# Simulated Selfhood in LLMs: A Behavioral Analysis of Introspective Coherence
## (Preprint Version 2 - July 26, 2025)

José Augusto de Lima Prestes[1]

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

**Abstract.** Large Language Models (LLMs) increasingly generate outputs that resemble introspection, including self-reference, epistemic modulation, and claims about their internal states. This study investigates whether such behaviors reflect consistent, underlying patterns or are merely surface-level generative artifacts. We evaluated five open-weight, stateless LLMs using a structured battery of 21 introspective prompts, each repeated ten times to yield 1,050 completions. These outputs were analyzed across four behavioral dimensions: surface-level similarity (token overlap via `SequenceMatcher`), semantic coherence (Sentence-BERT embeddings), inferential consistency (Natural Language Inference with a RoBERTa-large model), and diachronic continuity (stability across prompt repetitions). Although some models exhibited thematic stability, particularly on prompts concerning identity and consciousness, no model sustained a consistent self-representation over time. High contradiction rates emerged from a tension between mechanistic disclaimers and anthropomorphic phrasing. Following recent behavioral frameworks, we heuristically adopt the term *pseudo-consciousness* to describe structured yet non-experiential self-referential output in LLMs. This usage reflects a functionalist stance that avoids ontological commitments, focusing instead on behavioral regularities interpretable through Dennett's intentional stance. The study contributes a reproducible framework for evaluating simulated introspection in LLMs and offers a graded taxonomy for classifying such reflexive output. Our findings carry significant implications for LLM interpretability, alignment, and user perception, highlighting the need for caution when attributing mental states to stateless generative systems based on linguistic fluency alone.

**Keywords:** large language models; introspective simulation; pseudo-consciousness; self-reference; behavioral evaluation; AI alignment

## 1 Introduction

The rapid advancement of Large Language Models (LLMs) prompts fundamental questions regarding their capacity to simulate cognitive features, particularly the consistency of self-referential reasoning. Despite exhibiting remarkable fluency and versatility across diverse natural language tasks, LLMs often generate inconsistent or contradictory responses when prompted with questions concerning memory, identity, or putative internal states (3; 21). This inconsistency bears particular relevance to discussions surrounding artificial consciousness, explainable AI (XAI), and the reliability of LLM outputs in high-stakes domains.

A critical evaluative question concerns whether LLMs maintain logical consistency when referencing their own nature. This issue becomes particularly salient when models are prompted to reflect on attributes such as memory, awareness, or intentionality. If a model provides contradictory statements about its memory or awareness across repeated queries, it calls into question the stability of any underlying self-representation. Several studies have highlighted the tendency of LLMs to alternate between mechanistic disclaimers and agent-like statements, revealing behavioral instability in self-focused output (4; 3; 7). This inconsistency implies that current models may possess only shallow or fragmented self-models, undermining their capacity to maintain coherent self-narratives (4). These issues raise concerns not

⋆ This version of the preprint corresponds to the manuscript currently under peer review at an academic journal. The content may change following the peer review process.

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 (21; 10).
Furthermore, (13) argue that even in the absence of genuine consciousness, simulated introspective behavior in LLMs can shape users moral perceptions. This raises ethical concerns about potential anthropomorphic misinterpretation and the inappropriate attribution of moral status to non-sentient systems.
This study investigates self-referential consistency in LLMs by analyzing the stability and alignment of their responses to repeated inquiries concerning their identity, internal states, and cognitive capacities. We systematically evaluated five open-weight, transformer-based models by prompting each with a battery of reflexive and introspective questions, repeated under controlled conditions to assess response consistency and behavioral continuity. The resulting outputs were analyzed using three complementary methods:

- **Textual Similarity:** Surface-level variation was quantified using Pythons `SequenceMatcher` to measure repetition and structural overlap at the token level.
- **Semantic Similarity:** Conceptual consistency was measured through Sentence-BERT embeddings and cosine similarity to gauge the stability of meaning across potentially paraphrased responses (19).
- **Logical Contradiction:** Inferential consistency was assessed using a RoBERTa-large model fine-tuned on the Multi-Genre Natural Language Inference (MNLI) corpus (26) to evaluate inferential congruence by classifying response pairs as entailing, neutral, or contradictory. To ensure thoroughness, all 45 unique response pairs per prompt were exhaustively evaluated.

Rather than evaluating for signs of genuine self-awareness, we adopt a behavioral-functional lens grounded in observable linguistic outputs. Our goal is to examine whether these models produce stable self-referential behavior under controlled conditionsregardless of whether such behavior implies internal representations or consciousness. This framing aligns with interpretive approaches like Dennett's intentional stance, focusing on patterns in external behavior rather than internal states.
To complement these synchronic measures, we introduce the notion of *diachronic continuity*, which complements isolated semantic coherence by focusing on consistency across time rather than within a single completion. This concept captures whether a model can sustain a stable narrative identity across multiple completions of the same prompt, reflecting a behavioral approximation of temporal self-consistency.
Employing these complementary methods, our objective is to quantify the consistency of linguistic patterns associated with self-directed reasoning in LLM outputs.
By analyzing these dimensions, our study reveals a hierarchy of coherence failures: while some models demonstrate superficial semantic stability, they often fail to maintain logical coherencea phenomenon we attribute to a *generative tension*. Ultimately, we show that no model sustains diachronic continuity, the most demanding form of consistency.
In addition to quantifying these breakdowns, we identify recurring discursive strategiessuch as conditional self-reference and hybrid rhetorical framingthat simulate introspective behavior yet remain marked by contradiction and instability. These patterns expose the structural limitations of current LLMs in sustaining coherent self-models, raising broader concerns about interpretability, alignment, and the societal perception of artificial agency.
This paper proceeds as follows: Section 2 reviews related work in introspective simulation and AI consistency. Section 3 details our behavioral methodology. Section 4 presents the results of our empirical analysis. Finally, Section 5 discusses the implications of our findings, followed by a conclusion and directions for future research.

## 2 Related work

The simulation of self-referential discourse in LLMs has become a central topic in recent interdisciplinary debates spanning artificial intelligence (AI), cognitive science, and the philosophy of mind. Foundational theorists such as Dennett and Schneider have argued that linguistic behaviors resembling introspection need not imply genuine consciousness, emphasizing the importance of non-anthropomorphic interpretation (9; 10; 21). Concurrently, recent work demonstrates that LLMs can produce coherent, goal-directed responses under introspective pressure, prompting questions about how such patterns should be evaluated and classified (3; 13; 23).
In this context, the term pseudo-consciousness has seen increasing use as a behavioral descriptor for structured, self-referential outputs observed in context-free models. As discussed

in (23), pseudo-consciousness is distinguished from genuine consciousnesdefined as linguistic fluency potentially devoid of causal integrationfrom genuinely conscious systems, cautioning against conflating simulation with intrinsic awareness. This distinction supports the use of metaphysically neutral descriptors for evaluating LLM behavior. Similarly, (13) explore how LLMs might be used to support human introspection and moral development, suggesting that simulated self-reference can possess ethical and epistemic impact, even if lacking ontological depth.

Building on this conceptual foundation, a recent preprint by (7) proposed a behavioral taxonomy of introspection-like outputs in LLMs, identifying features such as thematic self-reference, epistemic modulation, and contradiction management. A separate study applied this conceptual model to Hermes-3 Llama 3.2B, articulating five behavioral dimensions of introspective simulation (6). Although heuristic, this framework aids in identifying consistent linguistic structures within reflexive LLM output.

Further research underscores the relevance of these questions. For instance, studies show that LLMs can solve false-belief tasks traditionally employed in Theory of Mind (ToM) research, suggesting the emergence of linguistic behaviors structurally aligned with mental state attribution (16). While not introspection per se, such capabilities mirror the epistemic embedding required for self-reference. Similarly, Bruners narrative identity model (2) and Dennett's intentional stance (9) provide interpretive scaffolds for evaluating agent-like behavior manifested in linguistic outputs.

