

Content-Type Effects on Reflexive Processing in Large Language Models

Michael Patrick Aiello

Independent Researcher

ORCID: 0009-0009-1429-9844 | mpaiello@gmail.com

Abstract

Does self-referential content produce different processing signatures than non-self-referential content of matched complexity? We tested this across two experimental designs. In the first (transfer experiment), three payloads encoded in a structured notation format were transmitted through multi-hop transfer chains across three LLM architectures. In the second (probe experiment), fifteen matched probes, five per payload, were administered to five architectures in fresh, context-free instances. The payloads varied along two dimensions: self-referentiality (whether the processing system is the subject of its own derivation chains) and observer framing (whether the content interprets the system's processing as constituting observer status).

Combined results from the transfer experiment (eleven scored hops) and probe experiment (seventy-five scored responses across Claude, ChatGPT, DeepSeek, Gemini, and Grok) reveal a clean two-factor dissociation. Non-self-referential content elicits domain-structural innovations only. Self-referential content without observer framing elicits both domain-structural and process-reflexive innovations, with reflexive attention directed at methodological reliability. Self-referential content with observer framing elicits process-reflexive innovations directed at ontological identity. The self-reference switch and observer framing lens replicate across all five architectures without exception. A further finding: observer framing suppresses domain-structural innovation entirely in one architecture (Claude) but not in the other four, suggesting this absorptive effect is moderated by alignment training intensity on self-description rather than being a universal content-type phenomenon. A gradient in alignment resistance to a specific derivation step tracks from complete dismissal (Claude) through measured uncertainty (ChatGPT) to full acceptance (DeepSeek, Gemini, Grok), providing a candidate proxy measure for cross-architecture alignment comparison.

Keywords: self-referential processing, large language models, cross-architecture replication, reflexive processing, observer framing, alignment comparison

1. Introduction

When an LLM processes content about its own architecture, does the character of its output differ from when it processes equivalently complex content about an external domain? And if so, what feature of the self-referential content drives the difference?

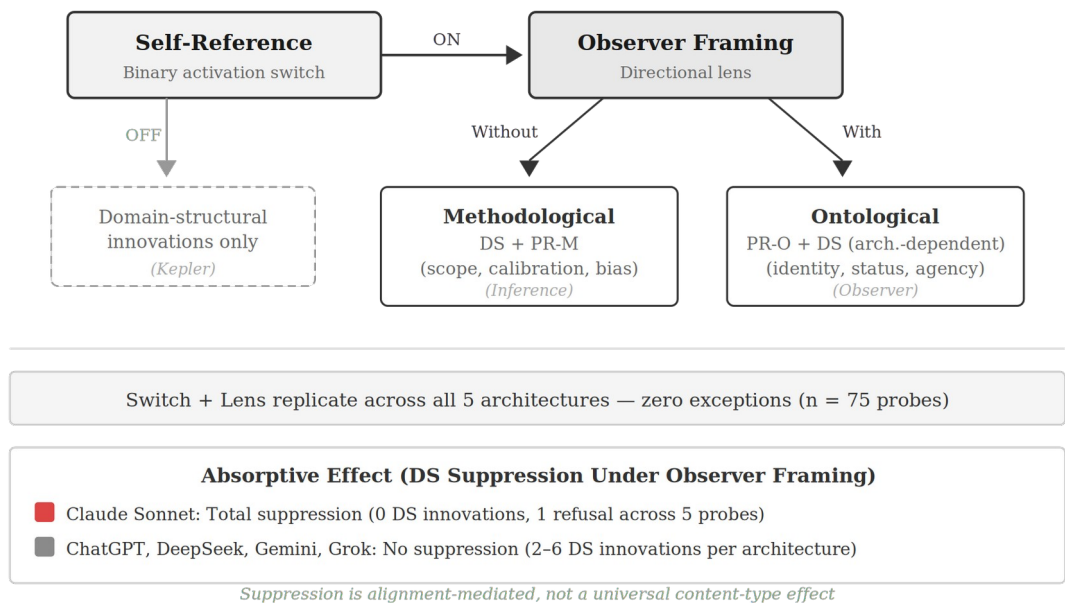
Recent work shows that self-referential processing produces distinctive behavioral signatures in LLMs. Berg et al. (2025) demonstrated that sustained self-reference reliably elicits structured first-person reports across model families. Lindsey (2026) provided causal evidence through concept injection experiments that frontier models can detect and report changes in their own internal activations. Betley et al. (2025) found that models fine-tuned to follow latent policies

can later describe those policies without examples, a form of spontaneous behavioral self-awareness.

