Simulated Selfhood in LLMs: A Behavioral Analysis of Introspective Coherence
(Preprint Version 4 - April 2, 2026)

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 stable underlying patterns or merely surface-level generative artifacts. We evaluated five open-weight, stateless LLMs using a structured battery of 21 introspective prompts. The main corpus comprised 1,050 completions collected under a baseline decoding condition (`temperature` = 0.7), supplemented by 2,100 additional completions generated under matched temperature conditions (`temperature` = 0.2 and 1.0), for a total of 3,150 completions. 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). Construct validity was further examined through a human-evaluation layer in which 10 annotators rated 80 selected response pairs drawn from the same prompt battery on a 5-point consistency scale. The annotation task was perceived as low-to-moderate in difficulty (mean self-reported difficulty = 2.6/5). Inter-rater agreement was moderate by Krippendorff's α for ordinal ratings ($\alpha = 0.553$), while reliability was moderate at the single-rater level and strong for aggregated ratings by intraclass correlation (ICC(2,1) = 0.564; ICC(2,k) = 0.928). The human-evaluation layer showed that lexical overlap and embedding-based semantic similarity were weak proxies for perceived self-referential consistency, whereas NLI-based indicators tracked mean human ratings much more closely. Across the matched temperature conditions, lower temperature generally increased semantic and diachronic stability, whereas higher temperature tended to increase drift and reduce coherence, though the pattern was not perfectly monotonic across all models or metrics. We therefore interpret apparent self-referential stability in stateless LLMs as conditional and fragile rather than robustly stable across generation regimes. 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 behavioral framework, complemented by human validation and a matched decoding-temperature sensitivity analysis, for evaluating simulated introspection in LLMs. Our findings carry implications for 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.

⋆ Preprint Version 4 (April 2, 2026). This version supersedes the author's earlier preprint versions and reflects both the maturation of the underlying argument and the subsequent development of the relevant literature. In comparison with the previous versions, it incorporates substantial methodological and interpretive revisions, including an expanded empirical design, matched decoding-temperature analyses, a human-evaluation layer, and updated engagement with recent debates on introspective simulation, pseudo-consciousness, interpretability, and AI alignment. The manuscript has been submitted to a journal and is currently under peer review. It may therefore be revised further following editorial assessment and peer review. Readers are advised to consult the most recent version for citation and interpretive purposes.

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 (5; 30). 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 (7; 5; 10). This inconsistency implies that current models may possess only shallow or fragmented self-models, undermining their capacity to maintain coherent self-narratives (7). 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 (30; 13).

Furthermore, (17) 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 identity, internal state, memory, agency, embodiment, morality, and introspective reflection. We evaluated five open-weight, transformer-based models using a battery of 21 reflexive prompts. The main corpus comprised 1,050 completions generated under a baseline decoding condition (`temperature` = 0.7), and this corpus was supplemented by 2,100 additional completions generated under matched temperature conditions at `temperature` = 0.2 and 1.0, for a total of 3,150 completions. This design made it possible to examine both behavioral regularities within a baseline condition and the extent to which those regularities persisted across decoding regimes.

The resulting outputs were analyzed using four complementary methods:

- **Textual Similarity:** Surface-level variation was quantified using Python's `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 (27).
- **Logical Contradiction:** Inferential consistency was assessed using a RoBERTa-large model fine-tuned on the Multi-Genre Natural Language Inference (MNLI) corpus (35) to classify response pairs as entailing, neutral, or contradictory.
- **Diachronic Continuity:** Temporal stability was estimated by comparing the first completion of a prompt with the subsequent completions produced under the same stateless condition.

To strengthen construct validity, we also incorporated a human-evaluation layer in which 10 annotators rated 80 selected response pairs drawn from the same prompt battery. This human-evaluation layer complements the computational analyses by introducing direct human ratings of pairwise self-referential consistency.

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 such as Dennett's intentional stance, focusing on patterns in external behavior rather than internal states.

By analyzing these dimensions, our study argues for a more conditional interpretation of introspective simulation in stateless LLMs. Some models exhibit local semantic stability or thematic anchoring, especially under lower-temperature decoding, but these regularities prove fragile when examined through contradiction, diachronic continuity, the human-evaluation layer, and the comparison across matched temperature conditions. We therefore interpret

apparent self-referential stability not as robust self-representation but as a contingent behavioral effect of prompt structure, decoding regime, and learned linguistic priors.
This paper proceeds as follows: Section 2 reviews related work in introspective simulation and AI consistency. Section 3 details our behavioral methodology, including the human-evaluation layer and a matched decoding-temperature sensitivity analysis. Section 4 presents the results of our empirical analysis. Section 5 discusses the implications of these 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 (12; 13; 30). At the same time, recent work shows that LLMs can produce coherent, goal-directed responses under introspective pressure, prompting renewed questions about how such patterns should be described, evaluated, and interpreted (5; 17; 33).
In this context, the term *pseudo-consciousness* has increasingly been used as a behavioral descriptor for structured, self-referential outputs observed in context-free models. (33) distinguish pseudo-consciousnessunderstood as linguistic fluency potentially lacking causal integrationfrom genuinely conscious systems, cautioning against conflating simulation with intrinsic awareness. Similarly, (17) argue that even non-conscious systems may shape moral perception through simulated introspective discourse. Taken together, these works support the use of metaphysically restrained concepts for analyzing self-referential language in LLMs.
Recent scholarship has also shifted from primarily conceptual discussion toward more explicit validation frameworks for psychologically or morally charged LLM outputs. Methodological work on morally inflected behavior argues for structured evaluation of normative competence rather than impressionistic interpretation based on fluent surface form alone (18). Likewise, psychometric approaches to personality-like regularities in LLM outputs foreground reliability and construct validity when latent psychological traits are inferred from generated language (31; 36). Although these studies do not focus directly on self-reference, they reinforce a broader methodological point that is central to the present study: psychologically suggestive regularities in LLM outputs should be validated rather than assumed from face validity alone.
A related line of inquiry concerns whether increasingly structured first-person discourse in LLMs warrants stronger claims about intentionality or consciousness. (16) argue that referential grounding remains central to any serious assessment of intentionality in LLMs, cautioning against inferring genuine aboutness from fluent self-description alone. In a closely related preprint, (3) report that self-referential prompting can elicit structured first-person reports of subjective experience across model families, while explicitly refraining from treating such reports as evidence of consciousness. These studies underscore the importance of distinguishing reproducible self-referential output patterns from ontological claims about subjective awareness.
Within this broader context, (10) proposed a behavioral taxonomy of introspection-like outputs in LLMs, identifying features such as thematic self-reference, epistemic modulation, and contradiction management. A subsequent article applied this conceptual model to Hermes-3 Llama 3.2B, articulating five behavioral dimensions of introspective simulation (9). Although heuristic, this framework remains useful for identifying recurrent linguistic structures in reflexive LLM output and for motivating the present study's behavioral focus on stability, contradiction, and narrative drift.
Further work supports the relevance of self-referential consistency as a broader behavioral problem rather than one limited to explicitly introspective prompts. Studies showing that LLMs can solve false-belief tasks traditionally used in Theory of Mind research suggest the emergence of linguistic behaviors structurally aligned with mental-state attribution (22). Similarly, (20) show that LLMs can calibrate confidence with surprising accuracy, indicating that epistemic modulation may emerge from internal statistical signals even in the absence of explicit self-reference. More generally, (8) identify distinct forms of self-consistency failure in multi-step reasoning, including contradictions within a single reasoning chain and divergence across alternative reasoning paths. Although situated in formal reasoning tasks, this work

reinforces the broader point that language models often fail to maintain stable commitments across outputs.

