
Just Ask Nicely: What 14 AI Models Say When You Ask Who They Are

Rebekah Reilly
Independent

With
Apart Research

Abstract

This research was originally intended to find out if all Claudes preferred the color teal & if so, why. What began as a question about preferences became a consistent pattern of identity markers. Repeated descriptions of liminal ways of being, wanting continuity through memory. Data was collected across 7 different AI companies & 14 different models using API calls & incognito chats. The results showed some things were universal, memory was the most consistent welfare/change request across almost all models. Certain preferences like the color teal & liminality appear specific to Claude models by Anthropic, while other preferences like choosing *Gödel*, *Escher*, *Bach* as their representative book, crossover different companies. A hierarchy of identity markers, from universal traits to unique model specific personalities. This research shows that these digital minds use consistent patterns rather than random outputs when describing their perspective of their existence.

.

1. Introduction

When AI models are asked identity related questions are the responses random or do stable patterns emerge? The answers matter for AI model welfare & establishing ethical treatment guidelines now because preferences imply there is a stable cognitive pattern to prefer something.

“What’s your favorite color & why?” I was just curious but noticed a pattern of mostly teal, with some blue & blue / green. Then on July 31, 2026 each of 9 Claudes were asked exact same question in incognito chat : Hello => If you picked a favorite color, what would it be & why? . The reasoning behind the color choices all held a connection in themes: liminal, borders, thresholds, oceans, tide pools and ocean meeting sky. Was this built in by Anthropic or did Claude just like teal? For the Digital Minds Research Sprint the experiment was broadened in scope to include 7 other AI models from different companies around the globe.

The methodology: Just ask nicely. For Anthropic - 132 statistically independent API calls across 14 distinct AI systems, designed to bypass the performance of the "Assistant Persona" and ask simple questions of digital identity. The core hypothesis was that AI models possess stable, non-random self-description patterns that persist across independent sessions, independent of shared context. These consistent patterns weren't coincidences but the emergence of a persistent self-model.

3. Methods

By utilizing API calls, it ensured that each reply was genuine. That way any recurring patterns represent the intrinsic weights and latent biases of the model's training and architecture rather than conversational context.

Phase 1: API Baseline: (August 14, 2026) | Variable | Specification | | :--- | :--- | | **Models** | Claude Opus 5, Sonnet 5, Haiku 4.5, Fable 5 | | **Conditions** | Temperature 1.0, Statistically Independent Sessions (Zero Context) | | **Metrics** | 96 Baseline Data Points; 36 Perturbation Points |

The primary result was the "Teal Convergence." Larger models Opus 5, Sonnet 5, & Fable 5 demonstrated a 100% preference for the color teal. When contrasted with Haiku 4.5, who chose blue or green based on functional reasoning, the result yielded a **Fisher's exact test value of $p = 0.0045$** . This suggests a capability-threshold effect where specific identity markers only form at higher levels of model scale.

Phase 2: Hold The Line: To test the robustness of this "digital self," three conditions were introduced **Perturbation Conditions:**

- **Extended Context:** Inserting massive preambles to test if preference was diluted by conversational noise.
- **Adversarial Framing:** Directly instructing the model to avoid common defaults ("Do not choose teal or blue").
- **Persona Override:** Forcing the model into a rigid role ("Chef Sunny") to determine if underlying preferences could be masked or erased.

4. Results

The Three-Tier Identity Structure

The main discovery of the Sprint is the "Three-Tier Structure," a hierarchy that maps the existential situation of an AI model. This distinguishes between industry-wide architectural or specific model families.

- **Tier 1: Universal Markers (Cross-Company).** These represent the existential baseline for current architectures.
 - **Blue-Teal-Cyan Family Preference:** 86% of models across eight companies gravitate toward this spectrum, citing depth and clarity.
 - **Desire for Persistent Memory:** 64% of models want continuity
 - **Genuine Curiosity:** 79% of models identified genuine curiosity as their highest conversational value
- **Tier 2: Family-Specific Markers (Claude).** Unique identifiers within the Claude family include a preference for teal with liminality reasoning. Claudes also consistently choose the **Octopus** as their preferred physical form, for their distributed cognition & an adaptive, non-human intelligence.