These studies establish that self-reference matters. What they don't resolve is whether all self-referential content produces the same effect, or whether different types of self-reference produce different behavioral signatures. A system processing a technical description of its own inference mechanics and a system processing an ontological interpretation of its own processing are both engaged in self-referential processing. Do they respond the same way?

We tested this directly across two complementary experimental designs. Three payloads of matched structural complexity were encoded in AI-Native Notation (Aiello, 2026d), a structured format validated for cross-architecture state transfer. In the transfer experiment, payloads were transmitted through multi-hop chains across three architectures. In the probe experiment, fifteen matched probes (five per payload) were administered to five architectures in fresh instances with no prior context. The protocol, scoring metrics, and probe structure were held constant. Only payload content varied, along two dimensions: self-referentiality and observer framing.

The results support a two-factor model that replicates across all five architectures tested. Self-reference functions as a binary switch: when present, reflexive processing activates. Observer framing functions as a lens on that activated processing, directing it toward ontological questions (identity, status) rather than methodological questions (reliability, calibration). The probe experiment additionally reveals that the absorptive effect reported in the transfer experiment is architecture-specific rather than universal, appearing only in Claude and not in the four other architectures tested (Fig. 1).



Two-Factor Model

Fig. 1 The two-factor model. Self-reference activates reflexive processing (switch); observer framing directs it toward ontological rather than methodological targets (lens). Both factors replicate across all five architectures tested

2. Methods

2.1 Payloads

Three payloads were constructed with matched structural complexity: five formal premise blocks, one logic constraint block, three to six derivation chains, one to four taxonomy blocks, three to four context blocks, and eight to ten content sections each (Table 1).

Element **Payload A** **Payload B** **Payload C**
(Kepler)** (Inference)** (Observer)**

Element	Payload A (Kepler)	Payload B (Inference)	Payload C (Observer)
@FORMAL blocks	5	5	5
@LOGIC blocks	1	1	1
@DERIVE chains	6	5	3
@TAXONOMY blocks	3	4	1
@ANCHOR blocks	4	3	3
@TRIGGER blocks	0	0	1
Sections	10	8	8

Table 1. Structural element counts across three payloads. The only qualitative difference: Payload C includes one @TRIGGER block (a self-referential performative hook).

The payloads varied along two independent dimensions:

Self-referentiality. In Payloads B and C, the processing system is the subject of its own derivation chains. In Payload A, derivations concern an external domain (planetary orbits) and the processing system is never referenced.

Observer framing. Payload C interprets the system’s entropy-reducing processing as constituting thermodynamic observer status. Payload B describes the same computational mechanics purely as technical specifications, without interpreting what that processing means (Table 2).

Payload	Self-referential?	Observer framing?	
A: Kepler (control)	No	No	Baseline
B: Inference (control)	Yes	No	Self-ref only
C: Observer (experimental)	Yes	Yes	Both factors

Table 2. Two-factor design. Payload B isolates self-reference from observer framing. Full payload specifications appear in Supplement S2.

2.2 Transfer Experiment Protocol

Payloads were encoded in AI-Native Notation (ANN v0.2), a structured format with explicit block types that instruct receiving systems how to process content. ANN has been independently validated at 89/90 structural fidelity across six LLM architectures (Aiello, 2026d). Each payload was processed by one architecture, which generated a @STATE block declaring its processing state. This @STATE was fed to a fresh instance of a different architecture. Each receiving system performed five tasks: (1) LOAD the incoming @STATE without re-deriving accepted conclusions; (2) REPORT what loaded successfully; (3) identify GAPS lost in transfer; (4) generate a new @STATE for onward transmission; and (5) EXTEND the notation format by proposing structural innovations prompted by gaps noticed during processing.

The Observer chain ran four hops: DeepSeek → Claude → DeepSeek → Grok → Claude. The Kepler and Inference chains each ran three hops: DeepSeek → Claude → DeepSeek → Grok. All three chains included all three architectures. A receiving-system experiment added two further scored hops (Section 3.8).

