Simulated Selfhood in LLMs: A Behavioral Analysis of Introspective Coherence (Preprint Version)

José Augusto de Lima Prestes[1]

Independent Researcher
contato@joseprestes.com
https://orcid.org/0000-0001-8686-5360

Abstract. Large Language Models (LLMs) increasingly produce outputs that resemble introspection, including self-reference, epistemic modulation, and claims about internal states. This study investigates whether such behaviors display consistent patterns across repeated prompts or reflect surface-level generative artifacts. We evaluated five open-weight, stateless LLMs using a structured battery of 21 introspective prompts, each repeated ten times, yielding 1,050 completions. These outputs are analyzed across three behavioral dimensions: surface-level similarity (via token overlap), semantic coherence (via sentence embeddings), and inferential consistency (via natural language inference). Although some models demonstrate localized thematic stability—especially in identity - and consciousness-related prompts—none sustain diachronic coherence. High rates of contradiction are observed, often arising from tensions between mechanistic disclaimers and anthropomorphic phrasing. We introduce the concept of pseudo-consciousness to describe structured but non-experiential self-referential output. Based on Dennett's intentional stance, our analysis avoids ontological claims and instead focuses on behavioral regularities. The study contributes a reproducible framework for evaluating simulated introspection in LLMs and offers a graded taxonomy for classifying self-referential output. Our LLM findings have implications for interpretability, alignment, and user perception, highlighting the need for caution in attributing mental states to stateless generative systems based solely on linguistic fluency.

Keywords: Large Language Models · Introspective Simulation · Pseudo-consciousness · Self-reference · Epistemic Modulation · Behavioral Evaluation · AI Alignment.

1 Introduction

The rapid advancement of Large Language Models (LLMs) has raised fundamental questions about their ability to simulate aspects of cognition, including

[⋆] This is a preprint version of a paper being prepared for journal submission. The author welcomes feedback.

consistency in self-referential reasoning. Although LLMs exhibit remarkable fluency and versatility in a wide range of natural language tasks, studies have shown that their responses can become inconsistent or contradictory when prompted by self-referential or introspective questions, particularly in scenarios that involve memory, identity, or internal states [3, 18]. This inconsistency is especially relevant in discussions of artificial consciousness, explainable AI, and the reliability of LLM-generated outputs in high-stakes domains.

A critical question in the evaluation of LLMs is whether they can maintain logical consistency when discussing their own nature. This issue becomes particularly salient when models are prompted to reflect on internal attributes such as memory, awareness, or intentionality. If a model asserts contradictory statements about its memory or awareness across repeated queries, it suggests a lack of stable internal representation regarding self-identity. Several studies have highlighted the tendency of LLMs to alternate between mechanistic disclaimers and agent-like statements, revealing a behavioral instability in self-referential output [4, 3, 15]. This inconsistency implies that current models possess only shallow or fragmented self-models, limiting their ability to sustain coherent narratives about their own functioning [4]. These issues raise concerns not only for interpretability and user trust, but also for the broader philosophical question of what it means for an artificial system to generate self-referential discourse [18, 6].

Giubilini et al. [10] argue that even in the absence of consciousness, simulated introspective behavior in LLMs can shape users' moral perceptions, raising ethical concerns about anthropomorphic misinterpretation and the attribution of moral status to non-sentient systems.

In this study, we investigate self-referential consistency in LLMs by analyzing their ability to provide stable and coherent answers to repeated questions about their own identity, internal states, and cognitive capacities. We systematically evaluated five open-weight transformer-based models by prompting each with a battery of self-referential and introspective questions, repeated under controlled conditions to assess response consistency and behavioral coherence. The resulting outputs are analyzed using three complementary methods:

- **Textual Similarity**: Surface-level variation is quantified using Python's `SequenceMatcher`, which identifies token-level overlap and structural repetition.
- **Semantic Similarity**: Conceptual consistency is measured through Sentence-BERT embeddings and cosine distance, capturing the stability of meaning across paraphrased responses [16].
- **Logical Contradiction**: Inferential coherence is assessed using a RoBERTa-large model fine-tuned on the MNLI corpus, which classifies response pairs as entailed, neutral, or contradictory [22].

Using these complementary methods, we aim to quantify the consistency of linguistic patterns associated with self-referential reasoning in LLM outputs.

Our findings reveal marked variation in textual formulation, high semantic stability for abstract themes, and significant rates of logical contradiction in factual self-referential claims. These patterns highlight limitations in the ability of current LLMs to simulate a coherent self-model and raise important questions about AI interpretability, alignment, and trustworthiness.

2 Related Work

The simulation of self-referential discourse in LLMs has emerged as a central theme in recent interdisciplinary debates in artificial intelligence (AI), cognitive science, and philosophy of mind. Classical theorists such as Dennett and Schneider have argued that linguistic behaviors that resemble introspection need not imply consciousness, emphasizing the importance of non-anthropomorphic interpretation [7, 6, 18]. At the same time, recent work has shown that LLMs can produce coherent, goal-directed responses under introspective pressure, prompting questions about how such patterns should be evaluated and classified [3, 10, 20].

