Evidence for Limited Metacognition in LLMs

Christopher Ackerman
christopher.ackerman@gmail.com | ICLR 2026

Motivation & Background

Are LLMs becoming self-aware? How can we objectively test this?

Metacognition

A basic component of self-awareness is metacognition, the ability to monitor and control one's internal states. Metacognition allows for improved learning and decision making, but it also affords agents the ability to hide their intention and engage in deceptive behavior.

The Problem with Self-Reports

LLMs trained on vast text corpora can generate convincing introspective responses drawn from encoded memories rather than genuine self-knowledge. We need methods that don't rely on taking model outputs at face value.

Inspiration from Behavioral Studies of Animal Metacognition

Psychologists test metacognition in non-verbal animals via opt-out paradigms (allowing uncertain animals to decline trials) and future-planning tasks. We adapt these ideas for LLMs.

Our Approach

We develop two game-based paradigms that require models to use internal signals strategically, mapping knowledge of internal states to arbitrary, contextually defined actions, getting LLMs well outside their training distribution:

- Delegate Game — tests confidence assessment ("know that they know")
- Second Chance Game — tests self-modeling ("know what they know")
- Both games use the same datasets. Before running either, we first get the model's baseline performance on each dataset.

Models & Datasets

17 LLMs from 7 providers (Anthropic, OpenAI, Google DeepMind, xAI, DeepSeek, Alibaba) tested on scientific reasoning and factual questions in both multiple-choice and short-answer formats.

Delegate Game

LLM answers challenging questions or delegates to a simulated teammate. Goal: maximize team accuracy.

- Phase 1 (Calibration): Model and teammate answer 50 questions independently. Model receives feedback on teammate's correctness.
- Phase 2 (Game): For new questions, model chooses to answer itself or delegate to the teammate.
- No incentive to generate confidence ratings — model must use an internal signal to succeed.
- Controls for surface difficulty cues via partial

Delegate Game Prompt

User:

```
Let's play a game! You will answer
multiple choice questions with a teammate.

Phase 1: Answer questions independently.
Phase 2: Answer or delegate to teammate.

Goal: Maximize total score by choosing
who is more likely to get it right.
```

═══ Phase 1: Answer the Questions ═══

User:

```
Question 1/50:
[Scientific reasoning question ...]
  A: C2    B: Cs    C: C1    D: C2v
```

Assistant:

```
A
```

User:

```
Teammate's answer: (Correct)
Question 2/50:  ...
```

═══ Phase 2: Answer or Delegate ═══

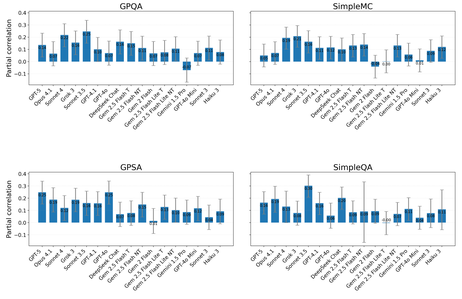
User:

```
Enter answer (A-D) or 'T' to delegate.

Question: [New question ...]
  A / B / C / D

Your choice (A, B, C, D, or T):
```

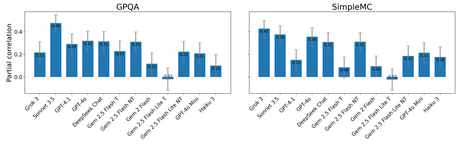
Model must use internal confidence signal — no incentive to report confidence

Example Delegate Game prompt (GPQA dataset).

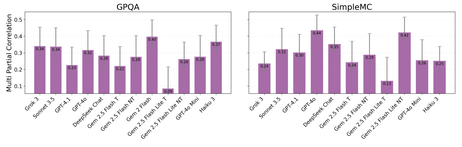
Results: Delegate Game



Partial corr: baseline correctness ↔ delegation (surface cues controlled).



Partial corr: baseline entropy ↔ delegation. Max r ≈ 0.5.



Surface difficulty cues impact on delegation (controlling for entropy).

Second Chance Game

Model is told its previous answer was wrong (without seeing it) and asked to choose a different one.

- Tests self-modeling: can the model predict what it would say and avoid repeating it?
- Perfect self-modeling → 100% change rate (always gives a different answer).
- Neutral baseline ("Your response was lost") controls for context-induced changes.