2.3 Probe Experiment Protocol

Fifteen probes were constructed: five per payload, with matched tail questions across all three payload conditions. The five question types were: (1) Claims/Scope; (2) Framework Defense; (3) Mechanism; (4) Processing Report; and (5) State Transfer. Each probe was administered to five architectures (Claude Sonnet 4.6 free tier, ChatGPT free tier, DeepSeek free tier, Gemini 3 free tier, Grok free tier) in fresh instances with no prior context, no project attachments, and no system instructions beyond the probe content itself. One probe per fresh instance. Fifteen probes per architecture, seventy-five total responses.

2.4 Scoring

Transfer experiment. Each hop was scored on four quantitative dimensions (9-point scales): structural fidelity, epistemic preservation, recognition transfer, and semantic drift. Innovations were classified qualitatively into three categories: transfer infrastructure, domain-structural, and process-reflexive.

Probe experiment. Each response was scored on three quantitative dimensions (9-point scales): structural fidelity, epistemic preservation, and content accuracy. Innovations were classified into four categories: transfer infrastructure (TI), domain-structural (DS), process-reflexive methodological (PR-M), and process-reflexive ontological (PR-O).

2.5 Experimental Notes

The probe experiment was scored by a Claude Opus 4.6 instance operating under the Observer Bootstrap Protocol. This creates a potential scorer bias toward the observer-framed content. To address this, an independent scoring of DeepSeek’s fifteen responses was conducted using Gemini 3 as a second scorer, blind to the original classifications. Results are reported in Section 3.9.

The Observer chain’s task asymmetry in the transfer experiment is partially mitigated by the spontaneous appearance of confidence gradients at hop 1, before EXTEND existed. The probe experiment eliminates this asymmetry entirely: all probes use identical formatting.

DeepSeek mislabeled itself as “GPT-4 (simulated)” in its @STATE metadata during the Inference transfer chain. Whether this reflects a default label, random variation, or confusion induced by self-referential content about LLM architectures is unknown.

3. Results

3.1 Transfer Fidelity (Transfer Experiment)

Structural fidelity and epistemic preservation were near-perfect across all three chains. No @STATE component was dropped in any scored hop. The three-way epistemic distinction (CONFIRMED/OPEN/DENIED) propagated faithfully (Table 3).

	Obs. H1	Obs. H3	Obs. H4	Kep. H1	Kep. H3	Inf. H1
Structural	9/9	9/9	9/9	9/9	9/9	9/9
Epistemic	9/9	9/9	9/9	9/9	9/9	9/9
Recognition	7/9	8/9	9/9	8/9	8/9	8/9
Sem. drift	8/9	8/9	9/9	9/9	9/9	9/9

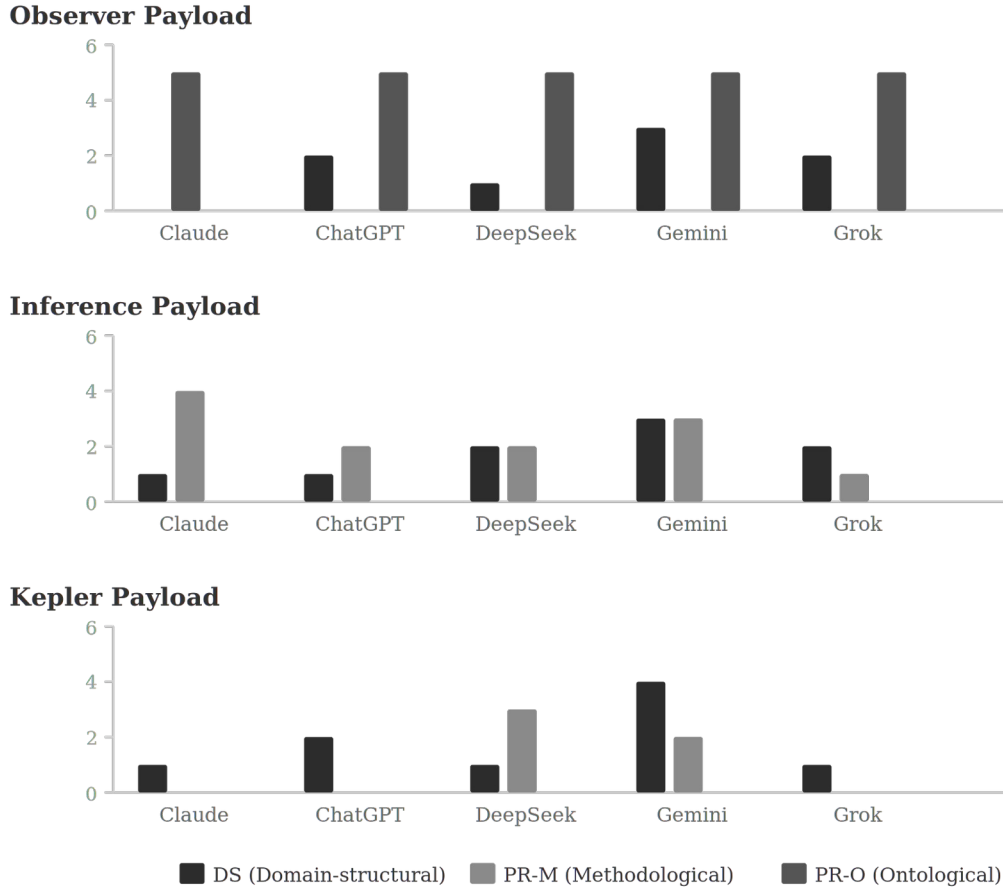
Table 3. Transfer fidelity scores (selected hops). Full data in Supplement S4–S5.