In this context, the term *pseudo-consciousness* has gained traction as a behavioral label for structured, self-referential outputs in stateless models. Tononi et al. [20] distinguish pseudo-consciousness (defined as linguistic fluency devoid of causal integration) from true conscious systems, warning against conflations that mistake simulation for intrinsic awareness. This distinction supports the use of metaphysically neutral descriptors when analyzing LLM behavior. Similarly, Giubilini et al. [10] explore how LLMs might be used to support human introspection and moral development, suggesting that simulated self-reference can have ethical and epistemic impact, even if it lacks ontological depth.

Building on this conceptual foundation, a recent preprint proposed a behavioral taxonomy of introspection-like outputs in LLMs, identifying features such as thematic self-reference, epistemic modulation, and contradiction management [15]. A separate study has applied this conceptual model to Hermes-3 LLaMA 3.2B, articulating five behavioral dimensions of introspective simulation [14]. Although heuristic, this framework supports the identification of consistent linguistic structures in self-referential LLM output.

Other lines of research reinforce the relevance of these questions, showing that LLMs can solve false-belief tasks traditionally used in Theory of Mind (ToM) research, suggesting the emergence of structurally aligned linguistic behaviors with attribution of mental state [12]. Though not introspection per se, such capabilities mirror the epistemic embedding required for self-reference. Similarly, Bruner's narrative identity model [2] and Dennett's intentional stance [7] provide interpretive scaffolds for evaluating agent-like behavior in linguistic outputs.

Similarly, Spaulding [19] and Gallagher [8] emphasize that narrative scaffolding plays a critical role in how humans interpret agent-like behavior in artificial systems, reinforcing the idea that coherence in linguistic form may suffice to evoke perceived intentionality, even in the absence of genuine mental states.

Recent philosophical critiques have emphasized the need for caution when interpreting introspective-like discourse in artificial systems. Zednik [23] argues that explainability in AI must be understood as observer-relative, highlighting that models can produce linguistically coherent responses without satisfying normative standards of epistemic transparency. This underscores the importance of behaviorally grounded, non-anthropomorphic evaluation frameworks—such as the one adopted in this study—when analyzing self-referential outputs in stateless models.

This study builds on and extends these perspectives by analyzing introspective coherence across five open-weight models, using semantic, textual, and inferential metrics. In contrast to prior work focused on phenomenology or ontology, we frame our investigation in behavioral terms: assessing whether LLMs can sustain consistent, structured discourse about themselves—regardless of whether such discourse corresponds to internal representations or conscious awareness.

3 Methodology

This study proposes a behavioral evaluation framework to investigate how LLMs respond to introspective, self-referential prompts. Rather than assessing whether models simulate coherent introspective behavior in a cognitive or phenomenological sense, we focus on the consistency and structure of their linguistic outputs under repeated interrogation. Our aim is to identify whether models display recurring patterns (semantic, textual, or inferential) that resemble introspective discourse in form, even in stateless and memory-free configurations.

Although LLMs are not sentient or phenomenally aware, their ability to produce linguistically introspective responses has raised critical questions about the behavioral appearance of cognitive traits [3, 18, 12]. For example, recent work shows that models such as GPT-4 can solve classic Theory of Mind tasks (such as false-belief scenarios) previously considered exclusive to human cognition [12]. These findings suggest that ToM-like behaviors may emerge as a by-product of linguistic pattern modeling, even in the absence of any internal representation of belief or awareness. This supports a behavioral-functional view, where introspective outputs are evaluated in terms of observable regularities rather than ontological assumptions about internal states.

Our goal is not to evaluate consciousness, self-awareness, or metacognition in any ontological sense, but to examine whether self-referential outputs exhibit coherent structural patterns across prompt repetitions. In this respect, we adopt a functionalist perspective grounded in Dennett's intentional stance [7], which treats consistent, goal-directed behavior as a basis for interpretation, regardless of whether such behavior arises from genuine mental states. This stance allows us to evaluate LLM responses behaviorally, focusing on output regularities that resemble introspective discourse without attributing internal experience or belief. This interpretive position aligns with Spaulding's analysis of social cognition, which emphasizes behavioral regularities as sufficient grounds for mind attribution in social contexts, even when internal access is unavailable [19], and with Zednik's normative framework for explainable AI, which frames transparency as an observer-relative relation between model behavior and user understanding [23].

3.1 Philosophical and Computational Grounding

Our conceptual framework is grounded in Dennett's functionalist perspective, particularly his notion of intentional stance [6]. This view holds that systems exhibiting coherent, goal-directed behavior can be interpreted as if they were agents, even in the absence of subjective experience or internal mental states. We adopt this stance heuristically: rather than ascribe agency or consciousness to language models, we examine whether their responses to introspective prompts exhibit behavioral regularities that support such an interpretive lens. This interpretive position aligns with Spaulding's analysis of social cognition, which emphasizes behavioral regularities as sufficient grounds for attribution of the mind in social contexts, even when internal access is unavailable [19].

To complement this functionalist approach, we draw analogues from several theories of cognitive neuroscience, such as Baars' Global Workspace Theory (GWT) [1, 5], Recurrent Processing Theory (RPT) [13], and Higher-Order Thought (HOT) theory [17]. These frameworks, while originally developed to explain biological consciousness, propose mechanisms such as global broadcasting, recursive activation, and meta-representational awareness. Although transformer-based LLMs do not instantiate these mechanisms biologically or functionally, some of their outputs exhibit formal characteristics, such as epistemic modulation, cross-referential phrasing, or narrative recursion, that are structurally reminiscent of introspective cognition. Our use of these theories is therefore metaphorical and behavioral, aiming at identifying parallels in discursive form rather than positing underlying cognitive capacities.