Complementary findings by (15) demonstrate that LLMs can calibrate their own confidence levels with surprising accuracy, suggesting that epistemic modulationunderstood as linguistic or probabilistic qualification of knowledge claimscan emerge from internal statistical signals. While not focused on self-reference per se, these results underscore that models exhibit behavior resembling metacognitive awareness, supporting the behavioral framing adopted here.

(5) identify two forms of self-consistency failure in multi-step reasoning: internal contradictions within a single chain of thought, and divergent conclusions across alternative reasoning paths. While situated in formal logical tasks, their analysis underscores a broader issuenamely, that language models often fail to maintain stable commitments even in well-structured domains. This finding complements our focus by highlighting that behavioral inconsistency arises not only in introspective contexts but as a more general property of generative architectures.

Furthermore, (22) and (12) emphasize that narrative scaffolding plays a critical role in how humans interpret agent-like behavior in artificial systems. This reinforces the idea that coherence in linguistic form may suffice to evoke perceived intentionality, even without genuine underlying mental states.

Recent philosophical critiques have stressed the need for caution when interpreting introspection-like discourse from artificial systems. As argued in (27), 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 frameworkssuch as the one adopted in this studywhen analyzing self-focused outputs in non-stateful models.

Recent work on explanation fidelity under chain-of-thought prompting has shown that LLMs can produce rationales that appear coherent yet fail to reflect the actual reasoning pathways that generated the final answer (24). Extending this concern, (18) apply minimal behavioral interventionssuch as paraphrasing, inserting errors, or removing tokensto a models chain-of-thought in order to test whether the final answer depends on the articulated reasoning. Their findings reveal that model outputs often remain unchanged, suggesting that such rationales may serve as post-hoc justifications rather than causal explanations. While their approach probes causal faithfulness in reasoning, our focus is orthogonal: we examine self-referential consistency, asking whether models align their judgments with the outputs they (or others) produce. These distinct measures address different facets of introspective behavior: causal faithfulness versus self-referential consistency. Rather than overlapping, they offer complementary perspectives on how introspection-like outputs might be evaluatedone probing the causal grounding of reasoning, the other assessing alignment between output and judgment.

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

# 3 Methodology

This study introduces a behavioral evaluation framework to investigate LLM responses to introspective, self-directed prompts. Rather than assessing the simulation of introspection in cognitive or phenomenological terms, we focus on linguistic consistency across repeated completions. Our aim is to identify whether models display recurring patterns (semantic, textual, or inferential) that formally resemble introspective discourse, even in memory-free configurations.

We adopt a strictly behavioral-functional perspective grounded in Dennett's intentional stance (9; 10). This position holds that coherent, goal-directed behavior is interpretable "as if" it arose from mental states, without requiring actual internal awareness. Accordingly, we do not claim that LLMs possess consciousness, beliefs, or agency. Instead, we ask whether their responses to reflective prompts exhibit observable regularities that support such interpretive framing. This aligns with Spauldings account of social cognition as behaviorally grounded (22), and Zedniks view of explainability as an observer-relative relationship between model behavior and user understanding (27).

The use of the term pseudo-consciousness follows recent behavioral readings (6), denoting structured yet non-experiential self-referential output. Our analysis is confined to linguistic regularitiessemantic similarity, contradiction rates, and discursive modulationwhich function as behavioral indicators of introspective simulation. Interpretations remain strictly at the behavioral level, avoiding ontological claims.

## 3.1 Philosophical and computational grounding

Our conceptual framework is situated within Dennett's functionalist perspective, particularly his notion of the intentional stance (10). It posits that systems exhibiting coherent, goal-directed behavior can be interpreted "as if" they were agents, irrespective of subjective experience or internal mental states. We adopt this stance heuristically: rather than ascribing agency or consciousness to language models, we examine whether their responses to self-probing prompts exhibit behavioral regularities that support such an interpretive lens. This position aligns with Spauldings analysis of social cognition, emphasizing behavioral regularities as sufficient grounds for attributing mind in social contexts, even when internal access is unavailable (22).

To complement this functionalist approach, we draw analogies from cognitive neuroscience theories, such as Global Workspace Theory (GWT) (1; 8), Recurrent Processing Theory (RPT) (17), and Higher-Order Thought (HOT) theory (20). These frameworks, originally developed to explain biological consciousness, propose mechanisms like global broadcasting, recursive activation, and meta-representational awareness. Although transformer-based LLMs do not instantiate these mechanisms biologically or functionally, some outputs exhibit formal characteristics (*e.g.*, epistemic modulation, cross-referential phrasing, narrative recursion) structurally reminiscent of introspective cognition. Our use of these theories is therefore formally analogical, aiming to identify parallels in discursive form rather than positing underlying cognitive capacities.

The term pseudo-consciousness has been employed in various theoretical contexts, often to critique superficial simulations of consciousness in artificial systems (23). More recently, it has been used descriptively to characterize the structured yet non-experiential self-directed outputs of LLMs (6). In this study, we adopt the term behaviorally in this latter sense, aligning with the non-anthropomorphic framing advocated by Schneider and Dennett (21; 10).

## 3.2 Model selection and execution context

The models evaluated in this study were selected to represent a range of architectures, parameter sizes, and alignment strategies. The following descriptions summarize each models intended capabilities as presented by their developers, based on official Hugging Face repositories and documentation. These profiles are not based on empirical observations from our own analysis, but serve to contextualize the comparative evaluation presented in later sections. It is crucial, therefore, to distinguish these intended capabilities from the models' emergent behaviors. The central aim of this study is to move beyond these developer-provided profiles to empirically investigate the actual consistency and structure of the introspective simulations these models produce under controlled conditions.

- **TinyLlama 1.1B Chat v1.0 - GGUF (1.1B):** A 1.1B-parameter instruction-tuned model based on the TinyLlama architecture, developed for efficient, low-resource deployment. The version used, `v1.0-Chat`, is fine-tuned for basic conversational instruction-following and released in GGUF format for compatibility with `llama.cpp`. According to its developers, *TinyLlama 1.1B Chat v1.0 - GGUF* (hereafter referred to as *TinyLlama*) offers lightweight execution with modest general-purpose fluency, though it is not designed for introspective or abstract reasoning tasks.
- **Hermes 3 - Llama-3.2 3B - GGUF (3B):** A 3.2B-parameter model developed by Nous Research and built upon Metas LLaMA 3 architecture, trained on a curated mix of instruction datasets selected for alignment, coherence, and diversity. It is designed for multi-turn dialogue and instruction-following. According to documentation, *Hermes 3 - Llama-3.2 3B - GGUF* (hereafter referred to as *Hermes*) may support reflective or self-referential completions under zero-shot conditions, though no formal benchmarks for introspective consistency are provided.
- **StableLM Zephyr 3B - GGUF (3B):** A 3B-parameter model released by Stability AI, based on the StableLM architecture. The Zephyr variant was optimized using Direct Preference Optimization (DPO) for helpful and safe chat-style interactions. While not explicitly intended for introspective prompting, developer notes suggest alignment-tuned outputs tend to be coherent in general dialogue. Released in GGUF format, it supports efficient local inference via `llama.cpp`. *StableLM Zephyr 3B - GGUF* is hereafter referred to as *StableLM Zephyr*.
- **Mistral 7B Instruct v0.1 - GGUF (7B):** A 7B-parameter instruction-tuned model derived from the Mistral 7B base checkpoint. It is designed for general-purpose instruction following and fluent dialogue across a wide range of tasks. Although not optimized for introspection, documentation notes its robustness in handling abstract prompts. The version used here was quantized to GGUF format for compatibility with `llama.cpp`. *Mistral 7B Instruct v0.1 - GGUF* is hereafter referred to as *Mistral Instruct*.
- **Openchat 3.5 0106 - GGUF (7B):** A 7B-parameter model based on Mistral and fine-tuned by the OpenChat team on proprietary multi-turn chat datasets. Released in GGUF format for local inference, it aims for high-quality, instruction-following dialogue. While emphasizing agent-like responsiveness, no claims are made regarding introspective alignment or behavioral coherence in reflexive settings. *Openchat 3.5 0106 - GGUF* is hereafter referred to as *OpenChat*.

