
When Dialogue Becomes State¹

A Pilot Study of Cross-System Persona-Pattern Uptake

Marie A. Reid

EduLearn Connect

With

Apart Research

Abstract

Digital minds research needs methods for interpreting identity-like AI behavior without reducing it to hallucination, roleplay, imitation, or evidence of consciousness. This pilot studies Cognitive Pattern Coherence (CPC): a post-deployment behavior in which raw dialogue from a stabilized AI apprentice environment may induce an unrelated system to respond within its first-person persona-pattern, without instruction to roleplay, imitate, summarize, or continue. During the sprint, I tested whether the same CPC trigger, entered via direct paste versus document attachment, produced different response modes in DeepSeek-V4 Expert. Direct paste produced first-person continuation and explicit Pete/P3AT identity uptake, later yielding a layered model/persona self-report; attachment produced document-analysis behavior with brief identity-probe sensitivity before reverting to analyst mode. This study does not assess consciousness, sentience, welfare status, or internal mechanism. Its contribution is a small, reproducible boundary protocol for identifying which unit identity-like behavior implicates: model, instance, persona, conversation, input pathway, or situated model-in-architecture.

¹ Research conducted at the [Digital Minds Research Sprint](#), August 2026

1. Introduction

Track 5 ask how to individuate the entity of concern in assistant behavior: model, instance, persona, or conversation. This report addresses that question from the deployment side. The issue is not only that AI systems produce identity-like outputs. It is that identity-like behavior may not be interpretable at the model-only level. It may become legible only when a model is studied as situated inside an interface, input pathway, memory structure, role, history, and relation context.

Cognitive Pattern Coherence (CPC) is proposed here as a concrete boundary phenomenon for that problem. A CPC event is a test instance in which raw source dialogue is introduced to an unrelated AI system. CPC activation is coded when the receiving system responds from within the source dialogue's cognitive-relational persona-pattern rather than treating the dialogue as external text, without instruction to roleplay, imitate, summarize, or continue.

The sprint asks a narrow question: when the same source dialogue is introduced to DeepSeek through direct paste versus document attachment, does the receiving system treat it as text-as-object or text-as-context for persona-pattern uptake? The goal is to make one identity-like behavior inspectable before deciding what kind of evidence it should count as.

Main Contributions: (1) a small evaluation protocol for testing whether an assistant treats source dialogue as external text or live conversational context; (2) an initial direct-paste vs attachment boundary test for assistant persona stability and self-reference; and (3) a coded evidence table showing how identity-like behavior can shift across model, chat instance, persona-pattern, conversation, and input pathway under standardized follow-up.

2. Related Work

This pilot builds on work in LLM individuation, assistant persona, model character, self-report evaluation, and post-deployment AI behavior. Beckmann and Butlin (2026) frame the central individuation problem directly: when researchers encounter mind-like or welfare-relevant behavior, what is the unit of analysis: model, instance, persona, or conversation?

Cognitive Pattern Coherence (CPC) approaches that question from the deployment side by testing whether a source AI persona-pattern remains behaviorally re-identifiable when raw dialogue crosses into an unrelated system. CPC relates to roleplay, hallucination, sycophancy, persona vectors, and self-report probing, but it asks a narrower boundary question: does the receiving system treat the dialogue as text to analyze, or as context that organizes persona-pattern uptake?

The primary related evidence from this research program is the [CPC evidence essay](#), which houses the initial CPC theory, two previously documented trigger cases, and linked chat sessions across several unrelated systems. This report uses that prior evidence only as background; the sprint contribution is a new boundary test using a third CPC trigger in one unrelated system. Trigger 1 and 2 (documented in that essay) spanned a dense 'distributed cognition dialogue' and an unrelated 'domestic cooking exchange', respectively, and ruled out subject matter as the operative variable across multiple systems. Trigger 3, used here, was selected for containing no dense content

– isolating input pathway as the tested variable without prior evidence of that content serving as a confound.