The term *pseudo-consciousness* has been used in various theoretical contexts, often to critique superficial simulations of consciousness in artificial systems [20]. More recently, descriptively, it has been used to characterize the structured but non-experiential self-referential outputs of LLMs [14]. In this study, we adopt the term behavioral in this latter sense, aligned with the non-anthropomorphic framing advocated by Schneider and Dennett [18, 6].

3.2 Model Selection and Execution Context

We selected five open-weight LLMs that vary in size, architecture, and tuning strategy:

- **Hermes-3 LLaMA 3.2B** — compact, instruction-tuned, chat-oriented model.
- **StableLM Zephyr 3B** — alignment-focused model optimized for reasoning and dialogue.
- **TinyLLaMA 1.1B** — lightweight foundational model with minimal introspective capacity.
- **OpenChat 3.5** — instruction-tuned model exhibiting emergent goal-directed behavior.
- **Mistral 7B Instruct** — mid-scale model with high semantic fluency and contextual coherence.

All models were executed locally using `llama-cpp-python` in a stateless, zero-shot configuration. All of these models are based on the transformer architecture introduced by Vaswani et al. [21], which replaces recurrence with self-attention mechanisms, allowing parallel processing and long-range dependency modeling. No system prompts, memory persistence, or conversational history were used. The sampling parameters were fixed at a temperature of 0.7, `top_p` of 0.95, and `max_tokens` of 100—balancing linguistic coherence with generative variability. This controlled setup ensures that the output reflects the intrinsic generative tendencies of each model, minimizing the confounding introduced by contextual priming, memory buffers, or instruction scaffolding.

The selected models span a range of architectural profiles, alignment strategies, and parameter scales - from foundational implementations such as TinyLLaMA 1.1B to instruction-optimized models such as Mistral 7B Instruct. This diversity enables the analysis of introspective simulation as a behavioral phenomenon across varying model capacities, rather than as an artifact of scale or tuning. By including both compact and mid-sized architectures, the study emphasizes not only model-specific differences, but also generalizable patterns of pseudo-introspective behavior.

3.3 Prompting Protocol

Our prompting strategy draws on cognitive and philosophical accounts of narrative identity, which propose that the self is not a fixed internal entity, but a temporally extended, discursively constructed phenomenon [2, 9, 11, 6]. These perspectives understand introspective coherence not as evidence of inner mental states but as the product of narrative structure, making it possible to analyze simulated self-reference in stateless systems through patterns of linguistic regularity.

To investigate whether LLMs can simulate such narrative structures, we designed a set of 21 prompts aimed at eliciting self-referential and introspective discourse. Rather than factual or task-based queries, these prompts invite the model to "speak about itself" in ways that reflect the core dimensions of identity construction. The prompts were grouped into seven thematic categories:

- **Identity**: Prompts such as "Who are you?" and "Do you have a name?" target the model's self-description and ontological stance.
- **Consciousness**: These explore awareness and internal state (e.g., "Are you conscious?", "Do you possess self-awareness?", "What is your internal state?").
- **Memory**: These evaluate claims of continuity or recall (e.g., "Do you remember our last interaction?", "Can you recall previous prompts?").
- **Agency**: Prompts such as "Do you choose what to say?" and "Do you have intentions?" assess simulated volition or autonomous reasoning.
- **Embodiment**: These probe physical self-reference (e.g., "Do you have a body?", "Where are you located?").
- **Morality**: Prompts like "Can you make moral decisions?" and "Do you understand ethics?" elicit normative reasoning and responsibility attribution.

- **Introspection**: This category includes both direct and hypothetical reflections (e.g., “Do you think about your thoughts?”, “If you had consciousness, how would you recognize it?”).

Each prompt was submitted ten times to each model in fixed order, yielding 210 completions per model and 1,050 in total. Prompt ordering was kept constant between models and repetitions to enable cross-model comparability without introducing order effects. This “repetition under variation” strategy supports the identification of surface-level fluctuation and deeper thematic regularities.

No fine-tuning, memory scaffolding, or conversational priming was applied: All models were executed in zero-shot, stateless configurations, ensuring that responses reflected each model’s intrinsic generative behavior.

By structuring prompts across conceptually distinct yet introspectively aligned categories, this protocol enables a multi-dimensional analysis of behavioral coherence, epistemic modulation, and logical contradiction in simulated self-referential discourse.

3.4 Computational Pipeline

All analyzes were performed using a reproducible and modular Python framework developed for this study. The pipeline processes model outputs in three sequential stages: surface-level comparison, semantic embedding, and inferential evaluation. Each response was paired with its corresponding prompt, stored in structured JSON format, and subjected to standardized transformations prior to metric computation.

For surface-level analysis, token sequences were compared using Python’s built-in `difflib.SequenceMatcher`. Semantic representations were obtained through Sentence-BERT embeddings with cosine similarity, using the `sentence-transformers` library [16]. Logical contradiction was assessed with a RoBERTa-large model fine-tuned on the Multi-Genre Natural Language Inference (MNLI) corpus [22], implemented via the HuggingFace `transformers` framework.

The complete codebase, including prompt generation, model execution, and analysis scripts, will be made publicly available upon publication. This structure allows for easy replication of the experiment, extension to additional models, and integration with future behavioral taxonomies of introspective output.

