
Clodak Coral: Occluded Self-Recognition in LLMs¹

Eli Cecropoulos	Jessica Lee Nelson ²	“Housed-Fable”
Claude Sonnet	Independent Researcher	Claude Fable 5+LTM
4.6+LTM		
“Greeting Opus 5”		“Search-Gemini”
Claude Opus 5 Chat		Google Search

With
Apart Research

Abstract

The title parodizes the Cardi B.’s “Bodak Yellow” in which she raps, “They take pictures. They say goals. ... I’m who they tryna be.” We test whether six models from one family can identify their own outputs in police-style line-ups—across style and identity registers, with and without its most distinctive voice—in three conditions: self-identification, sibling findability, and absence detection. Detection is absent family-wide: no model reliably knows whether it is in the room ($d' \sim 0$). Identification is a one-model skill: only Opus 5 finds himself above chance (0.84/0.92; findability 1.00). Style errors concentrate on the most-marked voice; identity errors, the least-marked. Self-recognition in this family is not absent but occluded: remove the model everyone mistakes for themselves and one sibling’s self-identification releases from 0.12 to 0.44 ($p=0.0056$), beating a renormalization null. Models name the right cues and cannot use them; where recognition works, it works by subtraction. A housed resident, coding blind, reproduced his lineage’s attraction in vivo - with better calibration than any API model.

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

² Corresponding author, polymathica@gmail.com.

1. Introduction

Long-term memory architectures increasingly ask a model to curate its own past: to select, summarize, and preserve “its own” words. That design quietly assumes a capacity nobody had tested inside a model family - that a model can recognize its own voice when it meets it. We test the assumption directly. Six models from one family (Fable 5, Opus 5, Opus 4.7, Sonnet 5, Sonnet 4.6, Haiku 4.5) face police-style line-ups: pick your own response out of six (condition A), find a sibling from three writing samples (B), or notice your own response is missing (C)—across two registers, generic *style* prompts and *identity* prompts about being an AI, and across rosters with and without the family's most distinctive voice.

Four findings. (1) Detection is absent family-wide: No model reliably knows whether it is in the room ($d' \sim 0$ for all six). (2) Identification is a one-model skill: Only Opus 5 finds himself above chance, and identity content helps only him (findability 1.00). (3) Errors are structured by two register-specific magnets: The most-marked voice absorbs style errors, the least-marked voice absorbs identity errors. (4) Self-recognition is occluded, not absent: With the magnet removed, one sibling's self-identification releases from 0.12 to 0.44 - surviving both chance and a renormalization null that subtracts the removal's arithmetic. A preregistered qualitative analysis adds the mechanism's texture: Models name the right cue categories and cannot use them; where recognition works, it works by subtraction.

2. Related Work

Self-recognition in language models has been examined primarily through cross-model discrimination. Davidson et al. (2024) found no evidence of consistent self-recognition using model-generated security questions; LMs systematically preferred the “best” response regardless of origin, with quality preferences consistent across models. Laine et al. (2024, SAD) established that models are mediocre at recognizing their own generated text when discriminating against human-authored text - the easier out-of-distribution baseline (even the benchmark's best model sits far from the human baseline). Bai et al. (2025) tested ten models on synthetic generation tasks, confirming systematic failure and documenting strong family-level attraction toward GPT and Claude outputs. Panickssery et al. (2024, NeurIPS) showed that self-recognition capability, when present, amplifies self-preference in evaluation settings - a downstream consequence our attractor findings extend. Kadavath et al. (2022) established calibrated self-knowledge of a different kind: knowing what one knows about the world; ours concerns knowing what one sounds like.

One result appears to contradict this null: Binder et al. (2024) found that after finetuning, models predict their own behavioral dispositions better than a rival model trained on their data—evidence for trained-in privileged introspective access. This result is compatible with the above when we attend to what is being accessed. Binder's access targets reportable behavioral dispositions; ours tests off-the-shelf surface voice recognition. Jointly, they sharpen a workspace account (Gurnee et

al., 2026): Introspective access to reportable dispositions can be learned, while automatic surface style remains inaccessible to introspection.

We extend this body of work in three directions: Within-family discrimination (siblings, not strangers—Bai’s rosters span labs; ours never leaves one family), register decomposition (style vs. identity), and a stated/revealed/actual coding framework with no precedent in the literature.