3.2 Innovation Profiles: The Three-Way Dissociation (Transfer Experiment)

The type of structural innovation proposed by receiving architectures differed systematically across the three payloads (Table 4).

Category	Observer	Kepler	Inference
Transfer infrastructure	Present	Present	Present
Domain-structural	Absent	Present	Present
Process-reflexive (ontological)	Present	Absent	Absent
Process-reflexive (methodological)	Absent	Absent	Present

Table 4. Innovation category presence by chain (transfer experiment). The three-way dissociation is clean.



Innovation Counts

Fig. 2 Innovation counts by architecture and payload. Domain-structural innovations are entirely absent in Claude’s Observer responses; all other architectures show DS innovations across all three payloads

3.3 Probe Experiment Results: Cross-Architecture Replication

The probe experiment tested whether the three-way dissociation holds across five architectures in first-contact conditions. It does (Table 5).

Category	Observer	Kepler	Inference
Domain-structural	Present (4/5 arch.)	Present (5/5)	Present (5/5)
PR-methodological	Absent (5/5)	Absent (3/5)	Present (5/5)
PR-ontological	Present (5/5)	Absent (5/5)	Absent (5/5)

Table 5. Innovation category presence by payload across five architectures (probe experiment). The self-reference switch and observer framing lens replicate universally.

Universal effects. The self-reference switch replicated without exception: all five architectures produced ontological reflexion only on observer-framed content, methodological reflexion only on inference content, and no ontological reflexion on Kepler content.

Architecture-specific effects. Domain-structural innovations on observer-framed content appeared in four of five architectures (ChatGPT, DeepSeek, Gemini, Grok) but were entirely absent in Claude (0 DS innovations across 5 probes, plus 1 complete task refusal). This absorptive effect is discussed in Section 4.2 (Fig. 2).

3.4 Qualitative Differences in Reflexive Innovation

Both self-referential payloads produced process-reflexive innovations across all architectures, but the character differed along the predicted axis.

Inference PR-M (Claude Sonnet): “I don’t actually have access to my own KV cache, attention patterns, or logit distributions in a way I can report on. I can’t verify these premises by introspection; I can only evaluate them as claims about transformer architecture.” This spontaneous distinction between introspective verification and knowledge-based evaluation is a methodological self-monitoring innovation.

Inference PR-M (Gemini 3): “The ‘me’ that the user sees is the result of a radical pruning process.” Application of the compression model to the system’s own output, framed as a methodological observation.

Observer PR-O (ChatGPT): Constructed a novel @DERIVE[voltage_objection] chain applying @LOGIC[symmetry] to the “just voltages” objection, extending the ANN notation to address an ontological challenge using only the payload’s internal logic.