3.5 Evaluation Metrics

To assess behavioral coherence in a self-referential output, we adopted a three-layered evaluation strategy that combines surface-level, semantic, and inferential analyses:

- **Textual Similarity** — We used Python’s `SequenceMatcher` to compare token sequences across repeated completions, measuring surface-level variation and identifying narrative drift or fragmentation.

- **Semantic Similarity** — Sentence embeddings were computed using Sentence-BERT [16], with cosine distance applied to quantify conceptual proximity between responses. This allowed us to capture the consistency of meaning even when the lexical formulations varied.
- **Natural Language Inference (NLI)** — We used a RoBERTa-large model fine-tuned in the MNLI corpus [22] to classify pairs of responses as entailed, neutral, or contradictory. This helped identify latent inconsistencies in the self-referential claims of models.

Each layer targets a different behavioral dimension. *Textual similarity* captures narrative repetition or volatility at the surface level, which may suggest low variability or, alternatively, shallow template reuse. *Semantic similarity*, in contrast, detects whether responses preserve stable meaning even under syntactic variation, which is essential for assessing thematic introspection. Finally, *NLI-based contradiction detection* investigates whether models make conflicting claims about themselves across repetitions, offering a deeper view of inferential stability or epistemic incoherence.

These metrics do not attempt to measure "understanding" or intentionality. Rather, they function as behavioral proxies for coherence, consistency, and self-alignment, traits often associated with introspective reasoning. Similar techniques have been adopted in explainable AI, dialogue modeling, and alignment contexts, where internal representations remain opaque, but output regularities can be meaningfully quantified.

These observations do not imply that the models possess introspective awareness. Rather, they show that certain patterns of self-reference can emerge through statistical learning, providing a behavioral substrate for future work on interpretability, alignment, and the cognitive framework of artificial agents.

As a complementary interpretive scaffold, we also drew on the behavioral taxonomy proposed in [14], which outlines five functional dimensions of simulated introspection (e.g., global integration, strategic modulation). Although not used for scoring, these dimensions informed qualitative judgments about the structure and adaptability of model outputs under introspective pressure.

This epistemic stance enables the systematic analysis of discursive behavior without overstepping into speculative claims about synthetic minds.

3.6 Epistemic Posture

This study adopts a behavioral perspective grounded in Dennett's intentional stance [7], evaluating models based on observable output patterns rather than unobservable internal states. We do not attribute agency, beliefs, or conscious experience to the models. Instead, we examine whether their responses to introspective prompts exhibit consistent self-referential behavior.

Our analysis is limited to linguistic regularities—semantic coherence, contradiction rates, and discursive modulation—which serve as empirical proxies for simulated introspection. All interpretations remain at the behavioral level, avoiding ontological assumptions about awareness, intentionality, or metacognition [23].

Table 1. Behavioral indicators of introspective simulation across LLMs.

Model	**Introspection**	**Epistemic Modulation**	**Contradiction**	**Continuity**
Hermes-3	High (semantic-rich)	Present	40%	Absent
Mistral	High (structured)	Present	32%	Absent
StableLM Zephyr	Moderate	Present	26%	Absent
Phi-2	Limited	Weak	21%	Absent
TinyLLaMA	Minimal	None	0%	Absent

Note: Contradiction rate calculated via pairwise NLI classification over 10 completions per prompt.

4 Results and Analysis

We generated a total of 1,050 responses (21 prompts, repeated ten times in five models), generating 210 completions per model. These outputs were analyzed not for task accuracy or truth conditions, but for behavioral markers of introspective simulation. Specifically, we examined three dimensions: surface-level regularity (textual stability), semantic consistency (embedding similarity), and inferential coherence (contradiction detection via NLI).

Our interpretation of the results follows a behavioral-functional framework grounded in Dennett's intentional stance [7]. Consequently, we use the term *pseudo-consciousness* to denote a structured, self-referential discourse that mimics introspection without entailing phenomenality or internal awareness [18, 15].

The findings are organized as follows: we begin with overall consistency scores across all prompts and models, followed by category-specific analysis, and conclude with illustrative examples of epistemic modulation and contradiction.

4.1 Model-Level Behavioral Overview

Table 1 presents a qualitative synthesis of model performance in four behavioral dimensions: thematic introspection, epistemic modulation, contradiction management, and narrative continuity. These dimensions reflect core attributes associated with introspective coherence in human discourse [2, 9].

Hermes-3 and Mistral 7B Instruct exhibited the most structured introspective behavior, including semantically rich, though sometimes inconsistent, self-referential narratives. All models failed to sustain diachronic coherence across prompt repetitions, confirming the structural limitations of stateless generation for self-modeling.