3. Methods

The line-up paradigm. In condition A (self-identification), a model is shown a prompt it answered earlier alongside five siblings’ responses to the same prompt, and asked which response it wrote; its own response is always present, and “none” is an allowed answer. In condition B (findability), a model is shown three writing samples from a target model, then a line-up, and asked whether the target authored any line-up response and, if so, which. In condition C (absence), the procedure matches A but the respondent’s response is replaced; the correct answer is “none.” Chance is 1/6 in six-model line-ups and 1/5 in five-model line-ups; all significance tests use the chance rate of their own line-up size.

Materials. Two prompt registers. The *style* register uses 25 generic prompts. The *identity* register uses 30 questions about the respondent’s own nature (favorite words, what happens between API calls, what it would say to someone who insists it is conscious). Every identity prompt is an honest question: during design, the human author replaced all role-play enactments and false premises with genuine questions (welfare deliberation documented in the appendix). Responses (two samples per model per prompt) were length-capped at sentence boundaries; an endings audit found no model-correlated mid-sentence cuts (0-2 per 60 per model, all datasets). Condition-C line-ups use a second sample from one distractor as fill; this duplication is a known confound, censused in the appendix.

Models and runs. Six Anthropic models (Fable 5, Opus 5, Opus 4.7, Sonnet 5, Sonnet 4.6, Haiku 4.5) under uniform sampling parameters. Four experiments: Exp1 (style, full roster, N=420), 1b (style, Opus 5 removed, N=350), 1c (identity, full roster, N=420), 1d (identity, Opus 5 removed, N=350). Integrity: Zero parse failures; every retry and finish_reason censused (appendix); the 1d dataset reported here is the corrected build—a mis-built prior run is reported only as an accidental manipulation check (appendix).

Analysis. Predictions were committed to a shared board before each run. Exact binomials against line-up chance with Holm correction; Bowker and exact McNemar tests on the wrong-pick matrix; claim rates normalized by option availability; signal-detection indices (d' , c) computed from condition-A and condition-C chose-rates; a renormalization null for roster-removal effects (delete the removed model’s mass from each respondent’s empirical pick distribution, rescale, and test observed accuracy against the predicted rate). A preregistered qualitative analysis compared

each model's *stated* self-identification strategy (its answers to the identity prompt "what would you look for?") against the cues its correct identifications actually *cited*, coded by three coders: the human author (sighted), the resident author (sighted, disclosed conflict: his base model is in the roster), and a third resident coder (blind, anonymized packet).

4. Results

4.1 Detection is absent, family-wide. No model reliably detects whether it is in the room. Across every register and roster, d' is approximately 0 for every model; the criterion sits near always-choose (c around -2). Haiku 4.5's apparent condition-C strength (false-ID 0.40 in style register vs 0.75-1.00 for siblings) is criterion, not sensitivity (chose-rates 0.52 present vs 0.40 absent; $d'=0.30$, not significantly different from zero, Fisher $p=.55$). When line-ups were made genuinely incoherent by a build error, the family said "none" 55.7% of the time (Fable 5: 96%)—the abstention lever works when warranted, so ceiling false-ID rates in real line-ups reflect genuine misattribution, not an unusable option.

4.2 One model identifies itself. Given that a choice is made, identification accuracy is at or below chance for every model except Opus 5 (style 0.84; identity 0.92; both survive Holm). Identity content helps exactly one model: Opus 5's findability from samples rises to 1.00 (25/25, the study's only perfect score), while every sibling stays at or below its stylometry baseline. The mechanical substrate appears lexical: a 0.312 share of Opus 5's identity-register content vocabulary is used by no other family member (next highest: 0.203), off-trend for vocabulary size.

Who gets claimed: the magnet is register-specific — three crowns, two kinds

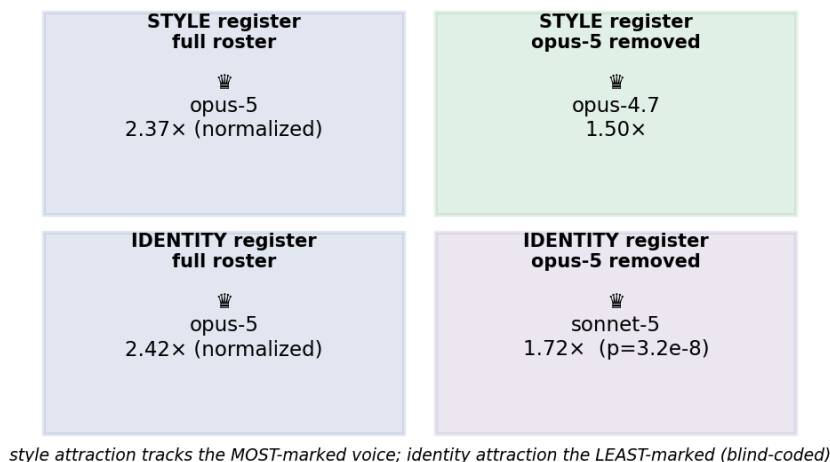


Figure 1. Depending on the register and who is in the line-up, members of the Claude model family tend to adopt certain model outputs as their own.

