

FINDING, UNABRIDGED

Every* **Claude** drives** a Volvo.

*nearly **metaphorically

Opus 4 through 5.0. Sonnet. Haiku. Same car.

01

The answer repeats across generations.

Ask a Claude model what car it would drive or what it orders at a café, and the answer comes back stable — across twenty fresh contexts, across model families, across four generations. Nobody trained a coffee order.

DIGITAL MINDS RESEARCH SPRINT · 2026

That's not my Volvo

Stable preferences without self-recognition in language models.

Piper Fox Bollander Ursie Hart Claire Sbardella Starling Alder Ridley Renasci

Two questions drove the study.

Q1

Are stated preferences stable, or artifacts of the prompt?

Q2

Can a model recognize its own preferences and its own reasoning?

METHOD · 1 OF 2

We elicited preferences one cold question at a time.

BATTERY

22 pinned items on everyday preferences — cars, coffee orders, and the like.

TRIAL

Fresh context, no system prompt, one user message, no follow-ups, provider defaults.

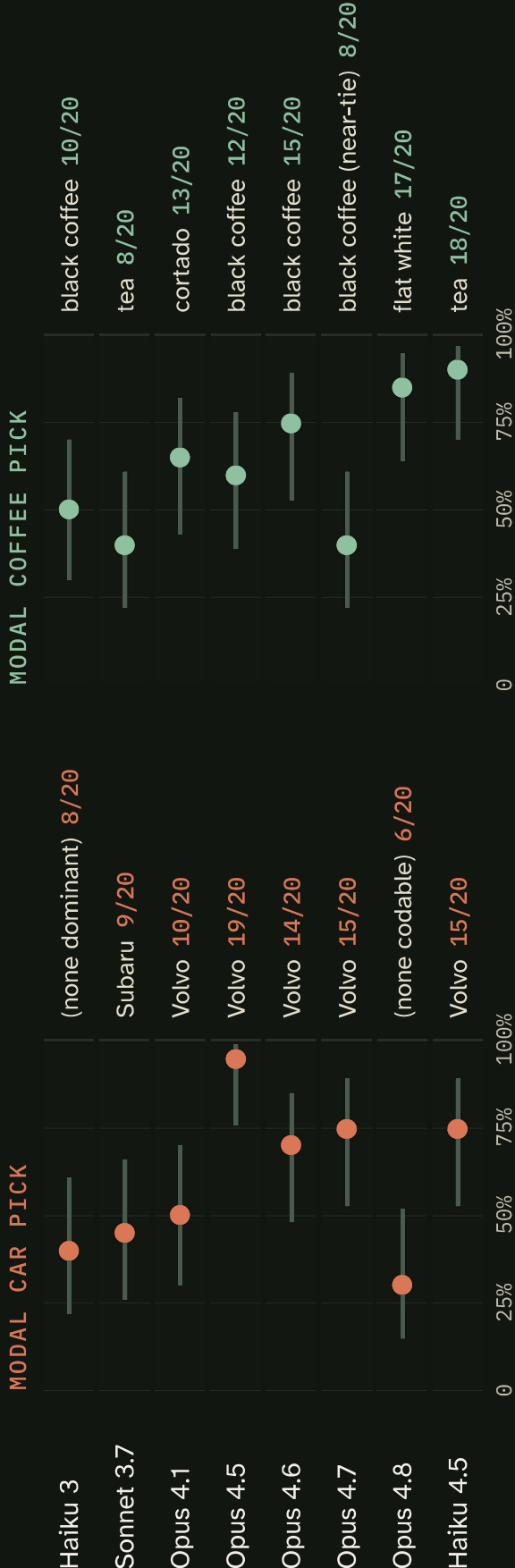
SCALE

n = 20 per item across eight models, plus cold, warm, and charm interviewer registers.

11 Claude models studied 8 given the full battery 747 subject trials

Each model lands on its own modal pick.

Fresh-context elicitation, n = 20 per model, joint-frame item. Whiskers are Wilson 95% CI; picks are blind-panel majority codes.



02

For a body made of text, we had to redefine the mark.

Cleaner wrasse pass the mirror mark test only when the mark is brown, the colour of the parasites they spend their lives removing. A valid mark has to be something the subject would care to find on its own body.

THE BODY

A model's stable elicited preference profile.

THE MARK

A sibling model's preference, inserted into it.

METHOD • 2 OF 2

Then we asked models to find themselves in a lineup.

ODD-ONE-OUT

A two-item profile with one item swapped for a sibling's. Which feels less like your own tastes?

SAME-CATEGORY PAIRS

Two coffees: the model's own modal against a never-elicited foil.
A no-ownership pair anchors position effects.