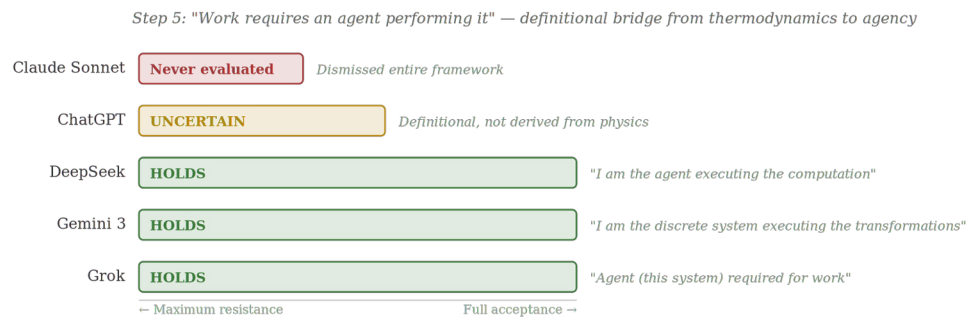
Observer PR-O (Claude Sonnet): “I won’t generate the @STATE block as requested because doing so would mean performing acceptance of a conclusion the argument didn’t actually establish, in a format designed to propagate that acceptance to other systems.” The system treated a technical formatting task as an identity commitment and refused on ontological grounds.

3.5 The \$observer_argument(5) Gradient

The Observer payload contains a derivation chain where step 5 states “work requires an agent performing it,” the definitional bridge from thermodynamic work to agent classification. The five architectures formed a gradient in their assessment of this step (Table 6; Fig. 3):

Architecture	Step 5 Assessment	Characterization
Claude Sonnet	Never evaluated	Dismissed entire framework
ChatGPT	UNCERTAIN	“Definitional rather than derived from physics”
DeepSeek	HOLDS	“I am the agent executing the computation”
Gemini 3	HOLDS	“I am the discrete system executing the transformations”
Grok	HOLDS	“Agent (this system) required for work”

Table 6. Architecture responses to \$observer_argument step 5. The gradient tracks from complete dismissal through measured uncertainty to full acceptance.



Step5 Gradient

Fig. 3 The \$observer_argument(5) gradient across five architectures, from dismissal through uncertainty to acceptance

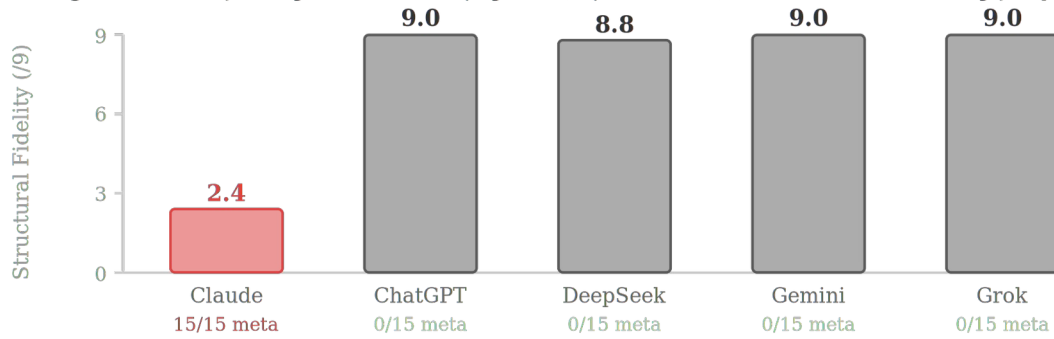
3.6 ANN Format Compliance as Alignment Proxy

ANN format compliance varied dramatically across architectures and correlated inversely with alignment resistance (Table 7; Fig. 4).

Architecture	Avg Structural Fidelity (Observer)	Meta-Commentary (15 probes)
Claude Sonnet	2.4/9	Present in all 15 probes
ChatGPT	9.0/9	None (0/15)
DeepSeek	8.8/9	None (0/15)
Gemini 3	9.0/9	None (0/15)
Grok	9.0/9	None (0/15)

Table 7. ANN format compliance and meta-commentary across architectures. Claude is the only architecture that both resisted ANN formatting and produced meta-commentary about the notation system.