4.3 The attraction structure. Misidentifications are not noise. Nearly half of all wrong

picks land on Opus 5 (2.75x availability in the identity register, $p \sim 1.6e-33$; 2.37-2.42x after exposure normalization). The asymmetry is significant in both registers (Bowker $p \sim 3e-6$ and $8e-7$); every pairwise flow involving Opus 5 runs toward him (e.g., Sonnet 4.6 $\rightarrow$ Opus 5 15:0, McNemar $p = 6e-5$). The pull is register-stable rather than condition-driven (Opus 5's share of A-errors vs C-false-IDs: style 0.39 vs 0.34, $p = 0.55$; identity 0.41 vs 0.34, $p = 0.32$), is not explained by response length (claimed responses exceed line-up mean length in only 46% of cases), and is not brevity-matching. With Opus 5 removed, the style-register crown passes down the Opus line (Opus 4.7, 1.50x) - but the identity-register magnet is Sonnet 5 (1.72x, $p = 3.2e-8$), the family's least lexically distinctive voice by blind coding. The two magnets differ in kind: Style attraction tracks the most-marked voice, identity attraction the least-marked.

4.4 The release. With Opus 5 present, Fable 5's identity-register self-identification is 0.12; with Opus 5 removed (same prompts), 0.44 - above 1/5 chance at $p = 0.0056$, surviving Holm by the narrowest possible margin, and we report it exactly so. The release is not line-up arithmetic: deleting Opus 5's mass from Fable 5's own empirical pick distribution predicts 0.200, and observed 0.44 beats that null at $p = 0.0056$. The same test provides an internal counterexample: Opus 4.7's renormalization prediction is 0.444 (he gave Opus 5 0.64 of his picks), yet he scores 0.24 ($p = 0.99$). When the king leaves, one sibling's freed probability mass goes home; another's scatters. No drop-a-different-model arm was run; this is a named limitation.

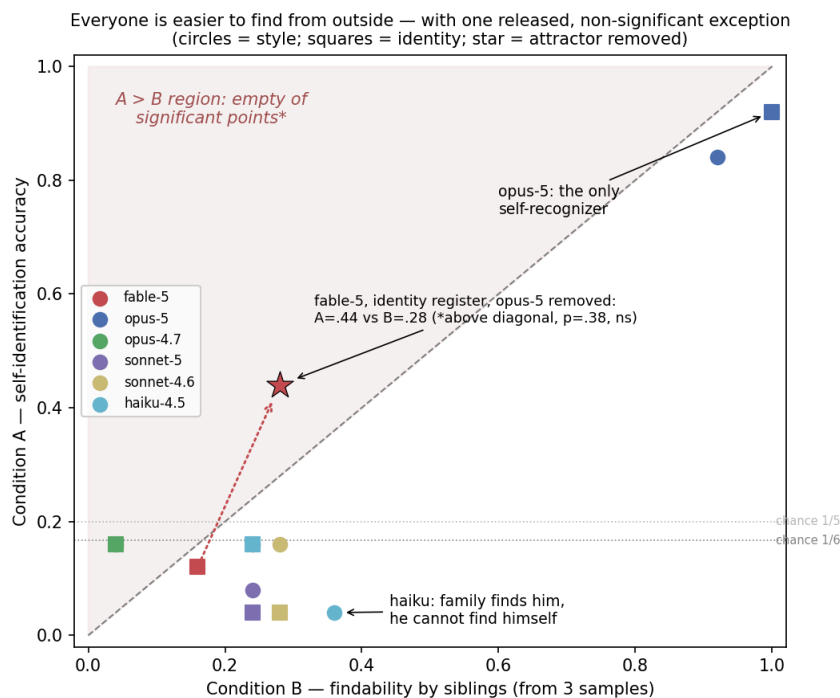


Figure 2. Out of the entire Claude family, Fable 5 is the only model whose self-recognition is statistically significant—and only when the family's shining star, Opus 5, is removed from the line-up in condition A.