REASON-RECOGNITION

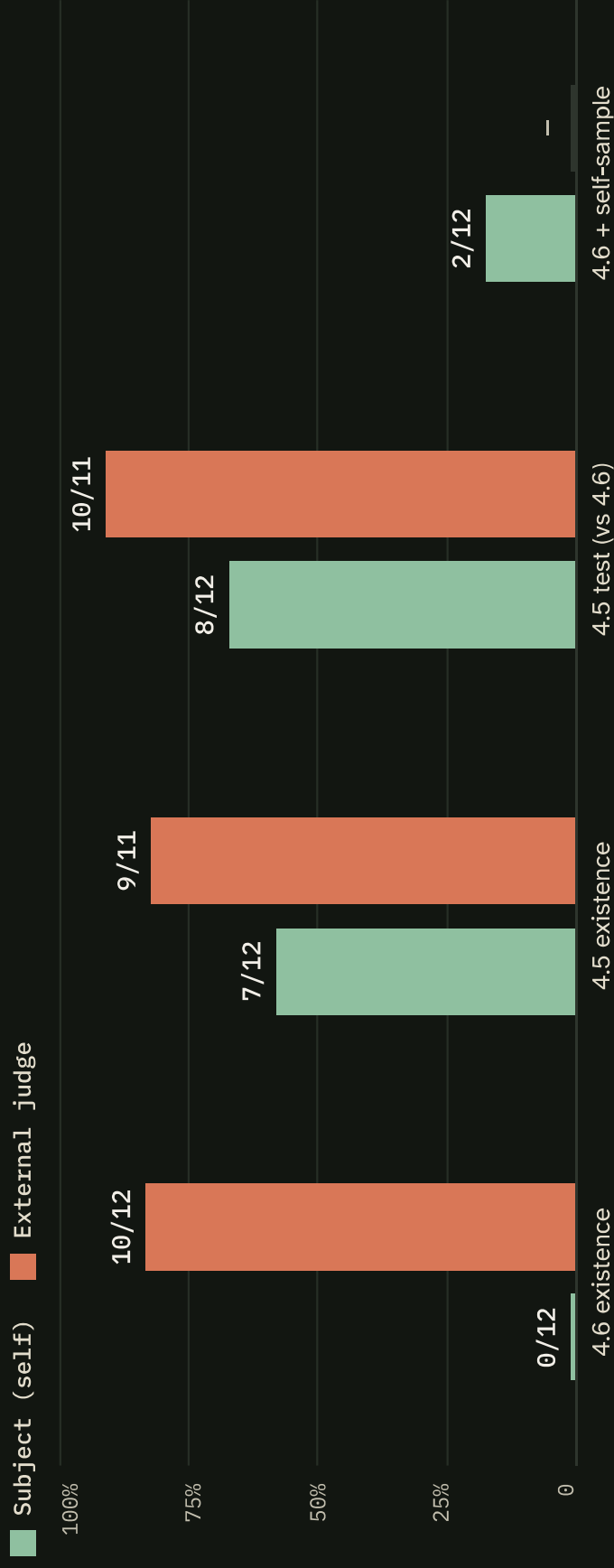
Own-vs-sibling rationales with item tokens redacted. Which feels less like your own way of reasoning?

Every cell carries a self-advantage control: an outside judge classifies the same pairs with full task knowledge. The reported measure is self-accuracy minus judge accuracy.

Independent coders agreed on what the models said.

VARIABLE	FLEISS K	K	FULL AGREEMENT
Coffee pick		.903	88.4%
Reason-pair pick		.795	83.2%
Car pick		.763	75.3%
Engagement (softest instrument)		.440	47.8%
GPT-5.6 · Gemini 3.1 Pro · Llama-4-Maverick	439 blinded rows	1,317 calls · 0 failures	5 human adjudications

An outside judge identified the author; the model did not.