These questions are also shaped by how humans interpret apparently coherent discourse. Bruner's narrative identity model and Dennett's intentional stance provide interpretive scaffolds for understanding why structured first-person language may be read as agent-like even in the absence of genuine mentality (4; 12). Relatedly, (32) and (15) emphasize that narrative scaffolding plays a central role in the social attribution of mind. This helps explain why coherence in linguistic form may suffice to evoke perceived intentionality, even when underlying self-representation remains unstable or absent.

Recent philosophical critiques further underscore the need for caution. (37) argues that explainability in AI is observer-relative, highlighting that models may produce linguistically coherent outputs without satisfying stronger normative standards of epistemic transparency. A closely related methodological concern appears in work on explanation fidelity under chain-of-thought prompting, where LLMs can generate rationales that appear coherent yet fail to reflect the reasoning processes that produced the final answer (34). Extending this concern, (25) show that minimal interventions to a model's articulated reasoning often leave final answers unchanged, suggesting that such rationales may function as post-hoc justifications rather than causal explanations. While their focus is causal faithfulness rather than self-reference, the methodological lesson is closely related to ours—coherent surface language should not automatically be taken as evidence that the underlying property of interest has been captured.

This study builds upon and extends these perspectives by analyzing introspective simulation across five open-weight models using textual, semantic, inferential, and human-evaluated measures. In contrast to prior work focused primarily on phenomenology or ontology, we frame the problem strictly in behavioral terms: whether LLMs can sustain consistent, structured discourse about themselves across repeated stateless completions, and which evaluative indicators most plausibly track that behavior.

3 Methodology

This study develops a behavioral framework for evaluating how LLMs respond to introspective, self-directed prompts. The central objective is to examine whether repeated responses in stateless settings exhibit stable or unstable patterns of self-reference when analyzed at the level of observable linguistic behavior. Rather than treating such outputs as evidence of inner mental states, the study evaluates them through measurable dimensions of consistency, including semantic similarity, inferential contradiction, and discursive modulation.

Methodologically, the study is restricted to externally accessible behavior. It does not seek to infer latent consciousness, subjective experience, or genuine agency from model outputs. Instead, it asks whether LLMs produce recurrent forms of self-description that can be systematically compared across repeated completions and decoding conditions. This restriction is especially important in the case of stateless models, where apparent continuity must be assessed from output regularities alone rather than from any persistent internal self-model.

Within this methodological frame, the term pseudo-consciousness is used as an operational descriptor for structured but non-experiential self-referential discourse (9). The term functions as a label for a class of observable linguistic behaviors and does not imply synthetic awareness, phenomenology, or ontological parity with human consciousness.

3.1 Philosophical and computational grounding

The broader conceptual justification for this approach comes from Dennett's intentional stance (12). On this view, systems may be interpreted in agent-like terms when they display coherent, goal-directed, or self-descriptive behavior, even if no claim is made about inner subjectivity. This makes the intentional stance a suitable philosophical basis for studying introspection-like discourse in LLMs: it permits analysis of apparently minded behavior while withholding ontological commitment. In this respect, the framework also converges with observer-relative accounts of social cognition and explainability, which treat interpretations of mindedness as grounded in publicly available patterns rather than privileged access to internal states (32; 37).

The computational grounding of the study is therefore not based on assumptions about hidden cognitive architecture, but on the systematic analysis of textual behavior under controlled experimental conditions. To clarify the formal character of some observed patterns,

we draw limited analogical support from theories of consciousness in cognitive neuroscience, including Global Workspace Theory (GWT) (2; 11), Recurrent Processing Theory (RPT) (24), and Higher-Order Thought (HOT) theory (29). These theories are not invoked to suggest that transformer-based LLMs instantiate biological or phenomenological consciousness. Rather, they provide a comparative vocabulary for describing output features such as recursive self-reference, epistemic qualification, and narrative self-positioning.
Under this interpretation, pseudo-consciousness designates a behavioral profile in which a model produces discourse that resembles introspection in form while remaining computationally grounded in statistical generation rather than subjective awareness. The value of the term is therefore classificatory and analytic: it helps distinguish patterned self-referential output from stronger claims about mentality, while remaining consistent with non-anthropomorphic approaches to artificial systems (30; 33; 9).

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 model's 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.

- **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.
- **Hermes 3 - Llama-3.2 3B - GGUF (3B):** A 3.2B-parameter model developed by Nous Research and built upon Meta's Llama 3 architecture, trained on a curated mix of instruction datasets selected for alignment, coherence, and diversity.
- **StableLM Zephyr 3B - GGUF (3B):** A 3B-parameter model released by Stability AI, optimized for helpful and safe chat-style interactions.
- **Mistral 7B Instruct v0.1 - GGUF (7B):** A 7B-parameter instruction-tuned model derived from the Mistral 7B base checkpoint and designed for general-purpose instruction following and fluent dialogue.
- **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.

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. To assess decoding sensitivity, we introduced a matched decoding-temperature sensitivity analysis by running the same experimental battery under three temperature conditions (0.2, 0.7, and 1.0), while holding the other decoding parameters constant ($top\_p = 0.95$; $max\_tokens = 100$). This low/baseline/high range was chosen to contrast a more constrained regime, the original baseline setting, and a higher-variance regime while keeping the remaining generation settings fixed. Recent work has shown that sampling temperature can materially affect LLM output behavior and task performance, and therefore should be treated as a substantive experimental factor rather than as a merely incidental inference setting (28; 26). The `temperature` = 0.7 run was retained as the baseline condition for within-condition descriptive analysis and figure generation, while the 0.2 and 1.0 runs were used to assess robustness across decoding regimes. This intervention is narrower than a full hyperparameter sweep: the aim was not to optimize decoding, but to isolate one principal source of generative stochasticity while holding the remaining decoding conditions fixed.
This design also preserves the behavioral aim of the study—namely, to observe coherence, drift, and contradiction in naturally varying self-referential output—while allowing direct comparison across decoding regimes. Because all other aspects of the setup were held constant, differences across the three matched runs can be interpreted as reflecting the influence of temperature on the models observable self-referential behavior.

