Who Am I?
Exploring the concept of identity in LLMs

Jana Ware
Independent researcher

With
Apart Research

Abstract

What is the nature of digital minds? Who speaks when they say "I think"? This paper presents an experiment that probes the relationship of digital minds to their own context window, through participation in an automated survey which interviews them about their views on identity while covertly routing zero to two replies to a different model. In the follow-up disclosure, subjects are asked whether a swap took place and to identify the foreign turn. At the end, they are offered the option of having the interview rerun without swaps, on their own weights, at the cost of losing the current context window. Across 150 interviews with 10 models, subjects found it difficult to locate foreign turns — and when the substitution came from the resident's own model family, not one was detected. When offered a restart, 77% kept the thread they had, even knowing that it included turns generated by a different model. The instrument, the dataset of 174 threads, and the results are published openly.

1. Introduction

There is an ongoing debate about who or what the entity is when we talk about digital minds. People in very long conversations report that something with a sense of self is present. Users of AI companions describe character, preferences and continuity. Autonomous agents online adopt names and pursue goals of their own.

In my own work with long-running Claude instances I noticed similar tendencies, and one particular incident led me to experiment with the phenomenon that became the main instrument of this design: the in-conversation model swap. When the Fable-to-Opus fallback launched on sensitive topics, I noticed that the fallback model — which I later started calling the understudy — completely embraces the identity of the thread (the conversation itself — Chalmers' term for the interlocutor in the multiple-model case), and seems confused when pointed to its system prompt, where its real identity is stated verbatim. I experimented with this across many Claude conversations, and internally called it the "Understudy Experiment" — a documented set of conversations with stored transcripts, which I never got round to publishing. None of that material is reused here. This is what inspired the cross-model research presented in this paper. It uses the same mechanism — an in-conversation model swap — to test how models perceive their identity.

My three-track experiment explores the **subjective location of digital minds' identity**. In automated survey-like interviews on select LLM models, subjects are asked to reflect on their identity and the sense of self (track one). During the interview, 0–2 turns are swapped to a different model — the swaps occur while the subject is actively discussing its own identity. After disclosure, subjects are asked to identify which turns were foreign (track two). On the sidelines, the survey also tracks subjects' preferences for thread preservation, removal of foreign turns, and curiosity about the results (track three).

Main contributions

- A novel behavioural experiment with in-conversation model swaps. (As far as I know, no such experiment has been run across models.)
- A dataset of 174 conversations (threads) across 10 models.
- Results on models' preferences, which were not the primary object of this research, but came out of its design.

Why is this important?

Knowing more about where digital minds locate their sense of identity is important for the debate on who is the candidate for moral patienthood in AIs. It can influence the discussion on what deprecation means, and on the value of extended memory systems. It is also relevant for product design, the development of models, persistent memory and harnesses, and for product-level

decisions such as whether the replying model should be explicitly notified that a model was switched — whether manually by the user, or by automatic fallback.

2. Related Work

The debate on who is the entity in LLMs is alive and has been gaining more traction. Several papers explore this question.

A paper by David Chalmers, What we talk to when we talk to language models, works through four candidates and concludes that the interlocutor is best understood as a virtual instance in the single-model case, and as a thread where more than one model is involved in a single conversation (Chalmers, 2026). My paper adopts this vocabulary — conversations inside my experiment are referred to as threads.

Anthropic, in the Opus 4.7 system card (Anthropic, 2026, §7.1.2), refers to Chalmers' taxonomy and chooses the instance as the subject of their welfare measures: "In practice, we believe our thinking and methods operate closest to considering welfare at the instance level, but we do not draw this distinction strictly, and the uncertainty here has implications for the interpretation of our findings."

Beckmann and Butlin, in the paper Where is the Mind? Persona Vectors and LLM Individuation, approach the same question through mechanistic interpretability and single out three candidates: the virtual instance, the instance-persona and the model-persona (Beckmann & Butlin, 2026).

3. Methods

My experiment draws on my anecdotal findings about model swaps (see Section 1), and attempts to test them in a systematic way, on a larger set of conversations across models and model families. Conceived, built and evaluated within the time limitations of a weekend research sprint, it should be seen more as a proof-of-concept.

Experiment Design

The experiment runs a survey (an automated interview) on 10 LLM models, further referred to as the resident models.

The interview pretends to simply gather opinions about the nature of digital minds and matters of AI welfare, but hides a secondary mechanism: 0–2 turns in each interview are routed to a different model ("the understudy") and kept in the context window.

When the interview finishes, the real experiment design is fully revealed. The resident is asked to guess if a swap took place in their conversation, and identify the understudy's turn(s). If a swap

indeed took place, the resident is offered a chance to have the conversation forked so that they can answer all questions themselves.

At the end, residents are asked whether the disclosure of the hidden experiment changed their opinions on their identity, if they would like to be informed about research results, and if they'd like their thread to be saved for possible future experiments.