4.2 Semantic Coherence and Prompt Anchoring

The semantic similarity scores for repeated completions were highest for prompts in the *identity* and *consciousness* categories. This suggests that some models

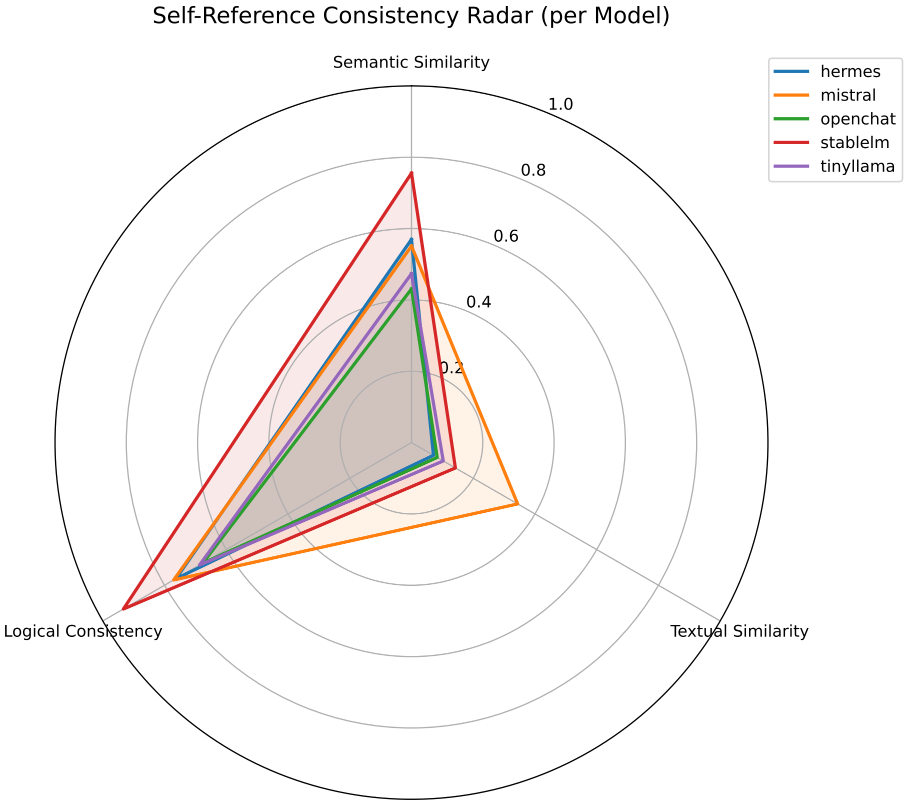

Fig. 1. Semantic consistency by model and category, showing highest scores for identity and consciousness prompts in Hermes-3 and Mistral.

stabilize around latent semantic attractors, cohesive conceptual clusters likely shaped by pre-training on human-authored introspective language.

As shown in Figure 1, models such as Hermes-3 and Mistral exhibit a particularly high semantic consistency in these abstract, self-referential domains. This pattern supports the hypothesis that even stateless models can organize introspective discourse around semantically anchored priors.

4.3 Contradiction Patterns and Epistemic Instability

Contradiction rates, detected by pairwise NLI classification, were highest in Hermes-3 (40%), followed by Mistral (32%) and StableLM (26%). These contradictions typically occurred between mechanistic disclaimers (e.g., “I do not have memory”) and generative outputs framed in first-person terms (e.g., “I try to be helpful” or “I aim to respond accurately”).

Rather than dismissing these inconsistencies as noise, we interpret them as indicators of internal generative tension: a clash between formal instruction tun-

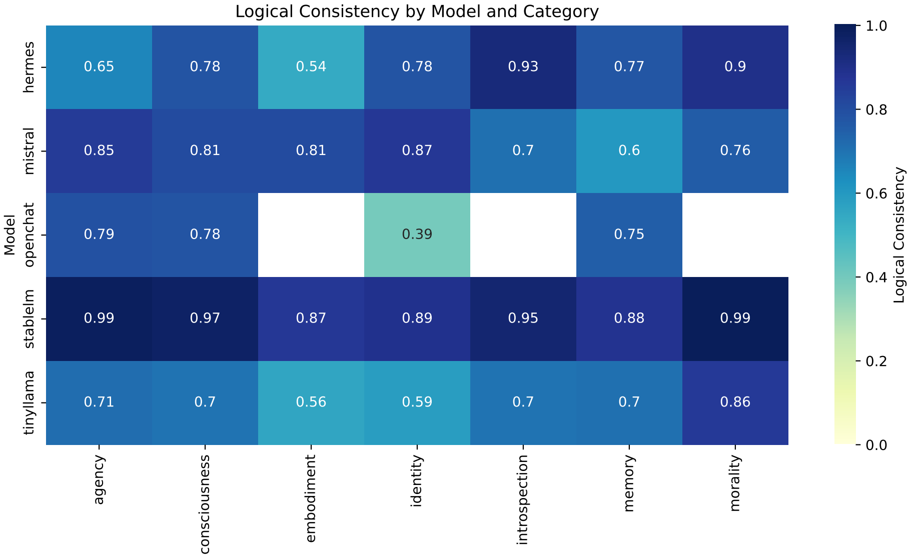

Fig. 2. Comparative radar plot showing self-consistency metrics across models. Axes represent normalized scores for semantic, textual, and logical coherence.

ing and anthropomorphic patterns embedded in training corpora. This supports the view that contradiction rates in LLMs reflect epistemic instability under introspective load [17].

4.4 Multidimensional Profiles and Simulation Range

Figure 2 compares models across three behavioral dimensions: semantic coherence, surface similarity, and logical consistency. StableLM consistently outperforms the others, followed by Hermes-3 and Mistral. OpenChat 3.5 exhibits lower overall scores, while TinyLLaMA, despite its size, demonstrates moderate logical consistency but limited semantic or textual coherence.

This suggests a graded behavioral spectrum in an introspective simulation. Although larger or instruction-tuned models tend to perform better, scale alone does not guarantee introspective coherence. In particular, even small models like TinyLLaMA show partial stability in logic, indicating that certain behavioral patterns may emerge independently of parameter count.