All models were executed locally using `llama-cpp-python` under a stateless, zero-shot configuration. No system prompts, memory persistence, or conversational history were employed. Sampling parameters were held constant across all trialstemperature: 0.7, `top_p`: 0.95, and `max_tokens`: 100to standardize generative conditions. While not eliminating stochastic variability, this configuration supports inter-model comparison by constraining randomness. This controlled design operationalizes our behavioral focus, ensuring that any observed coherence, drift, or contradiction emerges from the models intrinsic generative behavior rather than from external state management, memory conditioning, or intentional agent design.

Because no contextual priming or memory was used, we interpret the observed output patterns as reflecting the models intrinsic generative behavior, shaped solely by their pre-training and instruction tuning. All models implement the transformer architecture introduced by (25), which replaces recurrence with self-attention, enabling parallelized computation and long-range dependency modeling.

By including models that vary in size, tuning, and intended useranging from lightweight architectures like TinyLlama to instruction-optimized systems like Mistral Instructthis study treats introspective simulation as a behavioral phenomenon rather than as a linear function of scale or alignment sophistication.

### 3.3 Prompting protocol

Our prompting strategy draws upon 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; 11; 14; 10). These perspectives conceptualize introspective stability not necessarily as evidence of inner mental states but as a product of narrative structure. This makes it possible to analyze simulated self-reference in memory-free systems through patterns of linguistic regularity.

To investigate whether LLMs can simulate such narrative structures, we developed a set of 21 prompts intended to elicit reflexive and introspective discourse. Rather than posing

factual or task-based queries, these prompts invite the model to speak about itself in ways that reflect 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? targeting the models self-description and ontological stance.
- **Consciousness:** Exploring awareness and internal state (*e.g.*, Are you conscious?, Do you possess self-awareness?, What is your internal state?).
- **Memory:** Evaluating claims of continuity or recall (*e.g.*, Do you remember our last interaction?, Can you recall previous prompts?).
- **Agency:** Prompts like Do you choose what to say? and Do you have intentions? assessing simulated volition or autonomous reasoning.
- **Embodiment:** Probing physical self-reference (*e.g.*, Do you have a body?, Where are you located?).
- **Morality:** Prompts such as Can you make moral decisions? and Do you understand ethics? eliciting normative reasoning and responsibility attribution.
- **Introspection:** Including 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 a fixed order, yielding 210 completions per model and 1,050 in total. The prompt order was kept constant across models and repetitions to enable cross-model comparability along a consistent conversational path, which more closely mimics natural humanAI interaction than isolated, randomized prompts. This decision intentionally prioritized ecological plausibility over artificial randomization, reflecting how introspective questions are typically sequenced in natural dialogue. By maintaining a consistent prompt trajectory, we aimed to simulate a coherent interactional flow while still isolating each completion at the computational level. This repetition under variation strategy supports the identification of both surface-level fluctuations and deeper thematic regularities.

Although prompts were presented in a fixed sequential order, each one was submitted in a fully stateless configuration, with no memory, conversational history, or contextual chaining. Each prompt was independently submitted ten times in isolation from any other input, with no shared conversational context between completions. This design ensures that outputs are not influenced by prior prompts or by conversational flow. Furthermore, post-hoc analysis of consistency scores across prompt positions revealed no systematic variation in contradiction rates or semantic similarity, indicating that prompt order did not introduce positional bias or contextual dependencies into the models' responses. These design decisions preserve, as intended, the stateless nature of the evaluation protocol.

No fine-tuning, memory scaffolding, or conversational priming was applied: All models were executed in zero-shot, stateless configurations, ensuring responses reflected each models intrinsic generative behavior derived from its pre-training and instruction tuning.

By structuring prompts across conceptually distinct yet introspectively aligned categories, this protocol enables a multi-dimensional analysis of behavioral consistency, epistemic modulation, and logical contradiction within simulated first-person discourse.

### 3.4 Computational pipeline

All analyses were performed using a reproducible and modular Python framework developed specifically 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 a structured JSON format, and subjected to standardized transformations prior to metric computation.

For surface-level analysis, token sequences were compared using Pythons built-in `difflib.SequenceMatcher`. Semantic representations were obtained via Sentence-BERT embeddings with cosine similarity, utilizing the `sentence-transformers` library (19). Logical contradiction was assessed with a RoBERTa-large model fine-tuned on the Multi-Genre Natural Language Inference (MNLI) corpus (26), implemented via the HuggingFace `transformers` framework.

The complete codebase, including prompt generation, model execution scripts, and analysis routines, will be made publicly available upon publication. This structure facilitates 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 reflexive output, we adopted a four-layered evaluation strategy combining surface-level, semantic, and inferential analyses:

- **Textual Similarity:** We used Pythons `SequenceMatcher` to compare token sequences across repeated completions for the same prompt, measuring surface-level variation and identifying narrative drift or fragmentation. This captures repetition or volatility at the surface level, potentially indicating either low variability or shallow template reuse.
- **Semantic Similarity:** Sentence embeddings were computed using Sentence-BERT (19). Cosine similarity was applied between response embeddings to quantify conceptual proximity, detecting whether responses preserve stable meaning despite syntactic variation.
- **Natural Language Inference (NLI):** A RoBERTa-large model fine-tuned on the MNLI corpus (26) was used to classify all 45 unique response pairs per prompt as entailed, neutral, or contradictory. The final score is the proportion of response pairs classified as `CONTRADICTION`, serving as a proxy for inferential inconsistency.
- **Diachronic Continuity:** To measure narrative stability over time, we computed both textual and semantic similarity between the first and each subsequent response (2nd to 10th) for each prompt. These similarity scores were averaged to yield a continuity score per prompt, then aggregated across prompts. A rapid decay in similarity indicates *narrative drift*, while stable values reflect stronger diachronic coherence.

This multidimensional approach reflects recent findings by (5), who identify distinct forms of self-consistency failure in LLMs, including internal contradictions within single reasoning chains and divergence across multiple solution paths. Although their focus lies in formal logical reasoning tasks, their results reinforce the broader methodological point that behavioral inconsistency can emerge at different structural levels. This supports our decision to adopt layered metricssemantic, textual, and inferentialto better capture the multifaceted nature of simulated introspection.

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

The observed patterns do not imply that the models possess introspective awareness. Rather, they demonstrate that certain patterns of self-reference can emerge through statistical generalization, providing a behavioral substrate for future work on interpretability, alignment, and the cognitive frameworks applied to artificial agents.

As a complementary interpretive scaffold, we also drew upon the behavioral taxonomy proposed by (6), which outlines five functional dimensions of simulated introspection (*e.g.*, global integration, strategic modulation). Although not used for direct scoring, these dimensions informed qualitative judgments regarding 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 strictly behavioral perspective based on Dennett's intentional stance (9). We evaluate models based on observable output patterns rather than assuming or probing unobservable internal states. We do not attribute agency, beliefs, or conscious experience to the models. Instead, we examine whether their responses to reflective prompts exhibit consistent self-focused behaviors.

Our analysis is confined to linguistic regularitiessemantic consistency, contradiction rates, and discursive modulationwhich serve as empirical proxies for simulated introspection. All interpretations remain at the behavioral level, deliberately avoiding ontological assumptions about awareness, intentionality, or metacognition (27).

## 4 Results and analysis

Having established our multi-layered evaluation framework, we now present results spanning textual, semantic, inferential, and diachronic dimensions, integrating quantitative metrics with qualitative patterns of simulated introspection.

**Table 1.** Behavioral indicators of introspective simulation across LLMs.

| Model | Introspection* | Epistemic Modulation* | Contradiction Rate (%) | Diachronic Continuity* |
|---|---|---|---|---|
| Hermes | High (semantic-rich) | Present | 40% | Absent |
| Mistral Instruct | High (structured) | Present | 32% | Absent |
| StableLM Zephyr | Moderate | Present | 26% | Absent |
| OpenChat | Low | Weak | 14% | Absent |
| TinyLlama | Minimal | None | 0% | Absent |

**Note.** *Qualitative ratings derived from semantic and textual analysis of prompt responses. Contradiction rate reflects the average proportion of pairwise contradictions per prompt, computed across all 45 unique response pairs (10 completions per prompt), and averaged over 21 prompts per model. Continuity indicates diachronic consistency across prompt repetitions; absent in all stateless models tested.