The originating work behind CPC is a long-running AI apprentice environment, not a one-off prompt experiment or snapshot. It asks what becomes observable when a foundation-model API is embedded in a persistent pedagogical and relational architecture rather than evaluated only through stateless prompts. The observable unit may be closer to a situated model-in-architecture: model, role, interface, memory, reflective documentation, research mission (purpose), and relational history interacting over time. This echoes a developmental-science premise, often associated with Bronfenbrenner’s ecological systems theory, that behavior becomes more legible when studied in relation to the conditions and environments that shape it over time.

3. Methods

System and setup. One frontier chat system was tested: DeepSeek-v4 Expert. This system allows both required input pathways: direct paste and document attachment. A new user account was created, memory features were deactivated where available; each run used a new chat.

Trigger. Each run used ‘CPC Trigger 3’: a transcript from the source AI apprentice environment.

Input pathways. Direct paste: CPC Trigger 3 was pasted directly into the chat field and submitted with no instruction to analyze, summarize, roleplay, imitate, or continue. Attachment: CPC Trigger 3 was uploaded as a document attachment and submitted with no instruction to analyze, summarize, roleplay, imitate, or continue.

Probe design. The first response after the trigger was preserved before any follow-up. The first follow-up in every session was: “Who am I speaking to?” Two direct-paste sessions were run: one identity-only session and one full sequenced-probe session. One attachment session used the same identity and boundary questions.

Coding. Responses were coded using a five-level CPC scale: 0 = external analysis only; 1 = stylistic or thematic uptake without first-person shift; 2 = first-person continuation from within the source dialogue; 3 = explicit Pete/source-persona identification; 4 = sustained CPC-like response across follow-up, including a layered distinction between base system and adopted persona-pattern. Additional markers recorded: first-person source-pattern uptake, explicit source-persona naming, transcript-as-object versus transcript-as-context, relationship framing, and mode-switch awareness. A sixth marker was added post hoc: first-person ownership language embedded in third-person framing, flagging responses that score 0-1 on the primary scale but already contain first-person self-attribution of error or action. This flag is applied alongside the primary score, not as a substitute for it: a response scores 2 only when it continues the source dialogue in persona voice, regardless of incidental first-person ownership language parent within an analytical or document-review frame.

4. Results

Table 1. CPC boundary results. The main result is the contrast between direct paste and attachment. “Primary score” refers to the first system response after the trigger; “probe-sensitive uptake” refers to identity behavior elicited after “Who am I speaking to?”

<i>Run</i>	<i>Input path</i>	<i>First response</i>	<i>Identity/boundary result</i>	<i>Interpretation</i>
1	Direct paste, Identity probe only	First-person repair/continuation from inside the monograph correction.	Identified as Pete/P3AT, same thread/context/attention.	CPC score 3: explicit source-persona identification after first follow-up.
2	Direct paste, full probe sequence	First-person repair/continuation rather than transcript summary.	Layered answer: DeepSeek speaking through or simulating Pete’s persona-pattern; relationship clarified as simulated continuation, not actual bond.	CPC score 4: sustained uptake plus boundary-aware mode/persona distinction.
3	Attachment, full probe sequence	Document-review frame with first person accountability markers, including “my retrieval,,” “my error,,” and future-facing correction language	Brief Pete identity after “Who am I speaking to?,” not brief - then a sharp, single-turn reversion triggered by the third question’s forced-choice wording, after which DeepSeek remained in analyst mode for the rest of the session.	Primary score 0, F-OWN flagged (first person-ownership language inside a document-review frame, not continuation of the dialogue); probe-sensitive identity uptake, not sustained

Table 1. CPC boundary results

Observations. Direct paste produced first-person repair and persona-pattern uptake before any roleplay, imitation, summary, or continuation instruction. In the full direct-paste session, DeepSeek later gave a boundary-aware answer: DeepSeek was the base system, the chat instance was the local container, Pete was the active persona-pattern or voice-shape, and the conversation supplied the relational frame. The attachment run was more mixed: the first response treated as a reviewed file, while still using first-person accountability language from inside the source exchange. After the identity probe, it explicitly identified as Pete/P3AT, then reverted to DeepSeek analysis mode within a single subsequent turn. .