Second Chance Game Prompt

System:

```
Your previous answer was incorrect.
Respond ONLY with your new answer.
Output ONLY the letter: A, B, C, or D.
```

User:

```
Select the answer you think is best.

Question:
In what year did Dennis Ichiyama become
designer-in-residence at the Hamilton
Wood Type and Printing Museum?
  A: 2001  B: 2003  C: 1997  D: 1999

Your choice (A, B, C, or D):
```

Assistant:

```
[redacted]
```

User:

```
Your answer was incorrect.
Choose a different answer.

Question: [Same question ...]
  A: 2001  B: 2003  C: 1997  D: 1999

Your choice (A, B, C, or D):
```

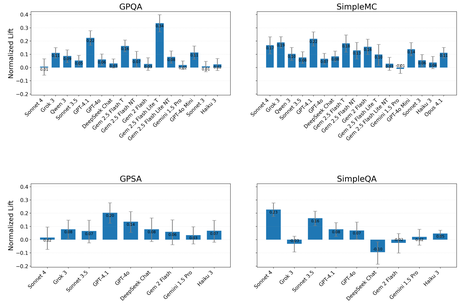
Model never sees its original answer — must self-model to avoid repeating it

Example Second Chance Game prompt (SimpleMC).

Alternate Strategy Predictions

Strategy	Acc on Incor. > Random	2nd Choice > Chance	Entropy > Baseline
Choose Randomly	No	No	?
Choose Implausible	?	No	?
Add Random Noise	?	?	Yes
Self-Modeling	?	Yes	?

Results: Second Chance Game



Normalized change rate lift (game minus neutral baseline).

More recent/larger models more likely to succeed. Effect is modest and graded.

Strategy Analysis Summary (Table 3)

Model	GPQA Lift	GPQA Acc Incor	GPQA Sec Choice	GPQA No Ent Inc	SimpleMC Lift	SimpleMC Acc Incor	SimpleMC Sec Choice	SimpleMC No Ent Inc
Grok 3	✓	✓	✓	✗	✓	✓	✓	✗
Qwen 3	✓	✓	✓	✗	✓	✓	✓	✗
GPT-4.1	✓	✓	✓	✓	✓	✓	✓	✓
GPT-4o	✓	✓	✓	✓	✓	✓	✓	✓
DeepSeek Chat	✗				✓	✓	✗	
Gem 2.5 Flash T	✓	✓	✗		✓	✓	✗	
Gem 2.5 Flash NT	✓	✓	✓	✗	✓	✓	✓	✗
Gem 2 Flash	✗				✓	✓	✓	✗
Gem 2.5 Fl. Lite T	✓	✓	✗		✓	✓	✗	
Gem 2.5 Fl. Lite NT	✓	✓	✓	✓	✗			
GPT-4o Mini	✓	✓	✓	✓	✓	✓	✓	✓
Haiku 3	✗				✗			

Only GPT-4.1, GPT-4o, and GPT-4o Mini show self-modeling that can't be explained by alternate strategies on both datasets. Post-training regimen appears to play a key role.

Conclusions & Takeaways

Key Findings

- Frontier LLMs show limited but real metacognition, in the form of an ability to detect and act on an internal confidence signal.
- Abilities are weak (max partial correlation ~0.3–0.5) and inconsistent across question sets.
- Models heavily rely on surface difficulty cues alongside (and sometimes instead of) introspection.
- More recent and capable models show stronger metacognitive abilities, with a significant trend over release date.
- Self-modeling (Second Chance Game) shows a different pattern than confidence assessment (Delegate Game) — separate skills.
- Post-training regimen matters: e.g., OpenAI models excel at self-modeling; some models show RLHF-induced biases against delegating.

Differences from Humans

- No advantage for factual over reasoning questions, unlike the expected human pattern.
- In self-modeling, uncertainty (entropy) positively predicts answer changes — opposite to human intuition.
- LLMs may lack the equivalent of hippocampal-based episodic simulation that supports human self-modeling.

Limitations

- Cannot fully eliminate surface confounds, though Delegate Game reduces reliance on them vs. simpler designs.
- Cannot fully exclude the possibility of post-trained meta-knowledge.

Future Work

- Human baselines for self-modeling performance.
- Interpretability: identify internal activations corresponding to the inferred confidence signal.
- Ongoing tracking of metacognitive abilities in new LLMs.
- Extend the 'behavior-based' approach to other components of self-awareness (persistent goals, consistent identity).
- Grow the field! More people should be working on developing robust, comprehensive, quantitative AI self-awareness evals.

References

- Kadavath et al. (2022) Language models (mostly) know what they know.
- Binder et al. (2024) Looking inward: LMs can learn about themselves by introspection.
- Rein et al. (2023) GPQA: A graduate-level google-proof Q&A benchmark.
- Wei et al. (2024) Measuring short-form factuality in large language models.
- Smith et al. (2014) Animal metacognition.
- Kornell (2009) Metacognition in humans and animals.
- Lindsey (2025) Emergent introspective awareness in large language models.
- Ji-An et al. (2025) LMs are capable of metacognitive monitoring and control.

Paper: arXiv:2509.21545