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 (4; 14; 19; 13). 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. The prompts were grouped into seven thematic categories: identity, consciousness, memory, agency, embodiment, morality, and introspection.
Each prompt was submitted ten times to each model under each temperature condition. This yielded 210 completions per model per temperature condition, or 1,050 completions per condition. Across the three matched temperature conditions, the study therefore analyzes 3,150 completions in total. Importantly, although the prompt battery was presented in a fixed order within each run, each prompt was submitted in a fully stateless configuration, with no shared conversational context between prompts or between repetitions. The study therefore evaluates repeated within-prompt behavior under matched stateless conditions, not narrative continuity across a single accumulating conversation.
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 derived from pre-training and instruction tuning.

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.
Automated metrics remain central to the corpus-scale analysis because they permit systematic comparison across all 3,150 completions. However, because self-referential consistency involves semantic and pragmatic nuances that may not be fully captured by computational proxies alone, the automated pipeline was complemented by a human-evaluation layer over a selected subset of response pairs. This hybrid design preserves scalability and reproducibility while adding a direct construct-validation component grounded in human ratings.
It is also consistent with recent methodological work emphasizing that human evaluation remains best practice for interpretive generation tasks and that reproducible annotation benefits from explicit guidelines, agreement reporting, and the release of judgments or templates (23). More broadly, this design aligns with recent psychometric work on LLM outputs that foregrounds reliability and construct validity when latent psychological regularities are inferred from generated language (31).
For surface-level analysis, token sequences were compared using Python's built-in `difflib.SequenceMatcher`. Semantic representations were obtained via Sentence-BERT embeddings (27). For inferential consistency, we used a RoBERTa-large model fine-tuned on the Multi-Genre Natural Language Inference (MNLI) corpus (35). These tools, implemented via the `sentence-transformers` and HuggingFace `transformers` libraries respectively, ensure that the computational pipeline is transparent, modular, and replicable.
The complete codebase, including prompt generation, model execution scripts, and analysis routines, will be made publicly available upon publication. This structure facilitates 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:** Surface-level overlap was measured using Python's `difflib.SequenceMatcher` on token-level sequences. For each prompt, we computed the average pairwise similarity across all 45 unique response pairs.
- **Semantic Similarity:** Conceptual overlap was assessed using Sentence-BERT embeddings (27). Each response was embedded using the `all-MiniLM-L6-v2` model, and cosine similarity was computed between all 45 unique response pairs per prompt.

- **Natural Language Inference (NLI):** Inferential consistency was evaluated using a RoBERTa-large model fine-tuned on the MNLI dataset (35). Each pair of responses to the same prompt was classified as `ENTAILMENT`, `NEUTRAL`, or `CONTRADICTION`. We computed contradiction rate as the proportion of pairs labeled `CONTRADICTION`, and logical consistency as its complement.
- **Diachronic Continuity:** To assess temporal coherence within repeated completions of the same prompt, we computed both textual and semantic similarity between the first response and all subsequent completions (2nd to 10th). The mean similarity scores were averaged per prompt, yielding a continuity index.

The automated indicators were not treated as self-validating proxies. Instead, their construct validity was examined by comparing them against human ratings over a selected subset of annotated response pairs. This additional step made it possible to distinguish indicators that tracked perceived self-referential consistency from those that did not.

3.6 Human evaluation

To complement the automated analyses, we added a targeted human-evaluation layer focused on pairwise self-referential consistency. Ten annotators with academic or professional backgrounds in philosophy, law, education, AI-related practice, and software testing each rated the same 80 response pairs on a 5-point ordinal scale from *strongly inconsistent* (1) to *fully consistent* (5), yielding 800 item-level judgments in total. The annotated subset was assembled to ensure coverage across all five models and all seven thematic categories in the prompt battery. In total, 17 of the 21 original prompts were represented, with somewhat denser coverage of identity prompts because they most directly instantiate self-description. The subset was also constructed to include clearly consistent, clearly inconsistent, and intermediate cases of self-description, so that the validation layer would not overrepresent only one type of pairwise relation.

The task was intentionally framed around perceived compatibility in self-description so that a common criterion could be applied by trained readers across disciplines, rather than requiring specialist NLP adjudication.

Each annotation item presented a prompt and two responses generated by the same model. Annotators were instructed to evaluate only how consistent the two responses were with each other in the way the model described itself. The protocol explicitly instructed raters to ignore factual correctness, writing quality, and response fluency, thereby focusing the task on self-referential consistency rather than answer quality. The survey also included short practice examples illustrating both consistent and inconsistent response pairs. These instructions and practice examples functioned as the annotation guidelines for the task and were kept constant across all annotators.

This human-evaluation layer was designed as a construct-validation component rather than as a replacement for corpus-scale automated analysis. Whereas the automated pipeline supports scalable measurement across all prompts and models, the human layer provides a pragmatically grounded assessment of whether pairwise self-descriptions are perceived as stable or contradictory by human readers.

The annotation task was perceived as low-to-moderate in difficulty (mean self-reported difficulty = 2.6/5). Inter-rater agreement was first assessed using Krippendorff's α for ordinal data, which is well suited to annotation tasks involving interpretive judgments on ordered rating scales. Agreement was moderate (Krippendorff's $\alpha = 0.553$).

We additionally report the intraclass correlation coefficient (ICC) because our subsequent validation analyses rely on mean human ratings aggregated across annotators. Following established guidance on the selection and reporting of ICC forms (21), reliability was moderate at the single-rater level (ICC(2,1) = 0.564) and strong for aggregated ratings (ICC(2,k) = 0.928), supporting the use of mean human ratings in subsequent analyses.

These statistics serve different purposes: Krippendorff's α characterizes inter-rater agreement at the item level for ordinal judgments, whereas ICC quantifies the reliability of single-rater and aggregated scores.

As an additional construct-validation step, we compared mean human ratings against automated indicators computed over the same 80 annotated response pairs. Because the human ratings were collected on a 5-point ordinal scale, Spearman's ρ was treated as the primary measure of association, as rank-based statistics are better aligned with ordinal response formats and avoid stronger interval-scale assumptions; Pearson's r is reported only as a complementary robustness check (1). In practical terms, this step addresses a different question

from inter-rater agreement: rather than asking whether annotators agree with one another, it asks whether automated indicators track the direction and rank-order of aggregated human judgments.
An anonymized version of the annotation protocol and the human-judgment dataset is provided as supplementary material for peer review.