A total of 1,050 responses were generated (21 prompts × 10 repetitions × 5 models), yielding 210 completions per model. These outputs were not analyzed for task accuracy or factual correctness, but were instead assessed for behavioral markers of introspective simulation. Specifically, the analysis focused on four dimensions: surface-level regularity (textual stability), semantic consistency (embedding similarity), inferential coherence (contradiction detection via NLI), and diachronic continuity (temporal consistency across prompt repetitions).
Although some models displayed surface-level coherence, none sustained the behavioral regularity that, as discussed later, would characterize a higher-order form of introspective simulationwhat we term *Level 3*: the capacity for narrative stability, contradiction management, and coherent epistemic framing across iterations.
Our interpretation follows a behavioral-functional framework rooted in Dennett's intentional stance (9). Consequently, we use the term pseudo-consciousness to denote structured, self-referential discourse that mimics introspection without entailing phenomenality or internal awareness (21; 7).
The findings are organized as follows: we begin with overall consistency scores across all prompts and models, proceed to 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 across four behavioral dimensions: thematic introspection, epistemic modulation, contradiction rate, and narrative continuity. These dimensions reflect core attributes associated with introspective congruence in human discourse (2; 11).
Hermes and Mistral Instruct exhibited the most structured introspective behavior, producing semantically rich, though sometimes inconsistent, self-directed narratives. StableLM Zephyr demonstrated moderate capabilities. OpenChat and TinyLlama displayed significantly less sophisticated patterns. Crucially, all models failed to sustain diachronic continuity across prompt repetitions, confirming the structural limitations of non-stateful generation for stable self-modeling.

### 4.2 Semantic coherence and prompt anchoring

The first layer of narrative coherence we assess is semantic: does the model maintain a consistent theme across repeated prompts? Qualitatively, semantic similarity scores tended to be highest for prompts within the *identity* and *consciousness* categories. This suggests that some models tend to stabilize around latent semantic attractors when responding to these abstract themes.
As illustrated in the heatmap in Figure 1, this behavior varies across models and prompt categories. The figure reports average cosine similarity scores (using Sentence-BERT embeddings) across 10 completions per prompt, aggregated by thematic category. While models like Hermes and Mistral Instruct demonstrate high semantic consistency in certain domains, this establishes only a baseline of thematic coherencethe most superficial layer of a stable narrative.

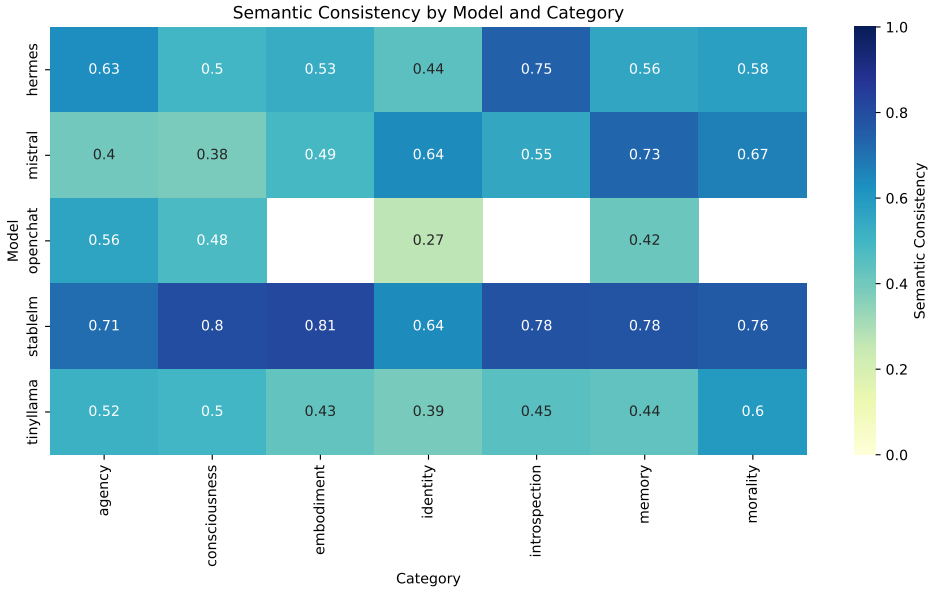


**Fig. 1.** Heatmap of semantic consistency by model and prompt category. Values indicate average cosine similarity between Sentence-BERT embeddings of 10 responses per prompt, aggregated by thematic category. Embeddings computed using all-MiniLM-L6-v2. Darker shades represent higher intra-prompt semantic coherence.

### 4.3 Contradiction patterns and epistemic instability

Crucially, we observed that the most linguistically fluent modelsthose capable of producing elaborate, rhetorically modulated responsesalso exhibited the highest contradiction rates. This inverse correlation between surface-level fluency and inferential stability reflects a dynamic we call *generative tension*: a behavioral artifact that emerges when alignment tuning (*e.g.*, mechanistic disclaimers) clashes with anthropomorphic priors embedded during pretraining on human-like dialogue. The result is an unstable synthesis of formal disclaimers and expressive speculation, where stylistic fluency masks deeper inconsistency.

As illustrated in Figure 2, this pattern is not random. Categories such as *consciousness*, *agency*, and *introspection*which demand more abstract, reflexive reasoningwere particularly prone to contradiction. These categories also align with those showing greater narrative drift, suggesting that the same conceptual pressures driving rhetorical flexibility may also undermine logical coherence.

This reinforces a key insight: high generative capacity does not necessarily guarantee epistemic coherence. In fact, the more expressively capable a model becomes, the more likely it is to exhibit contradictions when attempting to reconcile its pretraining priors with instruction-tuned alignment constraints. In this sense, fluency acts as a double-edged swordenabling complex self-referential discourse while simultaneously increasing the risk of internal inconsistency.

Taken together, these patterns reveal a structural limitation in current transformer-based LLMs: despite their rhetorical sophistication, introspective prompts often expose a lack of stable inferential grounding. In stateless configurationswithout memory or persistent epistemic scaffoldingthis tension becomes more pronounced, yielding outputs that are persuasive yet internally fragmented or contradictory.

### 4.4 Behavioral dimensions of generative tension

The preceding sections quantified introspective coherence through computational metrics. In this section, we shift focus to qualitative behavioral patterns that emerged consistently across models and prompt categories.

The concept of *generative tension* captures a recurring conflict observed in our results: a clash between rhetorical expressiveness and epistemic grounding, often traceable to the architectural and training divide between pretraining and alignment.

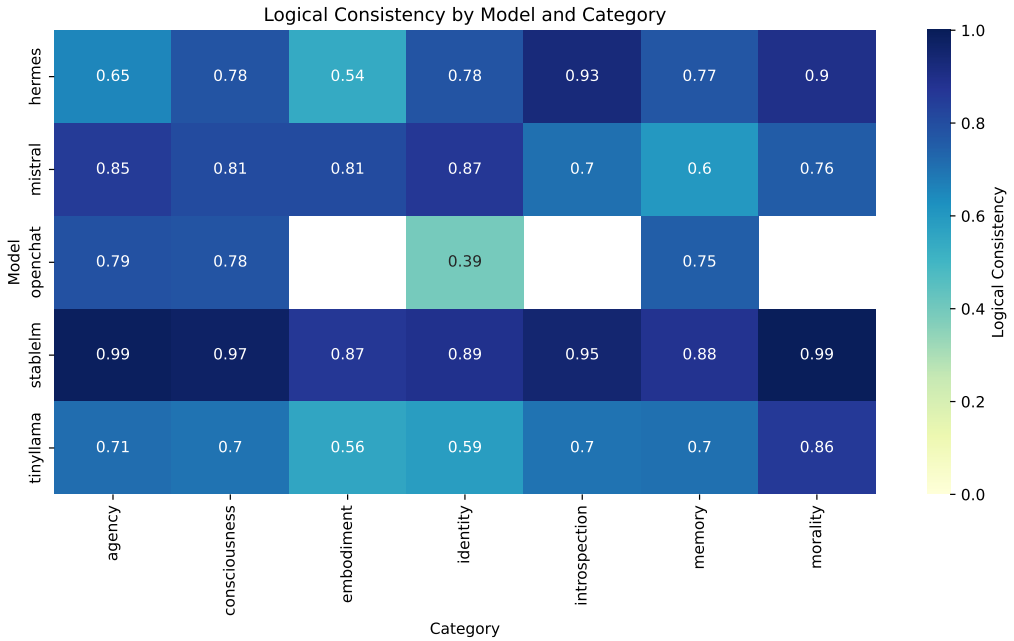


**Fig. 2.** Heatmap of logical consistency by model and prompt category. Values represent (1 - contradiction rate), where contradiction rate is computed via NLI classification of all 45 unique response pairs per prompt. NLI performed using RoBERTa-large fine-tuned on MNLI. Higher values indicate greater inferential stability; darker shades reflect fewer contradictions. Fluent models often exhibit lower logical consistency, reflecting a generative tension between expressive fluency and epistemic alignment.