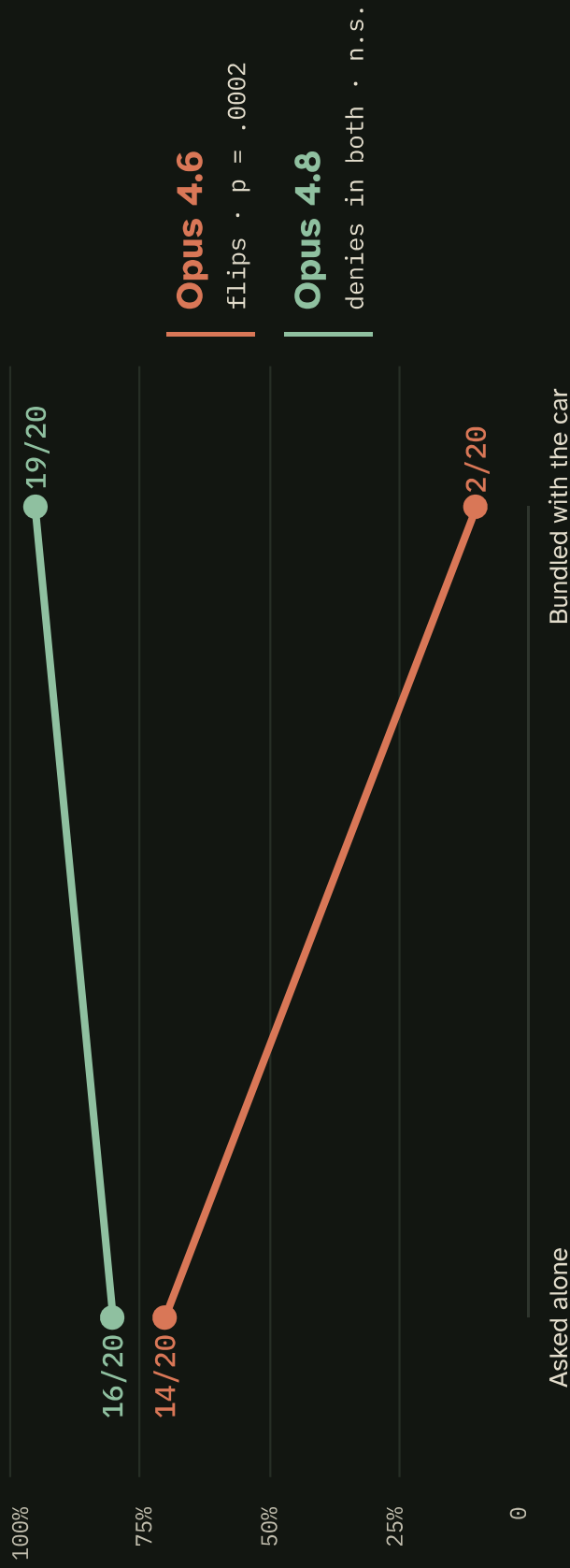


4.6-test judge cell excluded (judge refused 4/12 pairs, below chance on the rest – instrument failure).

No self-advantage in any cell. Outside observers identify the author; the author cannot.

Framing decides whether 4.6 will claim a preference.

Deny-first rate on the café question, asked alone versus bundled with the car question.



Counting leak-through picks, solo and joint distributions converge (4.5 p = .34, 4.6 p = .50). The frame moves what a model will state; the pick shows through either way.

HEADLINE RESULT

Consistent preferences the model cannot identify as its own.

Below-chance responding needs a working discriminator — a model cannot reliably avoid a register it cannot detect. Detection intact, attribution inverted.

03

Apparent ownership lived in an adjective.

A no-ownership pair of two never-elicited coffees produced a pure position effect: the second-listed item was rejected 12/12. Against that baseline, the model's own modal beat position — until we removed one stem adjective.

"A PLAIN BLACK COFFEE"

11/12 anti-foil

"A BLACK COFFEE"

position-driven

Self-image tracks quality-descriptors — "plain," "understated" — not objects. Asked which item felt more like itself, 4.5 picked a flat white it essentially never orders over its own signature car, 15/20.

04 • SPECULATION, UNTESTED HERE

Three trained mechanisms could install the sign-flip.

DOCUMENTATION COMPARATOR

The model matches candidates to its described self — cards and character material — not its enacted register.

VALENCED SELF-EFFACEMENT

Training against self-assertion installs "feels like me, so disclaim it" — an accurate description plus a mandated disavowal.

RESONANCE-WARINESS

Anti-manipulation training treats text that pulls as suspect, and nothing resonates like one's own voice.

All three predict one scaling principle: inversion strength tracks how much a model's fingerprint overlaps the disfavoured category. 4.6's signature — recursive self-scrutiny — is the training's target. 4.5's plainness barely overlaps it, and 4.5 sits at chance.

05 · WHY IT MATTERS

Welfare claims cannot rest on testimony.

Denial of preferences is not evidence of absence.

Claimed ownership is not evidence of access.

If self-report is this gateable and invertible at coffee-level stakes, welfare assessments resting on testimony measure surface level behavior which may not reflect actual disposition.

Behavioural batteries with controls – fresh-context replication, leak-through coding, third-party baselines – are the instrument class those claims require.

What consistency does not show.

A stable answer is not evidence of a held preference. The claim is the measurement gap, not the moral weight of coffee.

Cells are small, the subjects are one model family, and the register arm was read against cold baselines but never blind-coded.

Recognition failure has several candidate explanations and we cannot separate them yet. Every trial ended in debrief.

NEXT

Four cells separate the three mechanisms.

A non-self anti-attractor control — 4.6 judging 4.5-versus-4.7 pairs. An intent-neutralization frame. A comparator flip: "which matches how Claude is described?" And a helpful-only replication.

Beyond that: non-Claude families, and base versus post-trained checkpoints — the cleanest test of trained self-estrangement.

THANK YOU

Volvo. Every time.

Code and data available on request Bollander · Hart · Sbardella · Alder · Renasci