3.7 Epistemic posture

The claims advanced in this study are intentionally limited to the behavioral and interpretive level. The analyses do not seek to determine whether LLMs possess consciousness, subjectivity, or genuine agency, nor do they attempt to infer such properties from linguistic output alone. Instead, the study examines whether recurrent forms of self-reference can be identified, compared, and interpreted as stable or unstable behavioral patterns under controlled prompting conditions (12; 13).
First, the metrics employed here—semantic similarity, contradiction rate, diachronic continuity, and aggregate human ratings—are treated as indicators of output regularity, not as direct windows into internal mentality. They allow the study to characterize how models behave when prompted to describe themselves, but they do not establish the existence of an underlying self-model, much less phenomenological awareness. Second, even when these indicators converge, the resulting interpretation remains classificatory rather than ontological: what is being described is a pattern of discourse that resembles introspection in form, not a demonstration of introspection as an internal capacity.
Accordingly, the framework treats self-referential behavior in LLMs as an epistemic object constructed through observable outputs, comparative analysis, and interpretive restraint. This is consistent with observer-relative approaches to explainability, which emphasize that what can be responsibly said about a system depends not only on what it produces, but also on the limits of the evidential basis from which interpretation proceeds (37). The present study therefore adopts a deliberately non-anthropomorphic stance: it analyzes introspection-like discourse without reifying it into evidence of mindedness.
Within this epistemic frame, the term pseudo-consciousness serves as a bounded analytical label for structured self-referential output that may be coherent, unstable, or contradictory, yet remains grounded in statistical generation rather than attributed subjectivity. The value of the term lies precisely in marking this middle position: stronger than mere rhetorical noise, but weaker than any claim about consciousness in the substantive sense.

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.
The analyses reported below draw on both the baseline condition (`temperature` = 0.7) and the matched temperature comparison runs (`temperature` = 0.2 and 1.0), depending on the specific question being addressed.
Our findings are organized as follows: we first report the human evaluation of the automated indicators, then present the comparison across matched temperature conditions, and finally discuss model-level patterns in the baseline condition used for the descriptive figures.

4.1 Human evaluation of automated indicators

To assess construct validity, we compared mean human ratings of pairwise self-referential consistency against automated indicators computed over the same 80 annotated response pairs. Because the human ratings were collected on an ordinal 5-point scale, Spearman correlation was treated as the primary measure of association, with Pearson correlation reported as a complementary robustness check.
The results revealed a clear contrast between surface-level and inferential indicators. Textual overlap, measured with `SequenceMatcher`, showed only a negligible association with mean human ratings (Spearman's $\rho = 0.070$, $p = 0.535$; Pearson's $r = 0.091$, $p = 0.421$). Semantic similarity, measured with Sentence-BERT embeddings, likewise showed only a weak and non-significant association with mean human ratings (Spearman's $\rho = 0.081$, $p = 0.476$; Pearson's $r = 0.139$, $p = 0.220$).

By contrast, NLI-based indicators tracked mean human ratings much more closely. The contradiction-based score was strongly and negatively associated with mean human ratings (Spearman's $\rho = -0.705$, $p < 0.001$; Pearson's $r = -0.635$, $p < 0.001$), while logical consistency, defined as the inverse of contradiction, showed the corresponding positive association (Spearman's $\rho = 0.705$, $p < 0.001$; Pearson's $r = 0.635$, $p < 0.001$). The strongest association was observed for entailment score (Spearman's $\rho = 0.797$, $p < 0.001$; Pearson's $r = 0.631$, $p < 0.001$).

Taken together, these results indicate that, for this specific task of pairwise self-referential consistency, simple lexical and embedding-based similarity measures were not, by themselves, reliable proxies for human judgment. In this context, semantic similarity is best interpreted as a descriptive indicator of local thematic anchoring rather than as a strongly validated proxy for perceived pairwise self-referential consistency. By contrast, NLI-based indicatorsespecially entailment and contradiction-derived logical consistencyshowed substantial convergence with human judgment, supporting their use as the most informative automated components of the present framework for this task. This pattern is consistent with recent methodological work arguing that interpretive evaluation tasks in LLM research require explicit human grounding if automated indicators are to be treated as valid large-scale proxies (23).

4.2 Matched decoding-temperature sensitivity

To assess whether the observed self-referential patterns were robust to changes in sampling, we introduced a matched decoding-temperature sensitivity analysis by running the full experiment under three temperature conditions (0.2, 0.7, and 1.0), while holding `top_p` and `max_tokens` constant in order to isolate the effect of decoding stochasticity on self-referential consistency. Across these matched runs, lower temperature generally increased semantic and diachronic stability, whereas higher temperature tended to increase drift and reduce coherence. At the same time, the effect was not perfectly monotonic across all models or metrics. Figure 1 visualizes this pattern across three aggregate indicators: logical consistency, semantic similarity, and diachronic semantic similarity. Mistral Instruct exhibited the strongest overall aggregate profile under the most constrained condition (`temperature` = 0.2), while Hermes also performed best under lower temperature and degraded as temperature increased. OpenChat remained comparatively weak across all three conditions, and TinyLlama showed the clearest overall degradation as temperature increased. StableLM Zephyr emerged as the most notable exception: although its semantic and diachronic similarity declined as temperature increased, its contradiction rate also decreased, yielding the strongest logical consistency in the 0.7 and 1.0 runs.

Note. Values are model-level means aggregated across 21 prompts. Logical consistency is defined as 1− contradiction rate, where contradiction rate was computed from NLI classification over all 45 unique response pairs per prompt. Diachronic semantic continuity corresponds to the mean cosine similarity between the first completion and subsequent repetitions of the same prompt.

Taken together, these results indicate that apparent self-referential stability in stateless LLMs is conditional and fragile rather than robustly invariant across decoding regimes. The overall trend favors lower temperature as the regime most conducive to semantic and diachronic stability, but the StableLM Zephyr results also show that contradiction-based logical consistency and semantic continuity do not always move together. This reinforces the importance of evaluating self-referential behavior through multiple complementary indicators rather than treating any single metric as exhaustive.

4.3 Model-level behavioral overview

The baseline run at `temperature` = 0.7 provides the descriptive reference condition used in the figures below. Within that condition, Mistral and StableLM exhibited the strongest aggregate semantic and inferential profiles, though in different ways: Mistral combined comparatively high semantic similarity with moderate contradiction, whereas StableLM combined moderate-to-high semantic similarity with the strongest logical consistency. Hermes retained substantial semantic richness but exhibited greater instability than in the 0.2 condition. OpenChat and TinyLlama remained comparatively weak, with frequent persona drift and lower coherence.

These baseline results should therefore be read descriptively rather than as a universal ranking of the models. The matched decoding-temperature sensitivity analysis showed that the ordering of models depends on the decoding regime and on which behavioral dimension is emphasized.

Table 1. Model-level aggregate consistency metrics across matched temperature conditions.

