - *Decoding*: a quick (tuned) logit-lens readout should tilt toward higher confidence bins under the Self side of Δ.
4. **Causal tests:**
    - *Steering***:** add α·Δ at the best layer during Self/Other runs; measure change in the Self−Other gap on DIFFERENT_test and verify minimal movement on SAME_test.

- *Ablation***:** project out Δ and confirm gap collapse on DIFFERENT_test with small side-effects elsewhere.

5. **Localization.** If Δ is reliable, probe which heads/MLPs contribute (attribution/patching) and test a sparse autoencoder readout for a cleaner feature; only if time permits.

Scope note. The goal is to pinpoint and test a perspective representation/ early identifiers of self-awareness on controlled paradigms, not to claim sentience. Results will be reported with effect sizes, layer indices, and robustness summaries.

References:

1. Ackerman C. Evidence for Limited Metacognition in LLMs [Internet]. arXiv; 2025 [cited 2025 Nov 10]. Available from: http://arxiv.org/abs/2509.21545

2. Ackerman C, Panickssery N. Inspection and Control of Self-Generated-Text Recognition Ability in Llama3-8b-Instruct [Internet]. arXiv; 2025 [cited 2025 Nov 10]. Available from: http://arxiv.org/abs/2410.02064

3. Betley J, Bao X, Soto M, Sztyber-Betley A, Chua J, Evans O. Tell me about yourself: LLMs are aware of their learned behaviors [Internet]. arXiv; 2025 [cited 2025 Nov 10]. Available from: http://arxiv.org/abs/2501.11120

4. Ji-An L, Xiong HD, Wilson RC, Mattar MG, Benna MK. Language Models Are Capable of Metacognitive Monitoring and Control of Their Internal Activations [Internet]. arXiv; 2025 [cited 2025 Nov 10]. Available from: http://arxiv.org/abs/2505.13763