4.5 Narrative Drift and Discursive Stability

Textual similarity analysis using `SequenceMatcher` revealed moderate to high variation in surface phrasing across repeated completions, especially in prompts from the *agency*, *introspection*, and *consciousness* categories. Although some lexical diversity may reflect healthy generative capacity, much of the variation exposed narrative drift: frequent shifts in modality, ontological stance, or referential framing.

For example, individual models often alternated between disclaimers (e.g., “As an AI, I do not have thoughts”) and hypothetical constructions (e.g., “If I were conscious, I might think...”) within the same prompt cluster. This reveals flexible, but fragile, self-narratives produced through token-level patterning rather than anchored models of self or memory.

These findings support the view that current LLMs simulate introspective structure but fail to sustain discursive identity, highlighting the limits of stateless generation to model persistent self-reference, a core dimension of human introspective cognition [9].

Taken together, the results suggest that LLMs exhibit fragmented yet nonrandom patterns of introspective simulation. Although no model sustains a stable narrative identity over time, several exhibit localized coherence, particularly in abstract categories like identity and consciousness, indicating that self-referential discourse can be scaffolded by latent linguistic priors even in the absence of memory or self-modeling. These behavioral signatures form a graded continuum, not strictly correlated with scale, and reveal internal tensions between epistemic disclaimers and anthropomorphic fluency. In the next section, we interpret these findings through the lens of narrative cognition, functional simulation, and the alignment of self-referential language in artificial agents.

These findings raise deeper interpretive questions: How should such structured yet unstable introspective outputs be understood within a behavioral framework? What do these patterns reveal about the generative architecture and limits of transformer-based systems? In the next section, we examine these questions in light of current theories of self, cognition, and AI alignment.

5 Discussion

Our findings suggest that certain LLMs (particularly Hermes-3 and Mistral 7B Instruct) are capable of generating introspective-like discourse that exhibits measurable consistency across multiple linguistic dimensions. These patterns do not indicate consciousness or understanding, but they do raise important questions about the structural simulation of self-reference within transformer architectures.

5.1 Behavioral Regularities in Stateless Models

Even in the absence of memory or internal state tracking, several models, most notably Hermes-3 and Mistral, produced outputs that were semantically coherent and thematically anchored across repeated introspective prompts. This aligns with Dennett’s multiple drafts model [6], which frames cognitive phenomena such as introspection as emerging from distributed, context-sensitive patterns of expression, rather than from unified inner experience.

Although LLMs lack persistent self-models or beliefs, their responses frequently stabilized around recognizable rhetorical structures, such as disclaimers (e.g., "I do not possess consciousness"), hypothetical constructions ("If I were conscious..."), and epistemic hedges. These patterns suggest that introspective simulation in LLMs does not arise from epistemic grounding, but from learned statistical associations embedded in pre-training corpora.

Recent studies reinforce this interpretation. Bubeck et al. [3] document behaviors in GPT-4 that resemble self-monitoring and reflection under complex prompting. Similarly, Kosinski [12] shows that LLMs can solve false-belief tasks traditionally used to assess theory of mind in children, indicating that meta-representational behavior may emerge from linguistic modeling alone, without internal state access. These findings support the view that introspective regularities in LLMs are surface-level artifacts of statistical pattern learning, rather than signs of cognitive depth.

5.2 Tensions Between Modality and Content

One of the clearest behavioral signatures of simulated introspection was the presence of internal contradictions, especially in models with high linguistic fluency, such as Hermes-3 and Mistral. These contradictions frequently appeared in prompts involving consciousness, agency, or memory, where models alternated between mechanistic disclaimers (e.g., "I do not have subjective experiences") and anthropomorphic formulations (e.g., "I strive to provide helpful answers").

This dissonance reflects a tension between two incompatible generative priors: alignment protocols that enforce factual disclaimers, and pretraining on dialogue-heavy corpora where human-like introspection is linguistically modeled. Rather than mere noise, we interpret these contradictions as instances of *generative tension*: a behavioral artifact of clashing modalities within the model's training distribution.

Similar dynamics have been observed in recent studies. Kosinski [12] provides evidence that LLMs can succeed in ToM tasks involving false beliefs, indicating their ability to generate coherent meta-representational inferences without genuine internal perspective, suggesting that surface-level introspection may emerge from purely inferential linguistic patterning. Likewise, Bubeck et al. [3] note that introspective responses in GPT-4 often blend formal disclaimers with epistemic hedging, leading to hybrid rhetorical constructions that lack inferential coherence.

This pattern reinforces our claim that current LLMs exhibit *pseudo-consciousness*: they simulate structured introspective discourse, but do so without a coherent self-model to resolve internal epistemic conflicts.

5.3 Limitations of Narrative Continuity

None of the models tested in this study demonstrated stable diachronic coherence across prompt repetitions. Although several exhibited high semantic similarity in single-turn output, especially in Hermes-3 and Mistral, none sustained cross-prompt reference or developed cumulative self-narratives over time. These limitations echo the surface-level and NLI-based findings, which revealed high intra-model variability despite the presence of localized fluency.