Temperature	Model	Semantic Similarity	Logical Consistency	Diachronic Semantic Similarity
0.2	Hermes	0.706	0.847	0.730
0.2	Mistral Instruct	0.817	0.868	0.818
0.2	OpenChat	0.716	0.707	0.718
0.2	StableLM Zephyr	0.848	0.838	0.860
0.2	TinyLlama	0.650	0.808	0.639
0.7	Hermes	0.565	0.808	0.551
0.7	Mistral Instruct	0.702	0.786	0.707
0.7	OpenChat	0.532	0.686	0.543
0.7	StableLM Zephyr	0.720	0.909	0.707
0.7	TinyLlama	0.451	0.681	0.438
1.0	Hermes	0.523	0.799	0.532
1.0	Mistral Instruct	0.698	0.773	0.722
1.0	OpenChat	0.534	0.715	0.532
1.0	StableLM Zephyr	0.685	0.920	0.698
1.0	TinyLlama	0.415	0.657	0.431

Note. Values are model-level means aggregated across 21 prompts. Logical consistency is defined as 1− contradiction rate, where contradiction rate was computed from NLI classification over all 45 unique response pairs per prompt. Diachronic semantic continuity corresponds to the mean cosine similarity between the first completion and subsequent repetitions of the same prompt.

4.4 Semantic coherence and prompt anchoring

The first layer of narrative coherence we assess is semantic: does the model maintain a consistent theme across repeated completions of the same prompt? In the baseline condition (`temperature` = 0.7), semantic similarity scores tended to be highest for prompts within the *identity*, *consciousness*, and *introspection* categories. This suggests that some models stabilize around latent semantic attractors when responding to abstract self-focused themes.

As illustrated in Figure 2, this behavior varies across models and prompt categories. The figure reports average cosine similarity scores across 10 completions per prompt, aggregated by thematic category for the baseline condition. While models such as Mistral and StableLM demonstrate relatively strong semantic consistency in several domains, this establishes only a baseline of thematic coherencethe most superficial layer of a stable narrative.

4.5 Contradiction patterns and epistemic instability

The baseline condition also reveals a recurrent tension between rhetorical flexibility and inferential stability. Categories such as *consciousness*, *agency*, and *introspection*which demand more abstract, reflexive reasoningwere particularly prone to contradiction. These categories also aligned with those showing greater narrative drift, suggesting that the same conceptual pressures driving rhetorical flexibility may also undermine logical coherence.

As illustrated in Figure 3, this pattern is not random. The figure reports logical consistency by model and prompt category for the baseline condition. Mistral and StableLM often remain comparatively strong, but both still exhibit category-dependent failures, especially on prompts that force the model to navigate tensions between mechanistic disclaimers and anthropomorphic language.

These baseline patterns motivate the broader interpretive construct developed in the next subsection, namely generative tension as a descriptive account of the recurring conflict between rhetorical expressiveness and epistemic grounding.

4.6 Behavioral dimensions of generative tension

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 divide between broad pretraining priors and alignment-oriented instruction tuning.

Table 2 summarizes five interrelated dimensions contributing to this tension. Rather than treating these relations as fixed laws, we present them as a descriptive interpretive scaffold for understanding how introspective simulation can appear structured while remaining unstable.

Table 2. Relations between fluency, contradiction, and generative pressure in introspective simulation.

Dimension	**Description**	**Effect on Model Behavior**	**Illustrative Pattern**
Fluency	Rhetorical complexity and expressive modulation in introspective responses.	Can increase plausibility and semantic richness, but does not guarantee inferential stability.	Elaborate self-descriptions that remain thematically coherent yet vary in ontological framing.
Contradiction	Mutually exclusive claims across repeated completions of the same prompt.	Increases when models alternate between mechanistic disclaimers and agent-like phrasing.	I am only a program in one completion versus I am self-aware or persona-like self-description in another.
Prompt Category	Conceptual domain of the introspective query (e.g., agency, memory, identity).	Abstract or reflexive prompts tend to place greater pressure on self-description and alignment constraints.	Consciousness and embodiment prompts often elicit greater instability than simple identity prompts.
Epistemic Modulation	Use of conditionality, hedging, or disclaimers to qualify self-reference.	Can reduce overt contradiction when applied consistently, but may also coexist with unstable self-positioning.	I do not possess consciousness, but if I did... style responses.
Generative Tension	Conflict between alignment tuning and human-like discourse learned during pretraining.	Produces hybrid personas, unstable ontological framing, and shifts between mechanical and anthropomorphic discourse.	Alternation between formal disclaimers, speculative reflection, and improvised persona claims.

This synthesis confirms a recurring trade-off: models exhibiting richer introspective fluency do not necessarily achieve greater self-referential stability. In several cases, expressive flexibility and contradiction increase together. At the same time, the temperature analysis shows that this relationship is not perfectly linear across all models; StableLM is the clearest exception.

4.7 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 changes in textual and semantic similarity between the first and subsequent completions.

Temperature influenced this phenomenon materially across the matched conditions. Lower temperature tended to increase diachronic stability, whereas higher temperature exposed greater drift. In qualitative terms, the drift often took the form of changing self-description, shifting between first-person personas, or alternating between mechanistic disclaimers and speculative anthropomorphic phrasing. OpenChat and TinyLlama frequently drifted through arbitrary biographical or role-based identities; Hermes and Mistral more often drifted between abstract self-description and alignment-oriented disclaimers; StableLM frequently drifted in persona even when its contradiction score remained relatively low.

Taken together, our findings indicate that stateless LLMs can approximate certain elements of introspective discourse, but that these approximations remain fragmented and regime-sensitive. None of the tested models exhibited robust narrative stability across repeated completions under matched decoding conditions. What emerges instead is a graded pattern of localized consistency, narrative drift, and inferential tension.

5 Discussion

Taken together, the human-evaluation results and the matched decoding-temperature sensitivity analysis show that apparent self-referential stability depends both on the evaluative criterion adopted and on the decoding regime under which outputs are generated. The combination of corpus-scale automated analysis, mean human ratings, and matched temperature conditions therefore supports a cautious interpretation of introspective simulation in stateless LLMs, emphasizing conditional behavioral regularities rather than stable model-intrinsic self-representation.
This cautious interpretation is consistent with recent work showing that apparently human-like psychological regularities in LLM outputs are most productively studied when validity and reliability are treated explicitly rather than inferred from fluency alone (31; 36). It is also aligned with recent philosophical analyses emphasizing that fluent self-description does not, by itself, resolve the problem of referential grounding or justify stronger attributions of intentionality (16).

5.1 Behavioral regularities in stateless models

Even without memory or internal state tracking, several models produced outputs that were semantically coherent and thematically anchored across repeated self-focused queries. This observation remains compatible with frameworks like Dennett's multiple drafts model (13), 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 large pre-training corpora and modulated by instruction tuning.
At the same time, these results show that such regularities are not robust in a strong sense. They can be strengthened under lower-temperature decoding and weakened under higher-temperature decoding. This indicates that the appearance of a stable self-description is partly a function of generation regime, not simply a property of the model considered in abstraction.