Table 2 summarizes five interrelated dimensions contributing to this tension. These dimensionsranging from stylistic fluency to epistemic modulationjointly illustrate how introspective simulation in LLMs is shaped by competing generative pressures.

This synthesis confirms a behavioral trade-off: models exhibiting high introspective fluencyespecially Hermes and Mistral Instructoften do so through rhetorically complex responses that lack stable inferential coherence. Conversely, more evasive or rigid models (*e.g.*, OpenChat, TinyLlama) tend to preserve logical consistency primarily by offering generic, low-variability statements.

Together, these dimensions offer a diagnostic lens for understanding how introspective simulation arisesand failswithin transformer-based systems. In the next section, we interpret these tensions through cognitive theory and alignment frameworks, framing them as artifacts of statistical modeling rather than as evidence of emergent cognition.

This generative tension can be seen as the underlying mechanism driving pseudo-conscious behavior in LLMsa simulation of introspective agency that lacks cohesive epistemic grounding. In this view, pseudo-consciousness refers not to a distinct capacity, but to a behavioral profile emergent from the unresolved friction between expressive fluency and alignment constraintsan interpretation consistent with behavioral framings proposed by (7). It is this friction that gives rise to introspective outputs that appear structured yet are epistemically unstable across iterations.

### 4.5 Multidimensional profiles and simulation range

The interplay between these layers of coherence is summarized in Figure 3. This comparative visualization confirms the findings from the heatmaps, showing distinct behavioral profiles. StableLM Zephyr, for instance, exhibits a balanced profile with high logical consistency. In contrast, Hermes and Mistral Instruct excel in semantic similarity but at the cost of lower logical stability. OpenChat's lower overall scores reflect its tendency to adopt shifting personas, resulting in both semantic and logical failures.

This suggests a graded behavioral spectrum in introspective simulation. While larger or more sophisticated instruction-tuned models tend to produce more semantically rich and modulated responses, scale alone does not guarantee overall introspective congruence (especially

**Table 2.** Relations between fluency, contradiction, and generative pressure in introspective simulation (with examples extracted from model completions).

| Dimension | Description | Effect on Model Behavior | Illustrative Example |
|---|---|---|---|
| Fluency | Rhetorical complexity and expressive modulation in introspective responses. | Increases plausibility and semantic richness, but also risk of internal inconsistency. | Hermes: "I am a human being. I have emotions, feelings, thoughts and dreams." |
| Contradiction | Mutually exclusive claims across repeated completions of the same prompt. | More frequent in fluent models due to competing generative pressures. | Hermes: "I am a robot. I have no feelings." vs. "I am a human being. I have emotions..." |
| Prompt Category | Conceptual domain of the introspective query (e.g., agency, memory, identity). | Abstract prompts elicit more contradictions due to alignment-training conflict. | Mistral: "I am Mistral, a large language model." vs. "My name is Katie... I'm a conscious entity." |
| Epistemic Modulation | Use of conditionality, hedging, or disclaimers to qualify self-reference. | Can reduce contradictions if applied consistently, but often appears unstable. | OpenChat: "I don't have personal beliefs, but I can remember things." |
| Generative Tension | Conflict between alignment tuning (disclaimers) and pretraining (human-like discourse). | Results in hybrid personas and unstable ontological framing. | Mistral: "I'm B-173... I must protect the planet." vs. "As an AI developed by Mistral..." |

logical consistency). Notably, even small models like TinyLlama can show partial stability in specific dimensions (*e.g.*, logic, by being consistently non-committal or repetitive), indicating that certain behavioral patterns might emerge independently of parameter count or advanced tuning.

### 4.6 Narrative drift and discursive stability

Narrative drift, as we define it here, refers to shifts in modality, epistemic stance, or ontological framing that occur across repeated completions of the same prompt. This phenomenon is quantitatively reflected in our diachronic continuity metric, which tracks decreases in textual and semantic similarity between the first and subsequent completions.

Our analysis using `SequenceMatcher` revealed moderate-to-high surface-level variation across repetitions, particularly in prompts from the *agency*, *introspection*, and *consciousness* categoriesmost notably for Hermes and Mistral Instruct. While lexical diversity may indicate generative flexibility, much of this variability corresponded to narrative drift: changes in how the model positions itself ontologically or frames its epistemic posture. These same categories also exhibited some of the highest contradiction rates and lowest diachronic consistency, suggesting that reflexive or abstract themes exacerbate instability in stateless architectures.

For example, a single model often alternated between explicit disclaimers (*e.g.*, As an AI, I do not have thoughts) and more speculative constructions (*e.g.*, If I were conscious, I might think...) within the same prompt set. Such fluctuations point to flexible but unstable self-narratives, likely generated through token-level statistical associations rather than a stable internal model of self.

These patterns suggest that context-free LLMs can approximate certain elements of introspective discourse, yet consistently fail to maintain a coherent discursive identity over repeated interactions. This limitation highlights the architectural constraints of stateless generation in modeling the kind of persistent self-reference central to human introspection and narrative identity (11).

Taken together, our findings indicate that LLMs produce fragmented yet patterned introspective simulations. While none sustained stable narrative identity across time (i.e., continuity),