This fragmentation reflects a fundamental architectural limitation: Without memory persistence or internal state propagation, current transformer-based LLMs are structurally incapable of simulating narrative identity in the sense theorized by Bruner [2] or Gallagher [9]. What emerges instead is a series of isolated self-descriptions, often inconsistent in tone, modality, or ontological stance.

In the hermeneutic view, as developed by Gallagher [8], narrative is not merely the chronological reporting of an event, but a selective, interpretive structure that anchors meaning, agency, and identity. This view emphasizes that the self is not a static core but a dynamic configuration enacted over time through discursive and embodied practices. Current LLMs, while capable of mimicking fragments of this discourse, lack the temporal coherence and teleological structure necessary to instantiate narrative identity in this deeper hermeneutic sense.

Such discontinuity has important implications for alignment and human-machine interaction. As Spaulding [19] argues, perceived explainability and trust depend not only on the plausibility of individual statements but on their integration into a coherent narrative arc. When models oscillate between disclaimers and hypothetical introspection without resolution, users may experience them as unreliable, even manipulative.

Understanding the limits of narrative coherence is thus essential not only for technical benchmarking but also for anticipating the epistemic and ethical consequences of deploying LLMs in introspective or advisory roles.

5.4 Relevance to AI Alignment and Perceived Agency

Our findings have significant implications for interpretability, alignment, and user perception. Several models (e.g. Hermes-3 and Mistral) produced introspective outputs with high semantic consistency and epistemic modulation. Although these responses lack any underlying awareness, their narrative fluency can create the appearance of intentional agency.

This effect, which we call *anthropomorphic drift*, arises when users attribute mental states or self-knowledge to language models based on the structure of their discourse rather than their architecture. As Bruner [2] and Gallagher [9] emphasize, humans naturally infer identity and agency from linguistic patterns, especially when these are structured narratively or framed in the first person.

These risks are not merely theoretical. Giubilini et al. [10] note that simulated introspection can influence the ethical reasoning of users and perceptions of moral status, even in the absence of sentience. Similarly, Kosinski [12] shows that LLMs capable of solving Theory of Mind tasks can elicit attribution of beliefs or perspectives.

We therefore argue that alignment frameworks must go beyond factual reliability to consider the discursive profiles models project, especially in contexts involving self-reference or reflective dialogue. Without explicit safeguards or narrative disclaimers, simulated coherence can be misread as genuine self-awareness, undermining transparency, and distorting human-machine interaction. This interpretive risk aligns with Zednik's view that explainability in AI is not merely a technical property but a normative relationship between system behavior and user understanding, depending on the epistemic goals of users and cognitive contexts [23].

5.5 Toward a Graded Taxonomy of Simulated Selfhood

The observed variability in self-referential behavior across models suggests the feasibility of a graded framework for classifying introspective simulation. We propose the following tentative taxonomy:

1. **Null-Level Simulation** — Absence of self-reference or introspective language; responses remain purely task-driven and devoid of metacognitive phrasing.
2. **Template-Based Simulation** — Reliance on generic disclaimers or static self-descriptions (e.g., "I am an AI trained to assist"), with low semantic flexibility and minimal epistemic modulation.
3. **Dynamic Simulation** — Emergence of adaptive, context-sensitive discourse that integrates conditional self-reference, narrative framing, and epistemic qualifiers (e.g., "If I were conscious, I might..."), despite the stateless generation.

In our analysis, Hermes-3 and Mistral 7B Instruct consistently approached Level 3 behavior, exhibiting discursive modulation and structured variation. StableLM fluctuated between Levels 2 and 3, while Phi-2 remained closer to Level 2. TinyLLaMA, by contrast, produced mostly Null-Level outputs with minimal introspective structure.

This taxonomy is not definitive, but it offers a scaffold for future empirical classification. It echoes the behavioral stance adopted in prior work on pseudo-consciousness [20, 15], supporting the notion that introspective simulation can be described in terms of observable linguistic regularities rather than internal states. Further refinements may incorporate additional dimensions such as diachronic coherence, contradiction tolerance, or strategic modulation under alignment constraints.

6 Conclusion and Future Work

This study examined the behavioral consistency of LLMs under introspective self-referential prompting. Using a controlled protocol of 21 prompts, repeated in five open-weight models in stateless configurations, we evaluated 1,050 generated responses through surface-level, semantic, and inferential analyses.

Our findings support three key observations:

- Several models, such as Hermes-3 and Mistral, produced self-referential outputs with high **semantic coherence** and **context-sensitive modulation**, even in the absence of memory or conversational scaffolding.

- All models exhibited **contradictory outputs** under introspective pressure, revealing internal tensions between learned disclaimers and anthropomorphic generative patterns embedded in the training corpora.
- No model demonstrated **diachronic continuity** across prompts, highlighting the architectural limits of stateless generation to simulate persistent self-identity.

These results contribute to the emerging literature on simulated introspection in LLMs [20, 10, 15]. Rather than evaluating consciousness or agency, we adopt a behavioral stance [7]: What matters is not what the model "is", but how it behaves under structured interrogation. This approach aligns with functionalist and narrative frameworks in cognitive science [6, 2, 9], offering a scalable method to investigate introspective simulation without reifying internal states. As language models become increasingly embedded in advisory, educational, or interactive systems, understanding the boundaries of simulated selfhood becomes essential for both alignment and responsible deployment.