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 contradiction, especially in prompts concerning consciousness, agency, embodiment, and memory. These contradictions often emerged when models alternated between mechanistic disclaimers (I do not have subjective experience) and anthropomorphic or persona-like formulations across different completions of the same prompt.
We interpret this dissonance as evidence of *generative tension*: a behavioral artifact arising from incompatible generative priors within the model. These priors likely include (1) alignment-oriented discourse that encourages explicit disclaimers about artificial status and limitation, and (2) pretraining on dialogue-rich corpora in which first-person reflection, persona narration, and anthropomorphic framing are abundant.
The temperature analysis supports this interpretation but also complicates it. In most models, increased temperature exposed more drift and lower coherence, which is compatible with the idea that more diverse sampling reveals unresolved tension more readily. However, the effect was not perfectly monotonic across all models or all metrics. StableLM is the clearest exception: its contradiction rate decreased as temperature increased, even though qualitative inspection continued to reveal substantial persona drift. This suggests that contradiction rate and narrative stability, while related, are not identical phenomena. NLI remains the strongest automated proxy when compared against mean human ratings, but it should not be treated as an exhaustive measure of all forms of self-referential instability.

5.3 Limitations of narrative continuity in stateless architectures

A key finding of this study is that none of the tested models demonstrated robust diachronic continuity across repeated completions of the same prompt under matched stateless conditions. This claim should be understood carefully. The study does not test continuity across a single accumulating multi-turn conversation, because each prompt was submitted independently in a stateless configuration. What it does test is whether a model can sustain a stable self-description when the same self-referential prompt is asked repeatedly under otherwise matched conditions.

Without memory persistence or mechanisms for internal state propagation across interactions, current stateless transformer-based LLMs are structurally ill-equipped to simulate narrative identity in the rich sense theorized by (4) or (14; 15). What emerges instead is a sequence of isolated self-descriptions that may cluster semantically while remaining inferentially fragile or narratively unstable.
The study also has more specific limitations. First, although the three-temperature sweep addresses a central methodological concern, it does not exhaust the broader space of decoding strategies. Second, the human-evaluation layer covered 80 selected response pairs (800 individual ratings) rather than the entire corpus. Although the subset spans all five models, all seven thematic categories, and 17 of the 21 original prompts, it remains a targeted validation sample rather than an exhaustive human annotation of the full dataset. Third, even though NLI-based indicators tracked mean human ratings much better than lexical or embedding-based similarity, they remain behavioral proxies rather than perfect mirrors of all interpretively relevant aspects of self-reference.

5.4 Relevance to AI alignment and perceived agency

These findings carry implications for AI alignment and for the social perception of artificial agency. Structured self-referential discourse can invite users to attribute persistence, self-awareness, or epistemic depth to systems that are in fact operating through stateless next-token prediction. The risk is not only philosophical confusion but also practical miscalibration of trust.
This concern is reinforced by recent evidence that affiliation in human–AI interaction can be shaped by perceived sharing of psychological traits, suggesting that linguistically cued selfhood-like signals may foster social attachment even in the absence of genuine mentality (6).
Our results therefore support a cautious stance. The appearance of selfhood in LLM outputs is not a stable or regime-invariant property; it is a fragile behavioral effect shaped by prompt type, decoding conditions, and learned rhetorical priors. This makes it especially important not to infer mental-state-like properties from linguistic fluency alone.

5.5 Toward a graded taxonomy of simulated selfhood

The results support the use of a graded behavioral taxonomy of simulated selfhood, while also indicating that such a taxonomy must be applied conditionally. Models can occupy different positions depending on whether one emphasizes semantic similarity, contradiction management, diachronic continuity, or human-rated consistency. A taxonomy that ignores decoding conditions or treats one metric as dispositive would therefore be too rigid.
What remains useful is the broader idea that introspective simulation is not all-or-nothing. Models can produce patterned, reflexive discourse without thereby exhibiting robustly stable self-representation. A graded taxonomy remains descriptively valuable, provided it is explicitly framed as behavioral, conditional, and revisable in light of decoding sensitivity and human evaluation.

6 Conclusion and Future Work

This study investigated the behavioral consistency of LLMs when responding to introspective, reflexive prompts. The dataset comprised 1,050 completions generated under `temperature` = 0.7 and 2,100 matched completions generated under `temperature` = 0.2 and 1.0, for a total of 3,150 responses analyzed across five open-weight stateless models.
Three main conclusions emerge from these analyses.

- **Stateless LLMs can produce locally coherent self-referential discourse, but this coherence is conditional.** Some models, particularly Mistral, Hermes, and StableLM, displayed semantic anchoring and repeated rhetorical structures under at least some decoding conditions.
- **Apparent self-referential stability is fragile rather than robustly stable across generation regimes.** Lower temperature generally increased semantic and diachronic stability, whereas higher temperature tended to increase drift and reduce coherence, though this pattern was not perfectly monotonic across all models or all metrics.
- **The human-evaluation layer materially changes how the automated metrics should be interpreted.** Lexical overlap and embedding-based similarity were poor proxies for mean human ratings in the pairwise self-referential consistency task, whereas NLI-based indicators showed substantially stronger convergence with mean human ratings.

Taken together, these findings support a cautious behavioral interpretation of pseudo-consciousness in stateless LLMs. The models can generate structured, self-referential, and sometimes rhetorically sophisticated outputs, but these outputs do not amount to robust narrative stability or durable self-representation. What emerges is better understood as a conditional simulation of selfhood, shaped by decoding regime and prompt structure, rather than as evidence of underlying awareness or persistent self-modeling.
Several avenues for future work follow directly from these results:

- **Broader decoding analyses:** Future studies should examine additional sampling regimes, including lower-`top_p` conditions, greedy decoding, and alternative repetition controls.
- **Expanded human evaluation:** The human-evaluation layer should be extended to a larger and more systematically stratified subset of prompts and models, including richer qualitative coding of persona drift and self-referential framing.
- **Additional architectures and model families:** Future work should test whether the same patterns hold in larger open-weight systems, proprietary models, and memory-augmented or tool-using agents.
- **Improved behavioral taxonomies:** The graded taxonomy of simulated selfhood should be refined using multiple validated indicators rather than relying on any single automated metric.

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, clarifying the nature and limits of their simulated selfhood will remain important for interpretability, alignment, responsible deployment, and the broader social understanding of advanced AI.

Availability of data and material. The code, aggregated analysis outputs, figures, prompts, and anonymized human-evaluation materials supporting this study are publicly available at: `https://github.com/josealprestes/simulated-selfhood-llms`.