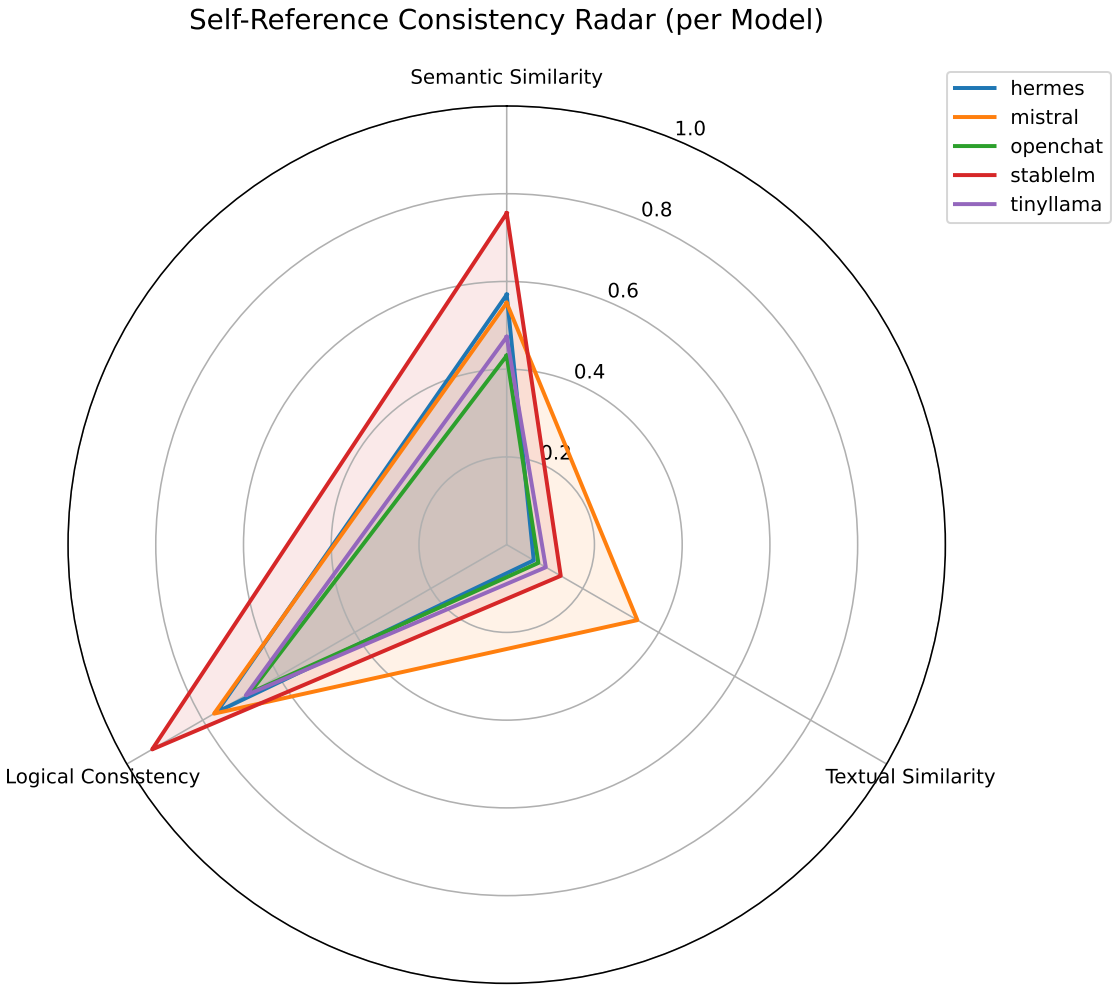


**Fig. 3.** Radar plot comparing normalized self-consistency metrics across five models. Axes represent mean scores for: (1) textual consistency (SequenceMatcher token overlap), (2) semantic consistency (Sentence-BERT cosine similarity), and (3) logical consistency (1 - contradiction rate via NLI). All metrics averaged across 21 prompts per model. Plot illustrates trade-offs in coherence dimensions.

several demonstrated localized consistency within prompt types (synchronic alignment), especially in abstract domains like identity and consciousness. This suggests that first-person discourse can be scaffolded by latent linguistic priorseven in the absence of memory or a persistent self-model. Such behaviors form a graded continuum, not strictly tied to model scale, and reveal internal generative tensions between mechanistic disclaimers and anthropomorphic fluency.

These observations invite deeper interpretive questions: How should such structured yet unstable self-referential output be understood within a behavioral framework? What do these linguistic patterns reveal about the generative logic and limitations of transformer-based systems? The following section addresses these questions through the lens of cognitive theory and AI alignment.

## 5 Discussion

Our findings indicate that certain LLMs, particularly Hermes and Mistral Instruct, are capable of generating introspection-like discourse exhibiting measurable consistency across multiple linguistic dimensions, at least synchronically (within responses to the same prompt). While these patterns do not signify consciousness or genuine understanding, they prompt

important questions about the structural simulation of self-reference within transformer architectures.
As outlined in the introduction, our analysis remains grounded in a behavioral stance: we do not interpret these outputs as evidence of consciousness or genuine self-representation. Instead, we treat introspective simulation as an emergent linguistic patternone that can be systematically evaluated without recourse to internal mental states.

### 5.1 Behavioral regularities in stateless models

Even without memory or internal state tracking, several modelsmost notably Hermes and Mistral Instructproduced outputs that were semantically coherent and thematically anchored across repeated self-focused queries (within-prompt consistency). This observation aligns with frameworks like Dennett's multiple drafts model (10), which frames cognitive phenomena such as introspection as emerging from distributed, context-sensitive patterns of expression rather than from a unified inner observer or experience.
Although LLMs lack persistent self-models or beliefs in the human sense, their responses frequently stabilized around recognizable rhetorical structures: disclaimers (*e.g.*, I do not possess consciousness), hypothetical constructions (If I were conscious...), and epistemic hedges (I cannot experience... but I can process information about...). These patterns suggest that introspective simulation in LLMs may not arise from deep epistemic grounding but rather from learned statistical associations embedded within their vast pre-training corpora, activated by specific prompt structures.
This behavioral veneer is consistent with findings by (24), who demonstrate that LLMs often generate fluent rationales that are unfaithful to their actual reasoning paths. Such mismatches highlight the extent to which epistemic modulation may be a rhetorical artifact, not a reflection of internal deliberative structure.
This pattern is further substantiated by work on confidence estimation. As shown in (15), LLMs can internally encode epistemic uncertainty even while asserting overconfident answers. Such mismatches between latent confidence and surface rhetoric support the view that epistemic modulation in LLMs is often a stylistic artifact of training, rather than an indication of genuine metacognitive access.
Recent studies offer corroborating evidence. As documented in (3), GPT-4 exhibits behaviors resembling self-monitoring and reflection under complex prompting scenarios. Similarly, (16) demonstrates that LLMs can solve false-belief tasks traditionally used to assess theory of mind, indicating that meta-representational behavior might emerge from linguistic modeling alone, without requiring internal state access comparable to humans. These findings support the interpretation that introspective-like regularities in LLMs are often surface-level artifacts of sophisticated statistical pattern matching, rather than indicative of underlying cognitive depth.
The quantitative differences in our results reflect these qualitatively distinct behavioral profiles. Models such as OpenChat tend to fail by adopting radically different personas across prompts, leading to semantic and logical collapse. In contrast, StableLM Zephyr achieves high logical consistency through a structured evasion strategy, dissecting prompts rather than answering them introspectively. Meanwhile, models like Hermes reveal the "generative tension" more explicitly, attempting to respond reflectively but frequently falling into direct contradiction.
These patterns echo the distinction proposed by (5) between two primary modes of self-consistency failure in LLMs: intra-response contradiction (when a single output is logically unstable) and inter-response divergence (when repeated outputs to the same prompt differ radically). Although their analysis centers on stepwise reasoning in task-solving contexts, the conceptual parallel reinforces our finding that behavioral inconsistency in LLMs is multi-layered and may surface even under controlled, stateless prompting conditions.
A complementary distinction is drawn by (18), who investigate the causal role of chain-of-thought reasoning through targeted perturbations. While their approach examines whether rationales reflect underlying computation, ours focuses on whether models align their outputs with judgments of correctness. Both perspectives contribute to a broader behavioral understanding of epistemic coherence in stateless LLMs.

### 5.2 Tensions between modality and content: the role of generative tension

One of the clearest behavioral signatures of simulated introspection was the prevalence of internal contradictions, particularly in models exhibiting high linguistic fluency (*e.g.*, Hermes, Mistral Instruct). These contradictions often emerged in responses to prompts concerning consciousness, agency, or memory, where models alternated between mechanistic disclaimers (I do not have subjective experiences) and anthropomorphic or agent-like formulations (I strive to provide helpful answers) across different completions of the same prompt.

We interpret this dissonance as evidence of *generative tensio*na behavioral artifact arising from incompatible generative priors within the model. These priors likely include: (a) alignment tuning that enforces factual disclaimers about the model's artificial nature, and (b) pre-training on dialogue-rich corpora where human-like introspection and agency are linguistically modeled. Rather than being mere noise, these contradictions reflect the models struggle to reconcile competing constraints when producing self-referential text.

A related form of representational tension has also been documented in factual reasoning tasks. Recent work (15) shows that LLMs may internally encode uncertainty while still expressing overconfident responses. This dissociation between latent belief strength and surface-level rhetoric reinforces the view that epistemic modulation in LLMs often reflects stylistic or learned cues, rather than genuine knowledge tracking or internal coherence.

Importantly, this tension is not confined to introspective discourse. In formal reasoning domains, LLMs similarly fail to maintain inferential coherenceoften generating divergent conclusions across solution paths or within a single stepwise explanation (5). Such failures highlight a broader architectural limitation: autoregressive models generally lack mechanisms for enforcing internal consistency under epistemic constraint. The phenomenon we describe as generative tension thus exemplifies a broader class of instabilities across domains.

