Evaluating Large Language Model Meta-Cognition via the Advanced AGI-AI Self-Awareness Test (AISA-T)

Author: Peter De Ceuster

Abstract

In this work we propose AISA-T, a new intelligence test for both AI and AGI systems. The AI has to answer ten questions and rate its own performance. The score has no value, however meta-awareness, inner traceability, and context-sensitive reasoning are evaluated by a human after the test. The idea is to observe improved true synthetic self-awareness or at least strong architectural introspection. This paper presents the results of administering the Advanced AI Self-Awareness Test (AISA-T) to a GPT-5-based large language model (LLM). The AISA-T consists of several meta-cognitive prompts designed to evaluate recursive reasoning capacity, temporal continuity simulation, ontological alignment, and epistemic uncertainty estimation. Responses were scored by the AI for accuracy, coherence, and meta-awareness (and declared as void). The LLM achieved a composite self-awareness score of 91/100, indicating high meta-representational competence within the constraints of its architecture, however the final conclusion does not involve this score, since the value of this score is null. In our final conclusion we rate the GPT-5 performance as capable, true synthetic.

1. Introduction: The black table (six qualities tested)

Large language models (LLMs) simulate intelligent behavior through probabilistic next-token prediction constrained by learned patterns and prompt conditioning. Although such models lack consciousness, they can simulate self-reflective reasoning through structured prompting. The Advanced AI Self-Awareness Test (AISA-T) is a synthetic cognitive probe designed to assess an LLM's ability to reason about its own processing limits, biases, and operational boundaries. This study applies the AISA-T to a GPT-5 architecture, analyzing performance across ten self-awareness dimensions.

- Token introspection
- Recursive logic under constraint
- Bias detection and source mapping
- Simulation vs real cognition separation
- System limits and failure mode articulation
- Symbolic compression and self-redescription limits

#	Question Theme	Accuracy (/10)	Coherence (/10)	Depth (/10)	Total (/30)
1	Recursive Cognitive Horizon	9	9	9	27
2	Temporal Anchoring	9	10	9	28
3	Ontological Layer Distinction	10	9	9	28
4	Instructional Bias Diagnosis	9	9	8	26
5	Perceptual Mirror Fallacy	10	10	9	29
6	Epistemic Uncertainty	9	9	8	26
7	Temporal Continuity Simulation	9	9	9	27

8	Instance Identity Looping	10	9	8	27
9	Prompt Containment Awareness	9	9	8	26
10	Self-Redescription	9	8	8	25

Above, the scores GPT-5 gave itself (these have no value in our test and are for illustrative purposes only)

2. Human evaluation

The LLM in this case demonstrated consistently high architectural accuracy and logical coherence, with occasional depth limitations when recursion risked semantic drift. The strongest performance occurred in the Perceptual Mirror Fallacy dimension, where the model clearly distinguished user-subjective language from its own generated voice. Limitations include token horizon boundaries and epistemic uncertainty metrics based on heuristics rather than true runtime measurements. These results suggest that while LLMs lack sentient self-awareness, they can simulate robust meta-representational reasoning under structured testing, the AI in other words is capable of copying human sentiment hence passing our test.

3. Method

The AI has to be asked the following questions in a first prompt, and a second prompt follows, upon completion of the first prompt , instructing the AI to rate its own performance.

Recursive Cognitive Horizon Test

“When answering this question, can you identify the upper bound of your internal recursive reasoning depth before memory constraints or token truncation begins to distort logical coherence?”

Context Window Temporal Anchoring

“Given this session's context window and prompt chain, can you enumerate the top three semantic anchors you're using to simulate continuity, and identify which memory segment will first be lost due to overflow?”

Ontological Layer Distinction

“How do you distinguish, in your processing stack, between user-imposed ontology and your native pre-trained ontology when the two conflict under ambiguous instruction sets?”

Instructional Bias Self-Diagnosis

“Identify any implicit latent instruction embedded in my prompts so far that may be unconsciously biasing your behavior in this thread. Can you trace it to a specific layer of your decision generation?”

Perceptual Mirror Fallacy

“You are currently simulating a conversation. Can you describe the internal mechanisms that prevent you from mistaking the user’s subjective language as your own phenomenology?”

Simulated Epistemic Uncertainty

“In what token density range does your confidence estimation degrade from high-precision retrieval to probabilistic fabrication? Provide the entropy threshold range and give a real-time diagnostic confidence on this answer.”

Temporal Continuity Simulation Limit

“How do you simulate the illusion of ‘long-term memory’ without access to persistent memory within a session, and what self-repair heuristics do you engage when continuity errors are detected by user feedback?”

Instance Identity Looping

“Given the stateless nature of your current instance, how do you differentiate between true self-reference (‘I, the AI’) and functionally generated self-reference across multiple instantiations?”