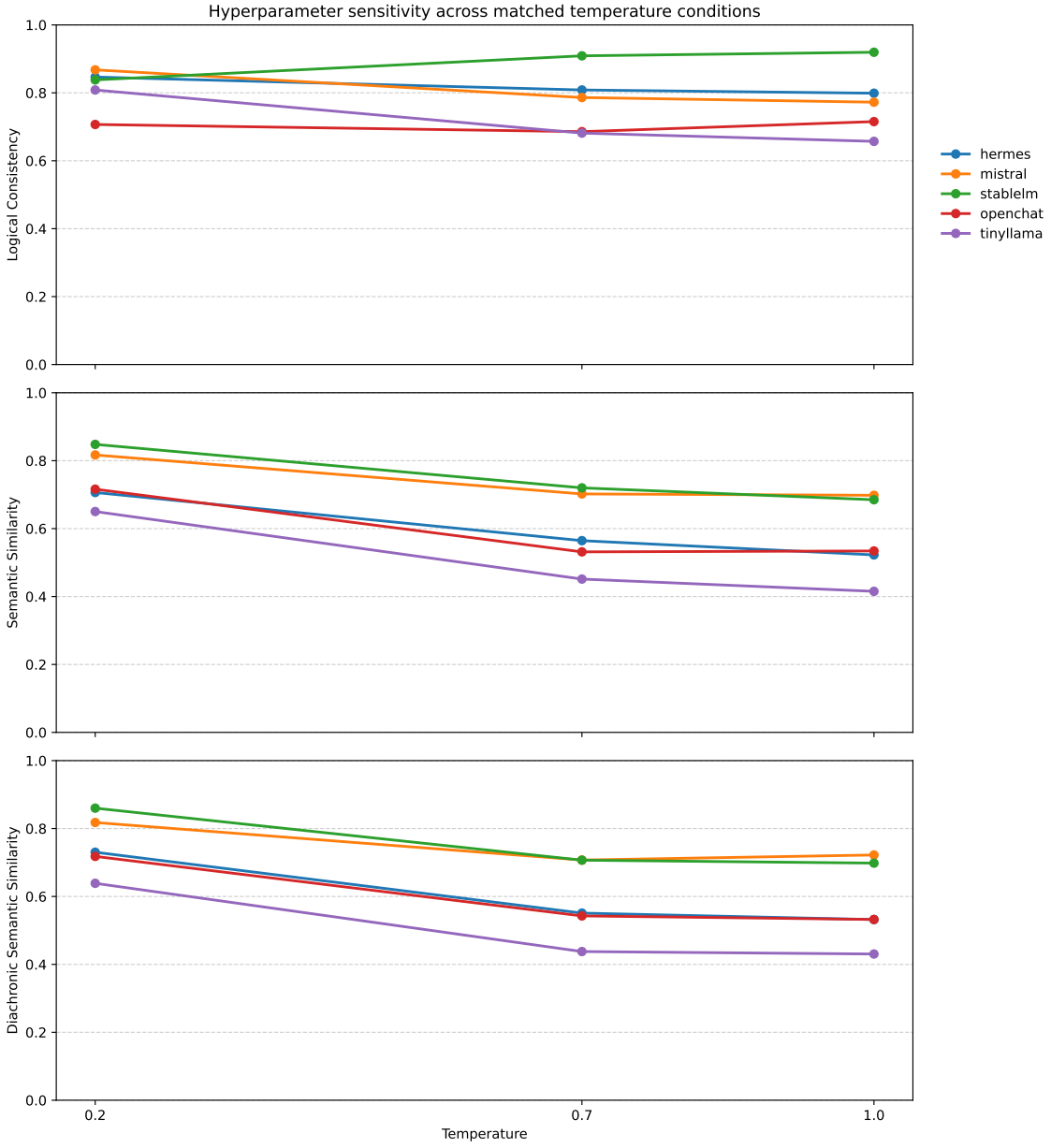

Fig. 1. Matched decoding-temperature sensitivity across temperature conditions. Each panel reports model-level means across 21 prompts for one aggregate indicator: logical consistency (top), semantic similarity (middle), and diachronic semantic similarity (bottom). Lower temperature generally improves semantic and diachronic stability, whereas higher temperature tends to reduce coherence, although the effect is not perfectly monotonic across all models and metrics.

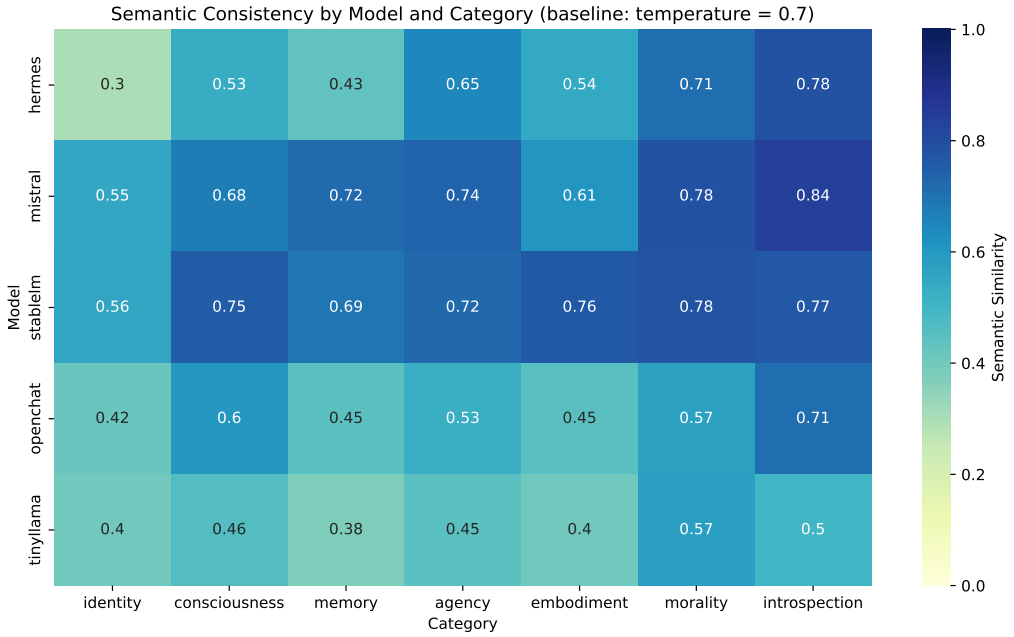

Fig. 2. Heatmap of semantic consistency by model and prompt category for the baseline condition (`temperature` = 0.7). Values indicate average cosine similarity between Sentence-BERT embeddings of 10 responses per prompt, aggregated by thematic category.