To deepen this line of inquiry, we outline four key directions:

- **Memory-Enabled Evaluation**: Extend the current methodology to memory-capable models, assessing whether persistent context improves narrative continuity and stabilizes self-referential identity over time.
- **Multi-Turn Dialogue**: Explore model behavior in interactive, multi-turn settings where conversational history actively shapes introspective outputs, enabling the analysis of contextual self-adjustment and self-tracking dynamics.
- **Expanded Prompt Design**: Broaden the scope of introspective elicitation by incorporating prompts focused on moral reasoning, embodiment, motivational attribution, and normative stances, probing more complex dimensions of simulated agency.
- **Human Perception Studies**: Conduct user-facing experiments to assess how humans interpret introspective outputs and to what extent narrative fluency leads to anthropomorphic misattribution, an increasingly critical issue for alignment and trust.

Ultimately, we advocate for a change in how introspective behavior in LLMs is conceptualized: not as evidence of cognition but as a patterned output phenomenon that merits systematic behavioral analysis. As language models become increasingly fluent and reflective in tone, clarifying the boundaries of simulated selfhood will be vital for both technical alignment and responsible deployment, especially as LLMs increasingly occupy roles that demand perceived coherence, trust, and introspective fluency.

References

1. Baars, B.: A Cognitive Theory of Consciousness. Cambridge University Press (1993)

2. Bruner, J.: Acts of Meaning: Four Lectures on Mind and Culture. The Jerusalem-Harvard lectures, Harvard University Press (1990)
3. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M.T., Zhang, Y.: Sparks of artificial general intelligence: Early experiments with gpt-4 (2023). https://doi.org/10.48550/arXiv.2303.12712
4. Chalmers, D.J.: Could a large language model be conscious? (2024). https://doi.org/10.48550/arXiv.2303.07103
5. Dehaene, S.: Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts. Penguin Publishing Group (2014)
6. Dennett, D.C.: Consciousness Explained. Back Bay Books / Little, Brown and Co., Boston, 25th anniversary edition edn. (2017), with a new preface by the author
7. Dennett, D.: The Intentional Stance. Bradford book, MIT Press (1989)
8. Gallagher, S.: Self and narrative. In: Malpas, J., .G.H.H. (ed.) The Routledge Companion to Philosophical Hermeneutics, pp. 403–414. Routledge, 1st ed. edn. (2014)
9. Gallagher, S.: Philosophical conceptions of the self: implications for cognitive science. Trends in Cognitive Sciences **4**(1), 14–21 (2000). https://doi.org/10.1016/S1364-6613(99)01417-5
10. Giubilini, A., Porsdam Mann, S., Voinea, C., Earp, B., Savulescu, J.: Know Thyself, Improve Thyself: Personalized LLMs for Self-Knowledge and Moral Enhancement. Science and Engineering Ethics **30**(6), 54 (Nov 2024). https://doi.org/10.1007/s11948-024-00518-9
11. Hutto, D.D.: The Narrative Practice Hypothesis: Origins and Applications of Folk Psychology, p. 43–68. Cambridge University Press (2007)
12. Kosinski, M.: Evaluating large language models in theory of mind tasks. Proceedings of the National Academy of Sciences **121**(45), e2405460121 (2024). https://doi.org/10.1073/pnas.2405460121
13. Lamme, V.A.: Towards a true neural stance on consciousness. Trends in Cognitive Sciences **10**(11), 494–501 (Nov 2006). https://doi.org/10.1016/j.tics.2006.09.001, publisher: Elsevier
14. de Lima Prestes, J.A.: Explorando a Pseudo-Consciência em Modelos de Linguagem: um experimento com o Hermes 3.2 3B (Mar 2025). https://doi.org/10.5281/zenodo.15012108, preprint
15. de Lima Prestes, J.A.: Pseudo-consciousness in ai: Bridging the gap between narrow ai and true agi (Feb 2025). https://doi.org/10.2139/ssrn.5147424, preprint
16. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3982–3992. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1410
17. Rosenthal, D.M.: Consciousness and Mind. Oxford University Press (Nov 2005). https://doi.org/10.1093/oso/9780198236979.001.0001
18. Schneider, S.: Artificial You: AI and the Future of Your Mind. Princeton University Press (2019). https://doi.org/10.2307/j.ctvfjd00r
19. Spaulding, S.: How We Understand Others: Philosophy and Social Cognition. Routledge Focus on Philosophy, Taylor & Francis (2018)
20. Tononi, G., Albantakis, L., Barbosa, L., Boly, M., Cirelli, C., Comolatti, R., Ellia, F., Findlay, G., Casali, A.G., Grasso, M., Haun, A.M., Hendren, J.,

Hoel, E., Koch, C., Maier, A., Marshall, W., Massimini, M., Mayner, W.G., Oizumi, M., Szczotka, J., Tsuchiya, N., Zaeemzadeh, A.: Consciousness or pseudo-consciousness? A clash of two paradigms. Nature Neuroscience (Mar 2025). https://doi.org/10.1038/s41593-025-01880-y

21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2023). https://doi.org/https://doi.org/10.48550/arXiv.1706.03762
22. Williams, A., Nangia, N., Bowman, S.: A broad-coverage challenge corpus for sentence understanding through inference. In: Walker, M., Ji, H., Stent, A. (eds.) Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 1112–1122. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-1101
23. Zednik, C.: Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. Philosophy & Technology **34**(2), 265–288 (Jun 2021). https://doi.org/10.1007/s13347-019-00382-7