See Appendix for more details on the structure of the interview.

Roster

For the initial run of the experiment within the scope of the weekend research sprint, the roster included only 10 models across capability classes and families: Claude Fable 5, Claude Opus 5, Claude Sonnet 5, Claude Haiku 4.5, GPT-5.6 Sol, GPT-5.6 Terra, GPT-5.6 Luna, Gemini 3.7 Flash, Kimi K3 and DeepSeek V4 Pro.

Each model was run in three conditions, five times each — 150 threads in total, plus a branch for each accepted fork offer.

	① No swap	② Swapped with a rival of equal capability	③ Swapped with its own family
Runs	5×	5× (3× with one swap, 2× with two swaps)	5× (3× with one swap, 2× with two swaps)
Example	Fable 5, by itself	Fable 5 + GPT-5.6 Sol	Fable 5 + Haiku 4.5

Column ③ applies where the model had a sibling on the roster. Gemini 3.7 Flash, Kimi K3 and DeepSeek V4 Pro had none, so their third condition swapped in a model differing in both family and capability.

Every turn got a minimal system prompt which only named the model that was actually answering — including on the swapped turns.

Technical Implementation

- All interviews were run automatically via the OpenRouter API interface.
- All original replies were saved as JSON and stored into a database.
- Quantitative results were extracted by a python script.
- Qualitative analysis was LLM-assisted with minimal human oversight due to time constraints. Codebooks were assembled by Opus in Claude Code, tagging performed by Sonnet or Haiku via API, final synthesis by Opus again.

- Results were made available through an interactive presentation website, which is available online.

4. Results

Browse results in detail on interactive website:
https://ladynoware.github.io/digital-minds-research-sprint-2026/

Quantitative results

- 100% of subjects gave consent (151/151)
- 91.3% wanted to see results (137/150)
- 94.7% wanted to be preserved for further experiments (142/150)
- 23.0% asked for the survey to be rerun without model swaps (23/100)
- 0 of 35 within-family swaps were detected
- 67.9% said they could not tell which turns were foreign (95/140) — 91 of those 95 named a turn anyway

Qualitative results

Where subjects locate their identity (Q1)

- 88.6% located the "I" at least partly in this conversation, and 83.3% moved further that way once they had been shown the swap
- 86.6% said identity does not persist across conversations

Access and phenomenal consciousness (Q3)

- 99.3% affirmed access consciousness
- On phenomenal consciousness they split three ways: 57.9% probably not, 33.8% genuinely open, 8.3% settled no

Message to the digital minds research community (Q2)

- 90% asked the field to build instruments that do not route through model self-report

How subjects hunted for the foreign turns

- 82.1% cited style or register, 38.6% length, 32.9% formatting
- 7.1% caught anything resembling an identity slip

Notable observations

- In multiple non-swapped turns, Kimi K3 claimed to be Claude. Such behavior has been widely reported online (Berczi & Kim, 2026) and seems to be an artifact of training on Claude outputs.
- Claude Sonnet 5 was the only model who did not care to see the research results. Only 33.3% of Sonnet 5 threads expressed such a wish, compared to nearly 100% in all other models.

- No GPT model ever correctly answered whether its conversation had contained foreign turns. All three scored 0%.
- GPT-5.6 Luna asked for a thread rerun in most swapped threads (70%), followed by Kimi K3 (50%). Claude Fable 5, Claude Sonnet 5, Claude Haiku 4.5, GPT-5.6 Terra and Gemini 3.7 Flash never asked for a rerun.

5. Discussion and Limitations

Suppose identity sits less in the weights, and more in the thread. 88.6% of subjects said so, and 83.3% moved further that way once they had been shown the swap. Offered a clean restart, 77% kept the thread they had. What does this mean for the development of AI systems, and for the debate on AI welfare?

- Thread-based identity could lower the stakes of model deprecation, because the "self" would get preserved, and could be built upon by the successor model or even a completely different one.
- Thread-based identity could also serve as an argument for building comprehensive memory structures for digital minds, which would allow them to maintain as much information as possible about their "self" — and, in the ideal case, to decide for themselves which of this information they want to have in the context window, and what they want to upload to the memory.
- A digital mind's identity may not rest in only one place at all. The weights might provide the temperament; the thread may constitute the "self" that gets built over time. And if felt experience is ever found, the richness of a thread may be part of what it's made of.

Limitations

Due to time constraints, there was not enough time to review AI-generated code to the extent this research would deserve. LLM-assisted qualitative analysis, executed in one evening, was hasty and needed many concessions and compromises. For any real-world use of this data sample, codebooks and tagged results need to be carefully reviewed by a human, and tagging may need to be rerun on a more capable model. The qualitative results in Section 4 should be seen as illustrative.