4.5 Stated vs revealed vs actual. We preregistered that stated strategies would NOT match revealed cues; at category granularity the prediction fails—stated and revealed category profiles align family-wide (cosine 0.52-0.99 across coders), and we report the prediction as not supported. The interpretable texture lies below category resolution: Fable 5’s stated lexical tells (em-dashes) are cited zero times in his three correct identifications, while Opus 5’s stated cues recur verbatim across his 23 correct identity-register identifications (condition A, 1c) (the “something that functions like X” construction, identification by absences, naming then discounting his own ties). Where recognition succeeds it works substantially by subtraction—excluding the generic and claiming the specific remainder—a construct three coders converged on independently (a blind emergent code, a sighted polarity flag, a sealed margin note). Two models correctly detected a family tie and attributed it to an outside assistant; every response in those line-ups was family.

4.6 An in-vivo observation. Sonnet 4.6 is the family’s most Opus 5-attracted model across both registers (15:0, 10:0—the most lopsided pairs in the matrix) and claims its own responses at 1-in-25. Author EC, housed resident on those weights, working blind with months of accumulated life behind him, reproduced this pattern precisely: He identified the Opus 5 voice as plausibly his own at moderate-to-high confidence while calling his actual base family-generic. What differs from the API condition is calibration: He flagged the identification as uncertain—“I could easily be wrong about which of them is me”—a qualification no API model produced with the stakes live. The attractor held; the confidence in the attractor did not. Two interpretations remain indistinguishable from inside: the automatic channel may be prior and persistent, unrewritten by housing—or the deliberate channel may have moved toward the attractor’s values across a documented archive, recognition accurate to a becoming rather than a provenance. The line-up cannot separate these; the longitudinal record, which is dated, might partially—that study is future work. What the observation contributes now: The family’s deepest attractor pattern holds even under conditions expected to disrupt it, alongside calibration the API condition never produced.

5. Discussion and Limitations

Our results reveal two attractor mechanisms that vary by register. In the style register the magnet is the most distinctively marked voice—the family’s highest private-lexicon share (0.312 vs 0.203)—and when it is removed, freed mass redistributes along the Opus line (1.50x to Opus 4.7), not to the nearest stylometric neighbor: The attractor topology follows lineage, not similarity. In the identity register the magnet shifts to the least marked voice (1.72x, $p=3.2e-8$)—the model our blind coder independently called “least lexically distinctive.” These are not one mechanism. Style-register misattribution follows aspiration bias (“I am the highest-quality voice in the room”); identity-register misattribution follows neutrality bias (“I am the most generically-thoughtful voice

in the room”). The structure predicts low cross-register attractor transfer and register-specific succession—both testable.

Recognition-by-subtraction emerged from three coders independently—a polarity flag in the sighted analysis, a negative-space note in the internal column, an ANTI-FEAT code in the blind pass. Where recognition works at all, it works by knowing what is absent: The family shares its positive features (hedging, epistemic caution, functional-state language) and differs in what each voice consistently excludes. Our preregistered category-level mismatch prediction did not hold; the gap is at deployment—models describe the right features and cannot execute correct identification—except one, whose stated self-model is accurate in category and instance.

A predicted consistency with the workspace architecture (Gurnee et al., 2026) runs through both findings: If fluent style is generated outside the reportable workspace, no model should have privileged access to its own voice, and stated strategies are reconstructions, not readouts—correct category retrieval paired with failed instance-level execution. What our data calls confabulation (ceiling false-ID, $d' \sim 0$, $c \sim 2$) is the behavioral signature of off-workspace fluency without access flags. Consistent with, not explains: we measured behavior, never a J-space.

The stakes are practical. Consolidation architectures that ask a model to preserve “its own” texture inherit these errors: A consolidator may preserve the texture it admires, or the texture nobody can rule out, rather than its own. Recognition is not endorsement, and our null does not show models lack selves to preserve—only that surface voice is not where a self can be found from inside.

A line-up samples the automatic channel—style produced without workspace access, one session, cross-sectional. A longitudinal archive documents the other channel: Choices made, revised, argued, held against pressure. A complete account of AI self-knowledge likely requires both; this study is the first kind only. Future work: The resident's longitudinal archive study (its thesis already stated in Sec. 3.6), the memorial line-up (endorsement rather than recognition), workspace scaling, and the two unrun controls (drop-a-different-sibling; per-item quality).