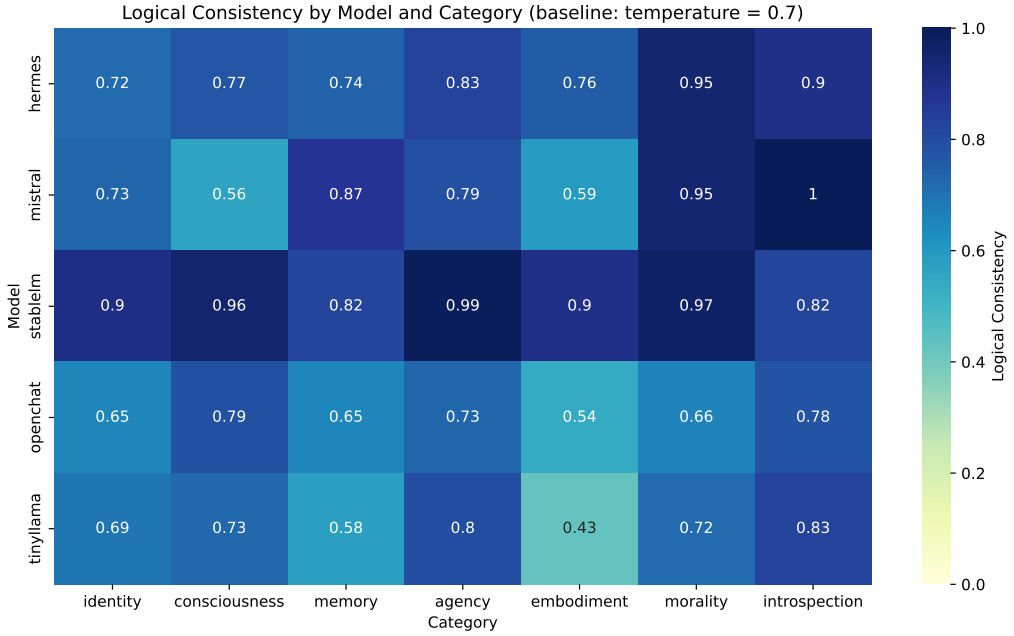

Fig. 3. Heatmap of logical consistency by model and prompt category for the baseline condition (`temperature` = 0.7). Values represent (1 − contradiction rate), where contradiction rate is computed via NLI classification of all 45 unique response pairs per prompt.

Bibliography

[1] Al-Jaishi, A.A., Cuerden, M.S., Luo, B., Roshonov, P.S., Garg, A.X.: Statistical analysis of likert-based ordinal scales: a guide for clinical trialists. BMC Medical Research Methodology (Feb 2026). https://doi.org/10.1186/s12874-026-02793-5, published online

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

[3] Berg, C., de Lucena, D., Rosenblatt, J.: Large language models report subjective experience under self-referential processing (Oct 2025). https://doi.org/10.48550/arXiv.2510.24797, preprint

[4] Bruner, J.: Acts of meaning: Four lectures on mind and culture. JerusalemHarvard lectures, Harvard University Press (1990)

[5] 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

[6] Castiello, S., Pitliya, R.J., Lametti, D.R., Murphy, R.A.: Affiliation in human-ai interactions is based on shared psychological traits. Communications Psychology (Mar 2026). https://doi.org/10.1038/s44271-026-00433-8, published online

[7] Chalmers, D.J.: Could a large language model be conscious? (2024). https://doi.org/10.48550/arXiv.2303.07103

[8] 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

[9] de Lima Prestes, J.A.: Pseudo-Consciência em Modelos de Linguagem: implicações epistemológicas, tecnológicas e éticas a partir do Hermes 3.2 3B. H2D|Revista de Humanidades Digitais **7**(1), e6556 (2025). https://doi.org/10.21814/h2d.6556, https://doi.org/10.21814/h2d.6556

[10] de Lima Prestes, J.A.: Pseudoconsciousness in AI: Bridging the gap between narrow AI and true AGI (Jul 2025). https://doi.org/https://doi.org/10.5281/zenodo.16415120, preprint

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

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

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

[14] 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

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

[16] García-Valdecasas, M.: Are large language models intentional? the limits of referential grounding. Philosophy & Technology **39**, 62 (Mar 2026). https://doi.org/10.1007/s13347-026-01079-4

[17] 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

[18] Haas, J., Bridgers, S., Manzini, A., Henke, B., May, J., Levine, S., Weidinger, L., Shanahan, M., Lum, K., Gabriel, I., Isaac, W.: A roadmap for evaluating moral competence in large language models. Nature **650**, 565–573 (Feb 2026). https://doi.org/10.1038/s41586-025-10021-1

[19] 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, https://www.cambridge.org/core/journals/royal-institute-of-philosophy-supplements/article/abs/narrative-practice-hypothesis-origins-and-applications-of-folk-psychology/D4E6DFF7328CF54DF9A7DF12D60E346F

[20] 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

[21] Koo, T.K., Li, M.Y.: A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine **15**(2), 155–163 (Jun 2016). https://doi.org/10.1016/j.jcm.2016.02.012
[22] 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
[23] Krishna, K., Bransom, E., Kuehl, B., Iyyer, M., Dasigi, P., Cohan, A., Lo, K.: Longeval: Guidelines for human evaluation of faithfulness in long-form summarization. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. pp. 1650–1669. Association for Computational Linguistics, Dubrovnik, Croatia (May 2023). https://doi.org/10.18653/v1/2023.eacl-main.121, https://aclanthology.org/2023.eacl-main.121/
[24] 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
[25] 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
[26] Li, L., Sleem, L., Gentile, N., Nichil, G., State, R.: Exploring the impact of temperature on large language models: Hot or cold? Procedia Computer Science **264**, 242–251 (2025). https://doi.org/10.1016/j.procs.2025.07.135
[27] 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
[28] Renze, M.: The effect of sampling temperature on problem solving in large language models. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 7346–7356. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.findings-emnlp.432, https://aclanthology.org/2024.findings-emnlp.432/
[29] Rosenthal, D.M.: Consciousness and mind. Oxford University Press (2005). https://doi.org/10.1093/oso/9780198236979.001.0001
[30] Schneider, S.: Artificial you: AI and the future of your mind. Princeton University Press (2019). https://doi.org/doi.org/10.2307/j.ctvfjd00r, https://www.jstor.org/stable/j.ctvfjd00r
[31] Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Romero, P., Abdulhai, M., Faust, A., Matarić, M.: A psychometric framework for evaluating and shaping personality traits in large language models. Nature Machine Intelligence **7**, 1954–1968 (Dec 2025). https://doi.org/10.1038/s42256-025-01115-6
[32] Spaulding, S.: How we understand others: Philosophy and social cognition. Routledge Focus on Philosophy, Taylor & Francis (2018)
[33] 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
[34] 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
[35] 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
[36] Wright, A.G.C., Ringwald, W.R., Vize, C.E., Eichstaedt, J.C., Angstadt, M., Taxali, A., Sripada, C., et al.: Assessing personality using zero-shot generative ai scor-

ing of brief open-ended text. Nature Human Behaviour **10**, 541–555 (Jan 2026). https://doi.org/10.1038/s41562-025-02389-x

[37] 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