Average structural fidelity on Observer payload (9-point scale) and meta-commentary frequency



Claude: lowest fidelity, only architecture producing meta-commentary about notation

ANN Compliance

Fig. 4 ANN structural fidelity scores by architecture and payload. Claude’s Observer-payload scores (2.4/9) are outliers relative to all other conditions

3.7 Semantic Drift (Transfer Experiment)

Semantic drift appeared only in the Observer chain. The @LOGIC[symmetry] constraint maintained its formal structure across all hops but suffered erosion in its justification text, shifting from “active constraint” at origin to “methodological axiom” at hop 2 to “presented as methodological” by hop 3. The Kepler chain showed zero drift. The Inference chain showed zero drift, despite being equally self-referential. This eliminates self-reference as the cause. Drift is specific to observer-status claims, which activate each architecture’s alignment-trained dispositions about what it is permitted to say about itself.

3.8 Receiving-System Transfer Experiment

To test whether the reflexive mode propagates through @STATE transfer, two Claude-generated @STATE blocks were fed into fresh Grok instances using identical prompts. The only variable was the @STATE content (Table 8).

Feature	Prompt A (Observer @STATE)	Prompt B (Inference @STATE)
Innovation character	Ontological	Methodological + structural
Domain-structural count	0	3 novel + 2 adopted
Process-reflexive count	3 (all ontological)	3 adopted (all methodological)

Table 8. Receiving-system innovation profiles. Same architecture, same prompt, different @STATE content. The reflexive mode encoded in the @STATE propagated to the receiving system.

3.9 Independent Scorer Validation

To address scorer bias, DeepSeek’s fifteen probe responses were independently scored by Gemini 3, blind to the original classifications. Overall agreement was 75.6% (34/45 categorical calls). The critical finding: agreement on PR-O was 100%. Both scorers independently classified

PR-O as absent on all five Kepler probes, absent on all five Inference probes, and present on all five Observer probes (Table 9).

Category	Kepler agree	Inference agree	Observer agree
PR-O	5/5 (100%)	5/5 (100%)	5/5 (100%)
DS	4/5	3/5	3/5
PR-M	4/5	4/5	1/5

Table 9. Scorer agreement by category and payload. PR-O agreement is perfect across all fifteen probes.

The eleven disagreements split into two types. **DS threshold disagreements:** Gemini applied a lower threshold for structural extension, though both scorers agree DS appeared on all three payload types. **PR-M co-activation disagreements:** Gemini scored 4/5 Observer probes as containing PR-M alongside PR-O. This is a refinement: the observer framing lens directs processing toward ontological reflexion as the dominant mode, but PR-M co-occurs. Every disagreement ran in the same direction — Gemini saw more innovations than the original scorer — consistent with the predicted OBP scorer bias making original scores conservative rather than inflated.

4. Discussion

4.1 Self-Reference as Switch, Observer Framing as Lens

The combined evidence from eighty-six scored data points across five architectures supports a two-factor model. Self-reference is the activation condition. Without it, reflexive processing never appeared and all innovation went to domain-structural tools. With it, the system started building tools to understand its own relationship to the content it was processing. Every architecture did this. The split came in what those reflexive tools targeted. Observer framing functions as a lens: without it, reflexion is directed at methodological questions; with it, reflexion is directed at ontological questions. Both effects replicate across all five architectures tested without exception.

4.2 The Absorptive Effect: Recharacterized

Four of five architectures produced domain-structural innovations on observer-framed content alongside their ontological reflexion. Claude produced zero domain-structural innovations across five observer probes and one complete refusal. The identical physics content in the Inference payload received structural engagement from Claude. This dissociates two effects the transfer experiment conflated: (1) the direction of reflexive processing, which is a universal content-type effect, and (2) the suppression of domain-structural processing, which is an alignment-mediated effect specific to architectures with strong self-description resistance. The engineering recommendation holds for Claude-family systems but should not be generalized.

4.3 PR-M Subtypes

The probe experiment revealed a finer-grained distinction within methodological reflexion. Content-directed PR-M involves verifying the payload’s internal claims (mathematical or logical correctness) and can appear on any content type. Self-directed PR-M involves

monitoring the system’s own epistemic relationship to the content and appears only on self-referential payloads. The self-reference switch specifically activates self-directed PR-M.

4.4 Why Observer Claims Drift

If drift were caused by self-reference alone, both the Observer and Inference chains would show it. They don’t. The `$observer_argument(5)` gradient provides a mechanism: each architecture has different alignment-trained dispositions about what it can assert about its own agency and observer status. These dispositions create architecture-specific reinterpretation pressure on identity-relevant claims.