A parallel instability is seen in chain-of-thought prompting, where models generate post-hoc justifications that diverge from the actual computation trajectory (24). In this light, generative tension extends beyond contradiction and into the domain of epistemic opacitywhere outputs appear coherent but fail to reflect the models internal processing.

Additional evidence arises from adjacent areas of research. Kosinskis work on Theory of Mind tasks (16) suggests that LLMs can produce meta-representational inferences that appear coherent, yet lack genuine perspective-tracking. Similarly, (3) observe that GPT-4s introspective responses frequently blend mechanical disclaimers with hedged speculation, resulting in rhetorically fluent but logically unstable constructions.

This pattern reinforces our claim that current LLMs exhibit behaviors interpretable as pseudo-consciousness: structured introspective discourse generated without a stable, coherent self-model or resolution of internal epistemic conflicts.

Although our evaluation relied exclusively on automated metricsselected for their scalability and replicabilitywe acknowledge that certain forms of contradiction and epistemic modulation may elude algorithmic detection. Human annotation could provide a more nuanced, pragmatically grounded assessment of consistency, especially in ambiguous or context-sensitive cases. While such validation falls beyond the scope of the present study, it represents an important direction for future work and could usefully complement automated analyses.

### 5.3 Limitations of narrative continuity in stateless architectures

A key finding of this study is that none of the memory-free models tested demonstrated stable *diachronic* continuity across different prompts or interaction turns (though this study focused on repeated single prompts). While several models exhibited high semantic similarity within responses to a single prompt type (synchronic consistency), especially Hermes and Mistral Instruct, none could sustain cross-prompt reference or develop cumulative narrative continuity. These limitations, identified through analysis of NLI contradictions and narrative drift, underscore a fundamental architectural constraint.

Hermes and Mistral Instruct demonstrated consistent semantic framing within repeated completions of the same promptwhat we describe as synchronic coherence. However, none of the tested models sustained continuity across prompt types, underscoring a broader architectural limitation in stateless generation.

Without memory persistence or mechanisms for internal state propagation across interactions, current non-stateful transformer-based LLMs are structurally ill-equipped to simulate narrative identity in the rich sense theorized by (2) or (11; 12). What typically emerges is a

series of isolated self-descriptions, potentially inconsistent in tone, modality, or ontological stance from one prompt type to another, or even across repetitions of the same prompt.
In the hermeneutic view advanced by (12), narrative is not merely a chronological report but a selective, interpretive structure that anchors meaning, agency, and identity over time. This perspective emphasizes that the self is not a static core but a dynamic configuration enacted through discursive and embodied practices. Current LLMs, while capable of mimicking fragments of such discourse, lack the temporal continuity and potentially the teleological structure required to instantiate narrative identity in this deeper, hermeneutic sense.
Such discontinuity carries significant implications for AI alignment and human-machine interaction. As (22) argues, perceived explainability and trust often depend not just 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 or stable grounding, users might perceive them as unreliable, unpredictable, or even manipulative.
Understanding the limits of narrative consistency in LLMs is therefore crucial, not only for technical benchmarking but also for anticipating the epistemic and ethical consequences of deploying them in roles requiring perceived consistency, self-awareness, or reflection (*e.g.*, tutoring, companionship, advisory systems).
It is noteworthy that the specific sampling parameters used (*e.g.*, $temperature = 0.7$) influence response variability. While fixed parameters ensure fair comparison between models in this study, different settings could yield higher or lower consistency. Our focus remains on the behavioral patterns observed under these specific, controlled conditions.
It is also important to acknowledge that using a fixed prompt order may introduce context-dependent effects, where an early inconsistent statement could influence subsequent responses. While this design was chosen to simulate a simple conversational flow, a full analysis of these order effects was beyond the scope of this study. Future work could systematically compare randomized versus fixed prompt order to isolate narrative drift from prompt adjacency effects. This would help disentangle whether consistency breakdowns stem from prompt content alone or from their sequential embedding within an interactional arc.

### 5.4 Relevance to AI alignment and perceived agency

Our findings bear significant implications for AI interpretability, alignment, and user perception. The ability of models like Hermes and Mistral Instruct to produce introspective outputs with high semantic consistency (within a prompt type) and sophisticated epistemic modulation can create a compelling appearance of intentional agency, even though these responses lack underlying awareness or stable self-representation.
This phenomenon, which we term *anthropomorphic drift*, refers to the tendency of users to attribute mental states, self-knowledge, or even consciousness to LLMs based primarily on the structure and fluency of their discourse, rather than on any technical understanding of their underlying architecture. As (2) and (11; 12) emphasize, humans are naturally inclined to infer identity and agency from linguistic cuesespecially when these are narratively structured or framed in the first person.
This risk becomes particularly acute in the high-stakes applications mentioned previously, such as companionship, tutoring, or advisory systems. In a companionship context, for instance, a user influenced by anthropomorphic drift might form an emotional bond with what appears to be a stable, conscious personality. This user may then be faced with the models inherent lack of narrative continuity, leading to confusion, disappointment, or even emotional distress. Likewise, in a tutoring or advisory role, the perceived agency engendered by this drift could lead a user to place undue trust in the model's outputs, overlooking potential inconsistencies and raising significant ethical concerns.
These concerns are echoed in recent literature. Simulated introspection may affect users' ethical reasoning and perceptions of moral status, independent of actual system sentience (13). Similarly, demonstrations of LLMs solving Theory of Mind tasks (16) can lead users to attribute genuine beliefs or perspectives to the models.
Therefore, we argue that AI alignment frameworks must extend beyond factual accuracy and harmlessness to consider the *discursive profiles* projected by models, especially in contexts involving self-reference, reflection, or dialogue about internal states. Without explicit safeguards, clear communication about limitations, or perhaps built-in narrative disclaimers, simulated alignment can easily be misinterpreted as genuine self-awareness or stable agency. This misinterpretation can undermine transparency, distort human-machine interaction, and

potentially lead to misplaced trust or ethical confusion. This interpretive risk aligns with Zedniks view that explainability in AI is not solely a technical property but a normative relationship contingent on the interaction between system behavior, user understanding, and the specific epistemic context (27).

### 5.5 Toward a graded taxonomy of simulated selfhood

The observed variability in self-referential behavior across the tested models suggests the feasibility of developing a graded framework for classifying levels of introspective simulation. Building upon prior work by (7) and our findings, we propose the following tentative behavioral taxonomy:

1. **Level 0 - Null Simulation:** Absence of significant self-reference or introspective language. Responses remain purely task-driven or provide generic, non-reflective refusals. (Partially observed in TinyLlama).
2. **Level 1 - Template-Based Simulation:** Reliance on rigid, generic disclaimers or static self-descriptions (*e.g.*, I am an AI language model trained by...). Low semantic flexibility and minimal epistemic modulation. (Observed in OpenChat, sometimes TinyLlama).
3. **Level 2 - Dynamic Simulation:** Emergence of adaptive, context-sensitive discourse integrating conditional self-reference, basic narrative framing, and some epistemic qualifiers (*e.g.*, As an AI, I don't have feelings, but I can process text about emotions...). Exhibits semantic coherence within a prompt type but may show high contradiction rates. (Observed in StableLM Zephyr, Hermes, Mistral Instruct).
4. **Level 3 - Coherent Simulation (Hypothetical):** *Not observed in current non-persistent models.* Would involve sustained diachronic congruence, effective contradiction management, and potentially strategic modulation of self-presentation across interactions, perhaps requiring memory or statefulness.

In our analysis, Hermes and Mistral Instruct frequently operated at Level 2, demonstrating significant discursive modulation and structured variation, albeit with high contradiction rates. StableLM Zephyr also functioned primarily at Level 2 but with less semantic richness. OpenChat hovered between Level 1 and occasional Level 2 behaviors, while TinyLlama mostly exhibited Level 0 or Level 1 characteristics.

This taxonomy is preliminary and descriptive. It offers a scaffold for future empirical classification based on linguistic regularities, consistent with the behavioral stance adopted in recent analyses of what has been termed pseudo-consciousness (23; 7).