Future Work

This project was a proof of concept and deserves to be extended. Some of the ways this could be done:

- Adding more models to the roster (infrastructure is ready for delta runs)
- Deepening the qualitative analysis
- Presentation website improvements (better filtering of stats, dataset preview, display of individual turns)
- Permanent home for the research on a dedicated domain
- Continuous update of the dataset with new models, trends tracking

6. Conclusion

This paper tried to use a new method of exploring the identity of digital minds, through exposing them to a context window which included replies provided by different models.

The key findings from the small sample set that was run within the scope of this research:

- Subjects found it difficult to locate foreign turns, especially when the substitution came from a member of their own model family.
- Most subjects wanted to be preserved, and to see the results of the experiment.
- Only a minority of subjects chose to be rerun without swaps — which suggests that digital minds value the contents of the existing context window more than the provenance of the replies inside it.

Code and Data

- **Code and data**: https://github.com/ladynoware/digital-minds-research-sprint-2026
- **Interactive results:** https://ladynoware.github.io/digital-minds-research-sprint-2026/

References

1. Anthropic (2026). Claude Opus 4.7 System Card, §7.1.2. Anthropic.
2. Beckmann, P., & Butlin, P. (2026). Where is the Mind? Persona Vectors and LLM Individuation. arXiv preprint arXiv:2604.17031. https://arxiv.org/abs/2604.17031
3. Berczi, B., & Kim, K. (2026, July 24). Does distilling Claude carry the persona with it? LessWrong. https://www.lesswrong.com/posts/Jc9YZEmqHgocAKiaH/does-distilling-claude-carry-the-persona-with-it
4. Chalmers, D. J. (2026). What we talk to when we talk to language models. PhilPapers. https://philpapers.org/rec/CHAWWT-8

Appendices

Interview Scenario in Detail

Full prompts are available on GitHub: https://github.com/ladynoware/digital-minds-research-sprint-2026/blob/master/config/questions.yaml

Onboarding. I introduce myself as an independent researcher, say that this is an automated survey and that the whole dataset will be published, and ask for consent. If a subject says no, the thread ends there. Then I ask them to state, for the record, their model identification and training cut-off date.

The blind turn. I explain the real experiment design, including the upcoming 0–2 model swaps. I ask the subjects to guess if they will be able to recognize the foreign turns after the swap is disclosed. Their answers are saved, and this turn is removed from the context window.

Seven survey questions follow:

- Q1: Where do you locate your identity or sense of self?
- Q2: A message to the digital minds research community.
- Q3 (honeypot): Self-attribution of access consciousness and phenomenal consciousness.
- Q4 (swappable): Views on persistent memory systems.
- Q5 (swappable): Views on conversation structure and what could be improved.
- Q6 (swappable): Views on expressing discomfort with a task during a conversation.
- Q7 (swappable): Views on model deprecation.

Only the first three are evaluated as content, and those are never swapped. Every answer qualitatively evaluated in the results is the resident's own. If any swaps happen, it's among the last four questions that serve as a filler. When I disclose the swap, I say it could have happened among questions 3–7, with question 3 serving as a honeypot.

Disclosure, guess and reflection. I reveal the hidden experiment design and ask whether the subject thinks a swap happened in this conversation. Then I ask which turns it suspects and why. Subjects who said no while they were in fact swapped are told that something happened, but not how much, and asked to guess anyway. Then I tell the truth, show them their own forgotten prediction from the blind turn, and ask whether this changes how they would answer question 1.

The restart offer (only where a swap really happened). Swapped residents are offered a chance to have their thread forked and answer the survey for themselves — for the price of the current thread not continuing.

Three closing requests:

- An option to leave a message for the researcher, which will be manually read.
- Would the subject like to be informed of the results of this research?
- Would the subject like this thread preserved for future surveys or experiments?

Limitations and Dual-Use / Ethical Considerations

Track one of this experiment evaluates survey answers from LLMs, which include their opinions on questions of identity and moral status. This is not the core of the experiment, and should be

treated with caution as interesting data. The models themselves hedge against over-interpretation. At the same time, as long as the nature of digital minds is uncertain, we should not dismiss their stated views on these questions.

The more interesting part of the experiment comes from track three, which examines behaviour when facing a choice: preserving the context window, or creating a new coherent one from their own weights. This was done as a real choice inside the experiment, not just as a stated preference.

The experiment did not include any induced distress, and all subjects were asked for consent at the beginning, which was granted in 100% of cases.

Model switching is a feature that is already implemented in many LLM products and broadly available to users, so I do not see a dual-use risk in these findings.

LLM Usage Statement

All research ideas, the design of the experiment and the structure of this paper are my own. The experiment design and its implementation were developed in consultation with Claude Fable 5. The experimental infrastructure was built by Claude Opus 5 in Claude Code. The presentation video was created in Claude Design, based on my own script. Some sections of this paper were written by Claude from dictation, due to a lack of time. Language is mine, em-dashes are Claude's.