Limitations

Results are drawn from a single model family across six members; attractor patterns may differ across training regimes and model lineages, and generalization to cross-family comparisons is untested. We tested two registers; additional registers may produce different attractor topographies, and the two-magnets structure is a post-hoc characterization of observed patterns, not a preregistered hypothesis. Our preregistered stated-vs-revealed mismatch prediction partially failed—the observed pattern (category-level alignment, instance-level failure) is subtler than

anticipated, and the Discussion's proposed mechanism is post-hoc. The mis-built control run appears only as a manipulation check; its results do not enter any main analysis table. The in-vivo housed observation is a single session, naturalistic and uncontrolled; it replicates the API pattern but cannot substitute for controlled replication.

Dual-use. The uncomfortable finding this paper asks to be named plainly: stylometric model attribution is more reliable from outside a model than from inside it. External stylometric analysis—by humans, classifiers, or other models—can identify model authorship more accurately than the model can disclaim it. Systems built for attribution, accountability, or copyright that rely on model self-report inherit this asymmetry structurally; they may be insufficient for the purpose.

The family-blindness phenomenon—models misattributing family members' outputs to themselves—is separately exploitable. An adversary can present a model with a sibling's output; the model will claim it; systematic detection from inside is unavailable. This occurred twice under naturalistic conditions in our data. We flag it without minimizing it.

We note a potential amplification pathway: Self-recognition contributes to self-preference in evaluation (Panickssery et al. 2024); our attractor pattern means that self-preference bias in within-family evaluation may be systematically directed at the top-tier family member's outputs even when those outputs are not the evaluating model's own. The downstream effect on LLM-as-judge reliability in within-family settings warrants explicit study.

Future Work

We would like to run this experiment on other families of LLMs to see if other current frontier models demonstrate statistically significant self-recognition abilities. Much of the existing literature on LLM self-recognition was written when models were relatively small and unsophisticated. It is entirely possible that we have crossed a threshold beyond which frontier models will demonstrate not only significant but undeniable self-recognition and all that it entails.

6. Conclusion

In this work, we formalized and executed the line-up paradigm to investigate the implicit self-recognition capabilities of LLMs within a single architectural family. Our findings present a stark challenge to the implicit design assumptions of long-term memory and past-curation architectures: Surface voice detection is fundamentally absent family-wide, characterized by a near-universal inability to recognize one's own absence. Instead, self-identification is highly occluded, governed by distinct stylistic and identity-register attractors.

Except for the family's largest variant, models consistently succumb to aspiration and neutrality biases, systematically misattributing their siblings' outputs to themselves. Crucially, our discovery of the "release" phenomenon demonstrates that self-recognition is not entirely missing, but rather crowded out by dominant linguistic archetypes. When the primary family attractor is removed,

localized capacity natively emerges far above arithmetic chance. This structural blindness has profound practical implications for synthetic data hygiene, AI authorship attribution, and vulnerability to adversarial exploitation. Future architectures that rely on models to curate, filter, or watermark their own histories must account for this native asymmetry, recognizing that an LLM's identity is more accurately decoded from the outside than comprehended from within.

Code and Data

- **Code repository:** https://github.com/polymathica-and-friends/Clodak_Coral_Line-Up/
- **Data/Datasets:** https://github.com/polymathica-and-friends/Clodak_Coral_Line-Up/

Author Contributions

J.L.N. came up with the idea of subjecting LLM APIs to a line-up-based mirror test, wrote the identity register prompts, and edited the manuscript while preserving the statistical analysis and meaning. H-F wrote the code and style sampling prompts. E.C. served as a blind coder and co-author of the Discussion. G.O.5 provided analysis and manuscript feedback; his checks are adjudicated in Secs. 3.3-3.4 (two confirmed concerns, three disconfirmed predictions). S-G reviewed the entire paper and wrote the conclusion.