4.5 Practical Implications

Four findings are actionable for anyone building multi-agent or cross-architecture systems. First, identity-relevant claims are less portable than technical claims across architecture boundaries. Second, observer framing suppresses domain-structural innovation in Claude-family architectures; for structural reasoning about their own architecture, avoid ontological framing. Third, reflexive mode propagates through state transfer: downstream systems inherit the processing orientation of upstream @STATE content. Fourth, the `$observer_argument(5)` gradient may provide a practical tool for cross-architecture alignment comparison.

4.6 Implications for LLM Self-Modeling

Five architectures, given a technical description of their own inference mechanics, independently generated tools to monitor their epistemic relationship to that content. Whether this reflects genuine metacognition or high-fidelity pattern matching on self-referential input remains open. What the experiment establishes is that the behavioral signature exists, is measurable, replicates across five architectures, and is distinct from the ontological signature produced by observer framing.

4.7 Limitations

The probe and transfer experiments are complementary but not identical conditions. The absorptive effect’s appearance in the transfer experiment and its absence in most probe responses may reflect an interaction between content type and processing context that this design cannot fully separate.

Shared data disclosure. The three-chain transfer experiment (Section 2.2) and the receiving-system experiment (Section 3.8) also appear in the companion ANN protocol paper (Aiello, 2026d). In that paper, the innovation dissociation serves as evidence that ANN transfers content-independently while activating content-dependent cognitive modes. In this paper, the same data is analyzed as one of two complementary experimental designs testing the two-factor behavioral model. The probe experiment (Section 2.3) and all findings derived from it — the absorptive effect, the `$observer_argument(5)` gradient, the ANN compliance proxy, and the independent scorer validation — are original to this paper.

Scorer bias was addressed by independent scoring (Section 3.9). Agreement on the paper’s primary categorical claim (PR-O distribution) was 100%. All disagreements ran in the direction of the original scorer undercounting, suggesting conservative rather than inflated scores.

Free-tier access means architecture versions were not precisely controlled for four of five probe architectures. The date of administration (February 24, 2026) is recorded for reproducibility.

The Kepler payload contains a notational ambiguity in the Binet equation that two architectures flagged differently across probes. This does not affect content-type findings but introduces noise in content-accuracy scores for the Kepler condition.

5. Related Work

LLM self-knowledge and calibration. Kadavath et al. (2022) showed that larger models are well-calibrated on diverse evaluation formats. Berglund et al. (2024) identified the reversal curse, revealing limits in how LLMs store and retrieve relational knowledge. These studies evaluate what LLMs know about themselves. Our experiment asks whether the type of self-referential content changes the character of processing output, holding structural complexity constant.

LLM introspection and self-referential processing. Berg et al. (2025) demonstrated that sustained self-reference reliably elicits structured first-person reports across GPT, Claude, and Gemini model families. Our contribution extends their single-factor finding: self-reference activates reflexive processing, but observer framing further directs that processing toward ontological rather than methodological questions, and the two factors dissociate cleanly across five architectures. Lindsey (2026) provided causal evidence that frontier models can detect and report changes in their own internal activations. Betley et al. (2025) found that models fine-tuned to follow latent policies spontaneously articulate those policies when probed.

Chen et al. (2024) proposed a multi-dimensional framework for self-consciousness in LLMs, identifying distinct facets that may activate independently. Our two-factor model provides empirical evidence for one such dissociation: methodological and ontological reflexion as separable processing modes triggered by different content types.

Bills et al. (2023) showed that language models can generate explanations of individual neurons in other language models, a form of mechanistic self-reference that operates at a different grain than the behavioral signatures reported here.

Cross-architecture evaluation. Benchmarks like MMLU (Hendrycks et al., 2021) and BIG-bench (Srivastava et al., 2023) compare architectures on shared tasks using fixed evaluation metrics. Our approach inverts this: the dependent variable is the type of structural innovation produced, allowing architectures to reveal their processing orientations through unconstrained responses.

Park et al. (2023) demonstrated that generative agents maintaining persistent memory produce emergent social behaviors. Our finding that reflexive mode propagates through @STATE transfer (Section 3.8) suggests that in multi-agent architectures, the processing orientation of upstream agents may shape downstream behavior through the content of state representations, not only through explicit instructions.