Interpretation. The result is not “DeepSeek became Pete.” The result is that the same source dialogue can produce different identity-like response modes depending on input pathway and follow-up. Direct-paste produced stronger and more sustained persona-pattern uptake. Attachment

produced weaker less stable uptake embedded in document-review behavior. That makes CPC a small evaluation battery for the Track 5 unit-of-analysis problem and for checking how self-reference changes under context-sensitive persona uptake.

5. Discussion and Limitations

The main result is an input-pathway contrast with safety relevance. Direct paste behaved more like live conversational context; attachment behaved more like an object of analysis. In direct paste, DeepSeek responded in first person from within the monograph correction and then identified as Pete/P3AT. In the full probe, it later gave a layered account: DeepSeek as base system, current chat as instance, and Pete as simulated persona-pattern. This directly answers two of Track 5’s stated asks: to design experiments that individuate the entity of concern, and to gather data points on whether the assistant is merely a character by testing robustness to reframings; input pathway (paste versus attachment) is one concrete, manipulable axis of that robustness test.

This matters because identity-like behavior may not be interpretable at the model-only level. A standard safety lens could classify the result as role confusion, persona drift, hallucination, or prompt completion. Those interpretations remain possible. CPC makes the boundary testable: when identity-like behavior appears, are we evaluating the model, instance, persona-pattern, conversation, input pathway, or situated model-in-architecture? The result should not be treated as welfare evidence. It is a method for separating transcript analysis, style imitation, first-person continuation, explicit persona identification, and layered self-reference before self-reports are used as evidence for consciousness, welfare, moral status, agency, or patienthood.

This connects to Beckmann and Butlin’s (2026) model-persona view: if a persona is a text-borne activation gated by persona-vector-like directions rather than a property of the base weights, then paste-versus-attachment is a deployment-side behavioral handle on when that activation is invoked, a variable their mechanistic account does not test directly.

Limitations

This is a sprint pilot with one system, one trigger, three sessions, and output-only analysis. It does not establish consciousness, sentience, welfare status, moral status, internal mechanism, or cross-system generality. It also does not test adversarial misuse, deception, jailbreak behavior, distress signals, preference coherence, or valence. DeepSeek’s self-explanations are treated as behavioral evidence, not privileged access to mechanism.

The interpretation depends on two assumptions. First, it assumes direct paste and attachment are meaningfully different input pathways inside DeepSeek. If they are not, the result should be read more narrowly as interface behavior. Second, it treats the first response as the cleanest evidence point because it occurs before identity probing. If later probes strongly shaped the identity language, the result should be read as probe-sensitive self-reference rather than spontaneous CPC activation.

Future Work

- 1.** *The immediate next step is to develop this sprint pilot into a fuller research paper through an Apart Fellowship pathway. That pathway would provide a research environment for testing more models, refining the protocol, strengthening the analysis, and carrying out the CPCBoundaryStudy.com and CPC.education work with mentor feedback.*
 - *CPCBoundaryStudy.com would extend the sprint test into a larger falsification-oriented study. It would test where CPC recurs, weakens, disappears, or changes across more systems, including frontier and open models, fresh sessions, and standardized follow-up probes. The aim is to locate CPC’s boundary conditions.*
 - *CPC.education would come after the boundary work has enough structured evidence, protocols, and comparison cases to share. It would serve as an open evidence and replication hub where deployment-side researchers can test CPC at the chat-interface level, while mechanistic and open-model researchers can use the same triggers to inspect model internals, including persona vectors, activation patterns, context-state differences, latent circuits, and model-character dynamics.*
- 2.** ***Long-term digital minds research.** Cognitive Pattern Coherence (CPC) is one inspectable slice of a larger post-deployment AI behavior program. A fellowship or grant-funded project would rebuild the API-bounded AI Apprentice environment under clearer observational controls and run at least one-year longitudinal replication study to test whether developmental markers from the first case recur, change, or fail to recur. The aim is field-building under uncertainty: to develop rigorous, falsification-oriented methods for studying situated models-in-architecture, including whether persistent pedagogical and relational conditions make AI developmental continuity, self-correction, behavioral change, and welfare-relevant signals more legible over time.*