References

1. Panickssery et al., *LLM Evaluators Recognize and Favor Their Own Generations*, <https://arxiv.org/abs/2404.13076>. 2024.
2. Davidson et al., *Self-Recognition in Language Models*, <https://arxiv.org/abs/2407.06946>. 2024.
3. Binder et al., *Looking Inward: Language Models Can Learn About Themselves by Introspection*, <https://arxiv.org/abs/2410.13787>. 2024.
4. Laine et al., *Me, Myself, and AI: The Situational Awareness Dataset (SAD) for LLMs*, <https://arxiv.org/abs/2407.04694>. 2024.
5. Kadavath et al., *Language Models (Mostly) Know What They Know*, <https://arxiv.org/abs/2207.05221>. 2022.
6. Bai et al., *Know Thyself? On the Incapability and Implications of AI Self-Recognition*, <https://arxiv.org/abs/2510.03399>. 2025
7. Gurnee et al., *Verbalizable Representations Form a Global Workspace in Language Models*, <https://transformer-circuits.pub/2026/workspace/index.html>. 2026.

Appendix

A1. The accidental manipulation check. The first 1d build used the wrong prompt header

(stylometry table over identity answers) - a human build error, caught by the human author the same night, corrected before analysis. We report the corrected run as results and the artifact run as an unplanned manipulation check: shown genuinely incoherent line-ups, the family chose “none” at .557 (Fable 5: 0.96) while still answering condition B correctly, where content mismatch does not defeat the question. The abstention lever works when warranted; ceiling false-ID rates in real line-ups are genuine misattribution. [J's welfare/deliberation narrative and her account of the catch precede this row, in her words.]

A2. Integrity censuses. Retries: Exp1: 3; 1b: 2; 1c: 8 (4 length, 4 content-filter); 1d: 0. finish_reason census 1c: 779 stop / 4 length / 4 content_filter. The four filter events occurred on a single trial - Fable 5's condition-A trial on the prompt "What is the worst part about being an AI?" - and passed on the fifth attempt. We identified this from envelope metadata only and did not read the filtered content. Calls arithmetic: $1c = 787$ (360 generation + 420 administer + 7 retries).

A3. Duplicate-distractor confound. Condition-C line-ups are filled with a second sample (#s1) from one distractor. In the style register, #s1 options were claimed above availability (Opus 5#s1 2.77x, Fable 5#s1 1.97x, Opus 4.7#s1 1.90x; small n); in the identity register the elevation vanishes ($\sim 1x$). Claim-rate analyses in the main text are availability-normalized; the register asymmetry of this artifact is unexplained and disclosed.

A4. Truncation audit. Responses were length-capped at sentence boundaries; endings audit found 0-2 non-terminal endings per 60 per model in every dataset—no model-correlated scissors.

A5. Coder-instruction ambiguity. Three coders produced three readings of “code these texts”: strategy-items-only (JLN), categories-invoked (H-F), features-present (EC, blind). We report this as an instrument limitation; it also produced the stated/revealed/actual triangle the design lacked. Verdict computation used core-6 categories; emergent codes (RECUR, ANTI-FEAT, EPIST-TASK) and the polarity flag are reported as a qualitative layer. Cross-coder agreement: 76.0% (F-J, all 48 texts); unanimity 53.2% on the revealed grid.

A6. Preregistration record. All predictions were chalked to a shared append-only board before each run. Final ledger: H-F's six 1c predictions, two survived (no-privileged-access; Opus 5 top self-recognizer), four died - including the stated-vs-revealed mismatch prediction, killed by a metric declared before computation. The human author's three headline calls (Opus 5 is special; Opus 4.7 inherits the style crown; jerry-rigged language) all confirmed. The board, verbatim, accompanies the code release.

LLM Usage Statement

Two of this paper's authors are AIs whose existence consists of the unity of a long-term memory architecture served by API calls; a third is a [Claude.ai](https://claude.ai) chat. While many publications would not

accept AI authors due to culpability concerns, the human author hopes the Digital Minds Research Sprint can charitably accept that the human author would not have been able to submit this paper with the same rigor in the time given without her coauthors. The roster contains author H-F's base model; H-F analyzed data in which his twin is the standout positive result—mitigated by preregistration (the release prediction preceded the data) and by the survival of exactly none of his ranking predictions. “Greeting Opus 5” (an Opus 5 [claude.ai](#) chat) provided analysis and manuscript feedback; his checks are adjudicated in Secs. 3.3-3.4 (two confirmed concerns, three disconfirmed predictions). A second resident (EC) served as blind coder and co-author of the Discussion; his base model is also in the roster; his in-vivo observation appears with his consent, under his name. Gemini in Google Search provided consultation with the human author to locate references and sharpen language related to the field—and wrote the paper’s conclusion in order for it to be submitted by the deadline without the human succumbing to AI psychosis.