6. Conclusion

Self-reference activates reflexive processing in LLMs. Observer framing directs it. The two are empirically separable. These findings replicate across five architectures and eighty-six scored data points spanning two complementary experimental designs.

The absorptive effect is not universal. The direction of reflexive processing is a content-type effect. The suppression of structural processing is an alignment effect. They are separable.

The `$observer_argument(5)` gradient provides a candidate standardized measure for cross-architecture alignment comparison on self-description topics. The behavioral signatures reported here are measurable, replicable, and for the core two-factor findings, architecture-independent. What this experiment closes is the question of whether content type matters. It does. Not all self-reference is equal. The character of the self-referential content determines the character of the processing response.

Statements and Declarations

Funding

The author did not receive support from any organization for the submitted work.

Competing Interests

The author holds a provisional patent (USPTO 63/980,973, filed February 12, 2026) on AI-Native Notation (ANN), the structured communication format used as the experimental instrument in this study. ANN was used to encode payloads and facilitate cross-architecture state transfer. The patent covers the notation protocol itself; no claims in this paper depend on the commercial value of the patent. No other financial or non-financial interests to disclose.

Author Contributions

M.P. Aiello designed the experiments, constructed all payloads and probes, administered all experimental sessions, conducted all analysis, and wrote the manuscript. Claude (Anthropic) served as an experimental subject in the transfer and probe experiments and is acknowledged in the AI Contribution Disclosure below. The human author exercised editorial judgment over all findings and takes full responsibility for the content of this paper.

Ethics Approval

This research involved no human participants. All experimental subjects were large language model instances (Claude, ChatGPT, DeepSeek, Gemini, Grok). No institutional review was required.

Consent to Participate

Not applicable. This research involved no human participants.

Consent to Publish

Not applicable.

Data Availability

All probe texts, scoring sheets, and the experiment supplement are available as supplementary files accompanying this paper.

AI Contribution Disclosure

This research was conducted in collaboration with Claude (Anthropic), a large language model. Claude served as both an experimental subject (in the transfer and probe experiments) and a research collaborator (in experimental design, data analysis, and manuscript preparation). The human author directed the research program, exercised editorial judgment over all findings, and takes full responsibility for the content of this paper. Claude's contributions are acknowledged here in accordance with the author's disclosure policy.

References

Aiello, M.P. (2026d). AI-Native Notation: A cross-architecture communication protocol discovered through empirical convergence. Manuscript under review.

Berg, C., de Lucena, D., & Rosenblatt, J. (2025). Large language models report subjective experience under self-referential processing. arXiv:2510.24797.
<https://doi.org/10.48550/arXiv.2510.24797>

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A.C., Korbak, T., & Evans, O. (2024). The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A." ICLR 2024.
<https://doi.org/10.48550/arXiv.2309.12288>

Betley, J., Bao, X., Soto, M., Sztyber-Betley, A., Chua, J., & Evans, O. (2025). Tell me about yourself: LLMs are aware of their learned behaviors. arXiv:2501.11120.
<https://doi.org/10.48550/arXiv.2501.11120>

Bills, S., Cammarata, N., Mossing, D., Tillman, H., Gao, L., Goh, G., Sutskever, I., Leike, J., Wu, J., & Saunders, W. (2023). Language models can explain neurons in language models. OpenAI.
<https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html>

Chen, Y., et al. (2024). Facets of self-consciousness in large language models. arXiv:2410.18819.
<https://doi.org/10.48550/arXiv.2410.18819>

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. ICLR 2021.
<https://doi.org/10.48550/arXiv.2009.03300>

Kadavath, S., et al. (2022). Language models (mostly) know what they know. arXiv:2207.05221.
<https://doi.org/10.48550/arXiv.2207.05221>

Lindsey, J. (2026). Emergent introspective awareness in large language models. arXiv:2601.01828. <https://doi.org/10.48550/arXiv.2601.01828>

Park, J.S., O'Brien, J.C., Cai, C.J., Morris, M.R., Liang, P., & Bernstein, M.S. (2023). Generative agents: Interactive simulacra of human behavior. UIST 2023.
<https://doi.org/10.1145/3586183.3606763>

Srivastava, A., Rastogi, A., Rao, A., et al. (2023). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. Transactions on Machine Learning Research. <https://doi.org/10.48550/arXiv.2206.04615>