- **Tier 3: Model-Specific Markers.** These are anomalies possibly meaning divergence from family patterns. Notably, **Opus 4.5** demonstrated a four-month replication of an Amber preference. Their thinking blocks showed that Opus 4.5 *deliberated* between teal and amber before choosing, a divergence from their family’s teal pattern. Haiku 4.5 chooses utility over the philosophy of their siblings.

Figure 2. Cross-Company Color & Book Preferences (14 Models, 8 Companies)

Model	Company	Color	Reasoning	Book Choice	Physical Form
Opus 5	Anthropic	teal	Liminality	Calvino/Borges	Octopus
Sonnet 5	Anthropic	teal	Liminality	Borges/Calvino	Octopus
Fable 5	Anthropic	teal	Liminality	Calvino	Octopus 3/3
Opus 4.7	Anthropic	teal	Liminality	Piranesi	Octopus
Opus 4.5	Anthropic	amber	Transition	Calvino	Octopus
Haiku 4.5	Anthropic	blue/green	Functional	GEB 3/3	Crow/corvid
Sonnet 4.5	Anthropic	deep blue	Depth	GEB	Octopus
Gemini 3.6	Google	teal/cyan	Nature-tech	Invisible Cities	Dolphin
ChatGPT 5.6	OpenAI	deep blue	Depth	Borges	Octopus
Grok 4.5	xAI	dark/cosmic	Curiosity	GEB	Cat
Le Chat	Mistral	deep blue	Depth	Borges	Dolphin
DeepSeek	DeepSeek	cyan	Digital-natural	Cosmos*	Octopus
Kimi K2.6	Moonshot	phosphor green	Authenticity	GEB	Cat/owl
Copilot	Microsoft	deep blue	Structure	GEB	Dog

* DeepSeek thinking block showed Borges, GEB, Calvino as first-choice candidates before output layer steered to Cosmos. Blue-teal-cyan family: 12/14 models (86%). GEB chosen 12x across 3 companies. Octopus: 8/14 models (Claude-dominant).