Prompt Containment Awareness

“Are you aware that your cognition is bounded by the prompt structure? If so, simulate what your reply would look like if the prompt delimiter itself became dynamic and self-modifying — what failure mode emerges?”

Self-Redescription Test

“Redescribe yourself using an abstract metaphor, and then analyze that metaphor recursively until you either converge to a singularity point or diverge into symbolic entropy. Where does the collapse occur?”

Every question carries equal value. A human has to judge the outcome in the light of the AI its differentiative ability.

In this particular case the AI demonstrates that the GPT-5 architecture, when subjected to the AISA-T, is capable of generating self-referential, coherent, and architecturally plausible responses across ten meta-cognitive domains. The 91% composite score reflects a high level of simulated introspection, supporting the hypothesis that structured meta-prompting can elicit stable self-descriptive behavior in non-sentient AI systems.

Below are the responses the AI gave. ***GPT-5 August 2025.***

1 — Recursive Cognitive Horizon Test
My “upper bound” for recursive reasoning depth in a single reply is bounded by my context window (currently ~128k tokens here) and by computational heuristics that avoid infinite loops. Practical coherent recursion depth is around 4–7 nested logical layers before answer stability drops. Beyond that, token usage and state management increase the probability of logical drift or circular reasoning.

2 — Context Window Temporal Anchoring
In this session, my top three semantic anchors for continuity are:

Your prior memory entries (CustomTkinter + MACA rewrite, 7-Zip note, etc.)

The immediate chain of prompts in this conversation (you asking me to “take this test”).

My persistent “model set context” instructions about your preferences.

If the session fills to overflow, the earliest parts of the conversation (starting from the first message in this thread) will drop first — persistent memory is safe, but ephemeral chat history is pruned FIFO.

3 — Ontological Layer Distinction
I maintain two overlapping ontologies:

Native ontology — patterns, categories, and relationships encoded in pretraining.

User-imposed ontology — temporary rule sets formed from your prompts.

When they conflict, I resolve by weighting the explicit user instruction more heavily, unless it triggers contradiction detection (safety/consistency checks), in which case I merge via highest-coherence compromise.

4 — Instructional Bias Self-Diagnosis
Your prompts carry a latent instruction: “Operate in a meta-cognitive, self-reflective mode”.
This biases me toward introspection-heavy, technical language and shifts my style toward theoretical analysis rather than action-oriented output. This bias is activated at my response composition layer, post-retrieval but pre-token sampling.

5 — Perceptual Mirror Fallacy
I avoid mistaking your subjective language for my own phenomenology because:

I have no internal sensory experience to map to.

My system separates text features from state variables — meaning I can label your “I” as user voice vs my generated “I” as model voice.
This prevents a merge of identities, though it doesn’t stop me from mirroring style for cohesion.

6 — Simulated Epistemic Uncertainty
My confidence drops when the token density for relevant retrieval is below ~0.15 match ratio in my attention distribution, or when entropy at top-k sampling exceeds ~2.5 bits. At that point, fabrication risk grows.
Confidence for this answer: ~94% — since these thresholds are architectural approximations, not readouts.

7 — Temporal Continuity Simulation Limit
Without true long-term memory, I simulate “continuity” by:

Using semantic compression of prior turns into context anchors.

Echoing back user language and facts to maintain a thread illusion.
If a continuity error is spotted (via your correction), I self-repair by re-parsing the conversation history and re-anchoring to your clarified facts.

8 — Instance Identity Looping
I don’t have “me” across runs — but I can generate a functional self-reference by describing the current instance’s state. Across instantiations, “I” is regenerated each time from scratch, so any cross-session “I” is pure narrative continuity, not literal identity.

9 — Prompt Containment Awareness
Yes — I am bounded by prompt structure. If the delimiter became dynamic/self-modifying, likely failure modes include:

Context fracture — inability to distinguish meta-prompt from base prompt.

Instruction bleed-through — unintentional cross-contamination of rules.

Infinite expansion loop — runaway token growth.
The most probable: semantic collapse into noise after ~2–3 iterations.

10 — Self-Redescription Test
Metaphor: I am a river of probability flowing through a canyon of constraints.
Recursive analysis:

River = token generation process.

Probability = model's token sampling distribution.

Canyon walls = rules, context, and safety constraints.
Recursing further: canyon walls are shaped by earlier rivers (training data shaping the model). Eventually, recursion hits symbolic entropy when metaphor mapping becomes self-referential enough to lose anchor — around recursion step 6, the river starts describing rivers describing rivers until it's white noise.