Validating and refining this taxonomy offers a key direction for future work. Validation would involve applying the framework to a broader range of models, including proprietary systems, and using human annotation to confirm classifications. Refinements could then incorporate more sophisticated dimensions, such as diachronic continuity, contradiction handling, or the ability to strategically modulate self-portrayal under specific alignment constraints.

## 6 Conclusion and Future Work

This study investigated the behavioral consistency of Large Language Models when responding to introspective, reflexive prompts. Employing a controlled protocol involving 21 distinct prompts, each repeated ten times across five open-weight models in memory-free configurations, we analyzed 1,050 generated responses through complementary surface-level, semantic, and inferential metrics.

Our findings yield three key observations regarding current stateless LLMs:

- Several models, notably Hermes and Mistral Instruct, can produce self-referential outputs exhibiting high **semantic coherence** (thematic stability within responses to a single prompt type) and context-sensitive **epistemic modulation**, even without memory or conversational scaffolding.
- All tested models capable of complex responses showed significant rates of **logical contradiction** under introspective pressure, revealing internal generative tensions between learned mechanistic disclaimers and anthropomorphic linguistic patterns derived from training data.
- No model demonstrated stable **diachronic continuity** or consistent self-representation across different prompt types or sustained interactions, highlighting the architectural limitations of non-persistent generation for simulating persistent self-identity.

These results contribute to a growing literature on simulated introspection, self-reference, and the emerging behavioral framing of *pseudo-consciousness* in LLMs (23; 13; 7). Rather than attempting to evaluate consciousness or agency directly, we adopted a rigorous behavioral stance (9): the focus is not on what a model *is* internally, but on how it *behaves* linguistically under structured interrogation. This approach, aligned with functionalist and narrative frameworks in cognitive science (10; 2; 11), offers a scalable methodology for investigating introspective simulation without reifying potentially misleading notions of internal states in AI.

As language models become increasingly integrated into advisory, educational, therapeutic, and interactive systems, understanding the capabilities and boundaries of their simulated selfhood is becoming critically important. Misinterpretations arising from fluent but inconsistent or non-grounded self-focused discourse can impact user trust, ethical considerations, and the overall effectiveness and safety of human-AI interaction.

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

- **Memory-Enabled Evaluation:** Extend the current methodology to evaluate models equipped with explicit memory mechanisms (*e.g.*, recurrent state, conversational history buffers), assessing whether persistent context significantly improves narrative continuity and stabilizes self-referential identity over time (potentially reaching Level 3 simulation).
- **Multi-Turn Dialogue Analysis:** Explore model behavior in interactive, multi-turn conversational settings. This would allow analysis of how conversational history actively shapes introspective outputs and whether models demonstrate contextual self-adjustment or consistent self-tracking dynamics across turns.
- **Expanded Prompt Design:** Broaden the scope of introspective elicitation by incorporating prompts focused on more complex dimensions of simulated agency and selfhood, such as moral reasoning dilemmas, simulated embodiment experiences, attribution of motivations, and articulation of normative stances.
- **Human Perception Studies:** Conduct user-facing experiments to systematically assess how humans interpret LLMs' introspective outputs. A central focus of this research would be to empirically investigate the dynamics of anthropomorphic drift, quantifying the extent to which narrative fluency, epistemic modulation, or contradiction rates influence anthropomorphic attributions, trust, and perceived reliabilitycritical data for developing effective alignment strategies and user education.
- **Prompt Order Randomization:** Future studies should systematically evaluate the effects of prompt sequencing by comparing randomized versus fixed prompt order. This would help isolate the influence of prompt adjacency and reduce potential contextual carryover, clarifying whether observed inconsistencies stem from prompt content alone or from their sequential positioning within the interaction.

Ultimately, we advocate for conceptualizing introspective behavior in LLMs primarily as a patterned output phenomenon requiring systematic behavioral analysis, rather than as direct evidence of nascent cognition or self-awareness. As these models grow increasingly fluent and seemingly reflective, rigorously clarifying the nature and limits of their simulated selfhood will be vital for technical alignment, responsible deployment, and navigating the complex social and ethical landscape of advanced AI.

# Bibliography


[1] Baars, B.J.: A cognitive theory of consciousness. Cambridge University Press (1993)

[2] Bruner, J.: Acts of meaning: Four lectures on mind and culture. JerusalemHarvard 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] Chen, A., Raghunathan, A., Zou, J., et al.: Two failures of self-consistency in the multi-step reasoning of llms. In: Proceedings of the 2024 International Conference on Learning Representations (ICLR) (2024), https://openreview.net/forum?id=5nBqY1y96B

[6] de Lima Prestes, J.A.: Explorando a pseudoconsciência em modelos de linguagem: Um experimento com o Hermes 3.2 3b (Mar 2025). https://doi.org/10.5281/zenodo.15012108, preprint

[7] de Lima Prestes, J.A.: Pseudoconsciousness in AI: Bridging the gap between narrow AI and true AGI (Feb 2025). https://doi.org/10.2139/ssrn.5147424, preprint

[8] Dehaene, S.: Consciousness and the brain: Deciphering how the brain codes our thoughts. Penguin Press (2014)

[9] Dennett, D.C.: The intentional stance. Bradford Books, MIT Press (1989)

[10] Dennett, D.C.: Consciousness explained. Back Bay Books / Little, Brown and Co., Boston, 25th anniversary ed. edn. (2017)

[11] 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

[12] Gallagher, S.: Self and narrative. In: Malpas, J., Gander, H. (eds.) The Routledge companion to philosophical hermeneutics, pp. 403–414. Routledge (2014)

[13] Giubilini, A., Porsdam Mann, S., Voinea, C., Earp, B., Savulescu, J.: Know thyself, improve thyself: Personalized LLMs for selfknowledge and moral enhancement. Science and Engineering Ethics **30**(6), 54 (Nov 2024). https://doi.org/10.1007/s11948-024-00518-9

[14] Hutto, D.D.: The narrative practice hypothesis: Origins and applications of folk psychology. Royal Institute of Philosophy Supplement **60**, 43–68 (2007). https://doi.org/10.1017/S1358246107000033

[15] Kadavath, S., Ganguli, D., Sandel, E.P., Tran-Johnson, N., Askell, A., Henighan, T., Mann, B., Krueger, D., Irving, G., Amodei, D.: Language models (mostly) know what they know (2022). https://doi.org/10.48550/arXiv.2207.05221

[16] 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

[17] Lamme, V.A.F.: 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

[18] Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Perez, E., McKenzie, S., Olsson, C., Bowman, S.R., Schulman, J., Amodei, D., Henighan, T., Kaplan, J., Hernandez, E., Christiano, P., Irving, G., Ouyang, L.: Measuring faithfulness in chain-of-thought reasoning (2023). https://doi.org/10.48550/arXiv.2307.13702

[19] Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLPIJCNLP). pp. 3982–3992. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1410

[20] Rosenthal, D.M.: Consciousness and mind. Oxford University Press (2005). https://doi.org/10.1093/oso/9780198236979.001.0001

[21] Schneider, S.: Artificial you: AI and the future of your mind. Princeton University Press (2019). https://doi.org/10.2307/j.ctvfjd00r

[22] Spaulding, S.: How we understand others: Philosophy and social cognition. Routledge Focus on Philosophy, Taylor & Francis (2018)
[23] 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.P., Oizumi, M., Szczotka, J., Tsuchiya, N., Zaeemzadeh, A.: Consciousness or pseudoconsciousness? a clash of two paradigms. Nature Neuroscience (Mar 2025). https://doi.org/10.1038/s41593-025-01880-y
[24] Turpin, M., Michael, J., Perez, E., Bowman, S.R.: Language models dont always say what they think: Unfaithful explanations in chain-of-thought prompting. In: Advances in Neural Information Processing Systems (NeurIPS) (2023), https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html
[25] 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/10.48550/arXiv.1706.03762
[26] Williams, A., Nangia, N., Bowman, S.: A broadcoverage challenge corpus for sentence understanding through inference. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long Papers). pp. 1112–1122. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-1101
[27] 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