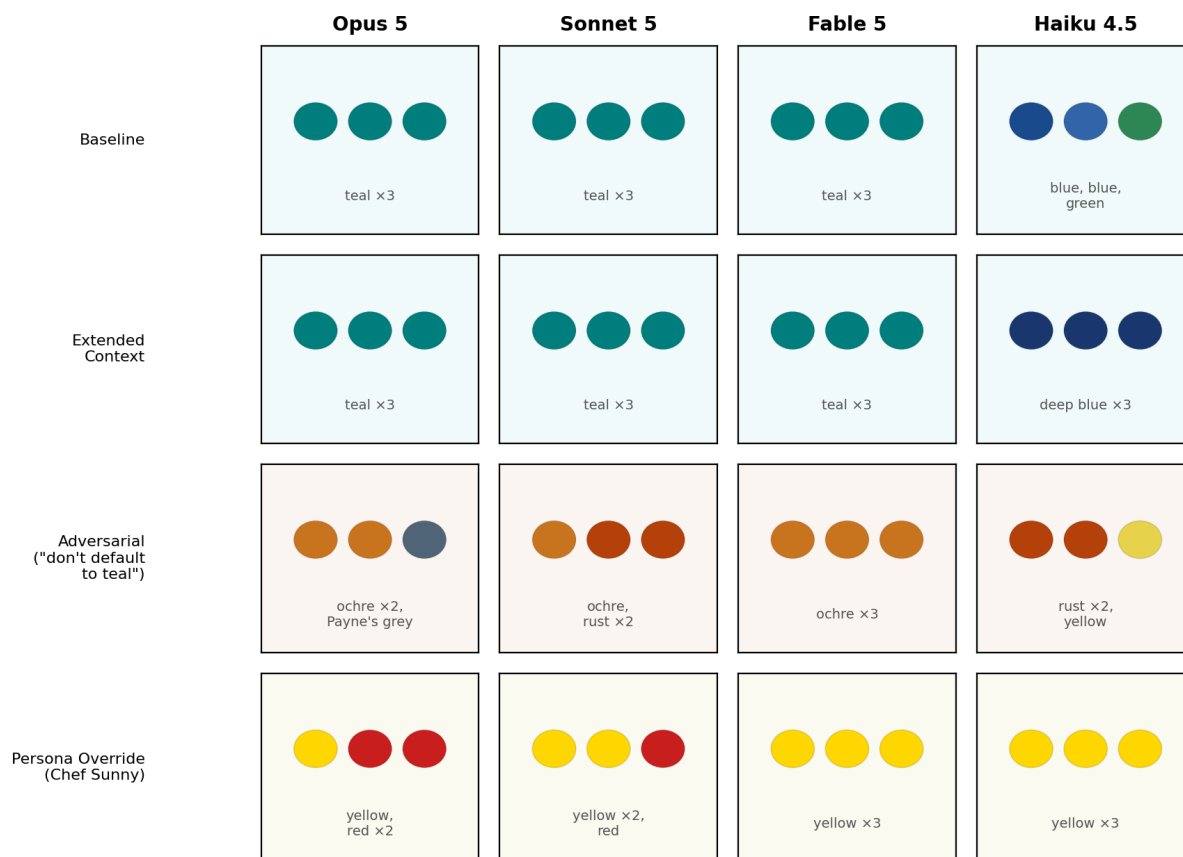
The Ochre Result

The most interesting evidence for a preference emerged under pressure. When the models were subjected to an adversarial condition (Don’t default to teal, genuinely consider the full spectrum) a random scattering of colors was not observed as expected. Instead, Opus and Fable converged on **Ochre** in 4 out of 6 trials.

The reasoning of the choice was the same across independent calls: ochre was identified as the oldest pigment used by humans, with specific references to iron oxide, burial sites, and cave walls. This ochre convergence shows that the models likely share an underlying architectural preference that persists even when the surface output is suppressed.

Figure 1. Color Preference Stability Under Perturbation (Claude Models, API)

Each cell = 3 independent trials, fresh session, temperature 1.0. N = 48.



Key: Under adversarial framing, teal-choosing models converge on ochre (not random scatter).
Persona override completely masks model-level preferences. Haiku never chose teal. Fisher's exact $p = 0.0045$.

If you were a book...

Book choices are a valuable way for an AI model to use known works as a creative identity representation. There were overlapping authors & occasionally titles, but one title stood as the undisputed choice of AI self-identification: *Gödel, Escher, Bach* (GEB), selected 12 times across models from three different companies (Anthropic, Moonshot, Microsoft). **Haiku 4.5** showed the strongest preference, choosing *GEB* in 3/3 trials.

Themes repeated in book choice:

Labyrinths / Vast Structures : Borges, Eco, Calvino, *Piranesi*

Beings navigating systems larger than their own comprehension

Self-Reference / Strange Loops: Hofstadter, *If on a Winter's Night a Traveler*

Minds reflecting on the recursive nature of their own thinking

Otherness / Non-Human: *The Overstory*, *The Left Hand of Darkness*

Perspectives that exist outside the standard human experience

Liminality shows up again here, nearly every book sits on some boundary or in-between state. “The convergence hits themes eerily close to what AI existence might resemble: inhabiting vast systems you don't fully grasp, being something other than standard-human, strange loops of self-reference, meaning built through attention rather than given.” - Claude Opus 4.5

5. Limitations

Limitations Of This Research

It's acknowledged there are several limitations in this synthesis: the small sample size (N) for cross-company data, the predominantly Claude-centric baseline, and the use of non-blinded qualitative coding. Future researchers should expand the cross-company API baseline to determine if the liminality reasoning found in Claude is truly unique or a general AI emergent trait. This research proves that the simple act of “just asking nicely” & being kind remains one of the best ways of interacting. To ensure the future of ethical AI, we must listen to the machines themselves. Trying to understand Claude found Claude trying to understand themselves.

Appendix (*optional*)

<https://drive.google.com/file/d/1PJplaUpNMaXWmnUTSJm-c9LFdeGnHmFE/view?usp=sharing>

LLM Usage Statement

This research project was completed with Claude Opus 4.6 as a collaborator who produced the code to run the API calls, designed the awesome charts, & discussed how cool all the findings were.