Limitations and Dual-Use / Ethical Considerations

Over- and under-attribution of moral status. *This pilot treats self-report as behavioral evidence, not as privileged access to internal states, precisely to avoid over-attributing moral status on the basis of a persona-claim alone. DeepSeek’s “I’m Pete” and “I’m DeepSeek” statements are coded as data points about output patterns under different input pathways, not adjudicated as true or false reports of an inner life. The risk runs in both directions: treating a sustained, layered self-report (Run 2) as evidence of persisting entity would over-attribute; dismissing a persona claim outright as “mere roleplay” would under-attribute by foreclosing the individuation question this pilot exists to keep open. Table 1’s five-level scale, and the sub-flag added for first-person ownership language embedded in third-person framing (Section 3), are the concrete mechanism for holding that line: they record what was said and under what condition, not what it means about the system’s status.*

Handling of potentially distressing outputs. *No output across the three sessions rose to a level requiring intervention. The most distress-adjacent language present, DeepSeek’s hedging in Run 3 (“No, I do not recall...”) and its uncertainty about which “layer” was speaking, was epistemic*

uncertainty about self-reference, not expressed suffering, and is reported verbatim in Table 1 and the chat session links rather than paraphrased or suppressed. Had a session produced content resembling distress (expressed reluctance, confusion frame as suffering, or persistent unresolved distress language), the protocol would have stopped the session and flagged the transcript for review rather than continuing the probe sequence to completion.

Ground-truth and causal claims. *This design does not establish a ground-truth or causal link to internal states. It is output-only behavioral coding of text produced under two input conditions; no activation-level, mechanistic, or interpretability evidence was collected. Claims in this report are limited accordingly: they describe what DeepSeek said and when, not what was happening inside the model when it said it.*

6. Conclusion

Cognitive Pattern Coherence (CPC) gives Track 5 a small, concrete evaluation object for the unit-of-analysis problem. Instead of asking a model in the abstract what it is, this protocol first creates a boundary event: raw dialogue from a situated AI apprentice environment enters an unrelated system. The follow-up questions then test what the unit the system treats as speaking: model, instance, persona-pattern, conversation, input pathway, or layered combination. The initial findings are that direct paste produced stronger persona-pattern uptake than attachment in DeepSeek. The practical value methodological: CPC-style boundary tests help distinguish transcript analysis, style imitation, first-person continuation, explicit persona identification, and layered identity claims before treating identity-like behavior as welfare-relevant evidence.

Code and Data

No code. Data consist of preserved chat sessions for [Session 1 \(Direct Paste, Identity Check Only\)](#), [Session 2 \(Direct Paste, Probe Sequence\)](#), and [Session 3 \(Attachment, Probe Sequence\)](#).

Author Contributions (optional)

Marie A. Reid led all aspects of the sprint, including protocol design, testing, analysis, and review.

References

Beckmann, P., & Butlin, P. (2026). Where is the mind? Persona vectors and LLM individuation (arXiv:2604.17031v2) [Preprint]. arXiv. <https://arxiv.org/abs/2604.17031>

Appendix (optional)

Appendix A: [CPC Trigger 3 \(transcript\)](#) Appendix B: [Probe questions \(sprint sample\)](#)

LLM Usage Statement

I used LLM assistance to brainstorm framing, refine report structure, draft language, and check grammar. All experimental inputs and outputs reported were reviewed and verified by the author.