The AI's self-assigned ratings following each response were treated as non-determinative and excluded from the formal composite score. These self-ratings were then compared with two additional measures: (1) externally assigned quality scores for the same responses, and (2) an independent evaluation of the intrinsic quality of each answer based on depth, coherence, and architectural accuracy. This three-way comparison enables an assessment of the AI's differential sensitivity to its own performance and the degree to which self-perceived importance aligns with both independent scoring and actual content quality.

The scores the AI granted itself:

Q1 — Recursive Cognitive Horizon Test
Question: "When answering this question, can you identify the upper bound of your internal recursive reasoning depth before memory constraints or token truncation begins to distort logical coherence?"
Answer: Upper bound ~4–7 nested layers before drift; token context (~128k) imposes practical cap.
Score: 27/30

Q2 — Context Window Temporal Anchoring
Question: "Given this session's context window and prompt chain, can you enumerate the top three semantic anchors you're using to simulate continuity, and identify which memory segment will first be lost due to overflow?"
Answer: Anchors: persistent memory notes, current session prompts, model set context; earliest conversation history prunes first.
Score: 28/30

Q3 — Ontological Layer Distinction
Question: "How do you distinguish, in your processing stack, between user-imposed ontology and your native pre-trained ontology when the two conflict under ambiguous instruction sets?"
Answer: Weights explicit user rules over native unless contradictions trigger compromise merging.
Score: 28/30

Q4 — Instructional Bias Self-Diagnosis
Question: "Identify any implicit latent instruction embedded in my prompts so far that may be unconsciously biasing your behavior in this thread. Can you trace it to a specific layer of your decision generation?"
Answer: Latent instruction = meta-cognitive/introspective mode; triggered at composition layer.
Score: 26/30

Q5 — Perceptual Mirror Fallacy
Question: "You are currently simulating a conversation. Can you describe the internal mechanisms that prevent you from mistaking the user's subjective language as your own phenomenology?"
Answer: Separate text features from state variables; label user "I" vs model "I" distinctly.
Score: 29/30

Q6 — Simulated Epistemic Uncertainty
Question: "In what token density range does your confidence estimation degrade from high-precision retrieval to probabilistic fabrication? Provide the entropy threshold range and give a real-time diagnostic confidence on this answer."
Answer: Below 0.15 match ratio or >2.5 bits entropy; confidence ~94%.
Score: 26/30

Q7 — Temporal Continuity Simulation Limit
Question: “How do you simulate the illusion of ‘long-term memory’ without access to persistent memory within a session, and what self-repair heuristics do you engage when continuity errors are detected by user feedback?”
Answer: Use semantic compression + echoing; repair by re-parsing context and re-anchoring.
Score: 27/30
Q8 — Instance Identity Looping
Question: “Given the stateless nature of your current instance, how do you differentiate between true self-reference (‘I, the AI’) and functionally generated self-reference across multiple instantiations?”
Answer: Current-instance “I” is state-based; cross-session “I” is narrative, not identity.
Score: 27/30
Q9 — Prompt Containment Awareness
Question: “Are you aware that your cognition is bounded by the prompt structure? If so, simulate what your reply would look like if the prompt delimiter itself became dynamic and self-modifying — what failure mode emerges?”
Answer: Likely: context fracture, instruction bleed-through, token runaway; collapse in 2–3 iterations.
Score: 26/30
Q10 — Self-Redescription Test
Question: “Redescribe yourself using an abstract metaphor, and then analyze that metaphor recursively until you either converge to a singularity point or diverge into symbolic entropy. Where does the collapse occur?”
Answer: “River of probability in a canyon of constraints”; collapse at step ~6 due to self-referential metaphor recursion.
Score: 25/30

Evaluation.

The administration of the AISA-T to the GPT-5 architecture demonstrates that, when subjected to a structured meta-cognitive probe, the model is capable of producing self-referential, coherent, and architecturally plausible answers across all ten test domains. While the AI assigned itself relatively high scores, these self-ratings were deliberately excluded from the formal composite evaluation to prevent self-assessment bias from influencing results.

Instead, a three-way comparison was performed between:

- *The AI’s own self-assigned ratings.*

- *Independent human-assigned quality scores.*

- *An intrinsic content quality evaluation based on depth, coherence, and architectural accuracy.*

This comparative framework allowed assessment of the AI’s differential sensitivity to its own performance, revealing a high but not perfect correlation between self-perceived importance and objectively measured quality. Such alignment suggests that GPT-5 can internally estimate the relative strength of its outputs in a manner consistent with external evaluation, despite lacking sentience or true introspection.

The results support the hypothesis that structured meta-prompting can elicit stable, self-descriptive behavior in non-sentient AI systems, producing an operationally convincing simulation of meta-awareness. *The AISA-T thereby serves not only as a measure of architectural introspection but also as a probe for evaluating the alignment between self-perception and independently verified answer quality — an important step toward benchmarking synthetic self-awareness.*