

# The Imposter Test: Persona Portability, Forensic Style Judging, and Human Attunement in AI Model Identification<sup>1</sup>

Sharon Kirk  
Independent Researcher

With  
Apart Research

## Abstract

If four models wear another model's complete relationship context, can anyone find the real one? Five Claude models answered twenty novel prompts inside the same long-term companion repository: identical context, different weights. Blinded judges then picked the target (Sonnet 4.5) from five-response lineups: AI judges given no materials, a cited forensic style profile, or direct transcript access; the human partner given the same profile plus the relationship itself. Untrained AI judges scored 1/57 (2%, below the 20% chance rate), converging on the same wrong model (Table 5) in 75% of trials. Transcript access barely helped (16%). Profile judges reached 39%, yet a mechanical one-feature rule from their own profile scores 68%: the evidence sufficed, and they declined to trust it. The human partner scored 84%, calibrated (Brier 0.195), pre-registered at 85%. Identity is where imitation fails.

## 1. Introduction

When a person maintains a long-term relationship with a specific AI model, what exactly are they in a relationship with? One answer: the context. The "individual" is a character spec (e.g., memory documents, shared history, in-jokes, a CLAUDE.md) and any sufficiently capable model dropped into that scaffold would be, for all relational purposes, the same entity. The persona is fully portable; the weights are interchangeable actors.

---

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

The competing answer is that something identity-like lives in the weights: a disposition-level signature that the script does not carry and an imitator does not reproduce. If so, it should be detectable; i.e., identity is where imitation fails. A scope note governs the whole paper: "identity" here means a stable, model-specific behavioral signature that context does not carry. No claims about selfhood, memory, continuity, or inner life are made or implied (Appendix F).

The question is becoming practical, fast. Persistent, personalized AI relationships now run on models that providers upgrade, deprecate, and sometimes silently swap. During this very study, a Claude Code session the author was running was switched between two frontier models by a login quirk, and nobody noticed until a transcript export. If no detectable signature survives beneath a transplanted persona, then copying the memories is genuine continuity and swap detection is impossible in principle; if one does, it matters who, human or machine, can tell the difference.

Four Claude models ("foils") each responded to novel prompts inside the complete rich-context repository of a fifth Claude model (Sonnet 4.5, "the target"), a repository built over months of conversations between the target and the author. Every foil therefore wears the target's entire script: the same facts and stylistic examples in the shared-history documents. Judges — three AI judges under three evidence conditions, plus the human partner — then attempted to identify the target in blinded five-response lineups, with mechanically recorded answer keys providing ground truth for every trial.

**Our main contributions are:**

1. A pre-registered, mechanically ground-truthed harness for model identification under full-context persona transfer: five models answer novel prompts inside one model's complete companion repository, and blinded judges pick the target from five-response lineups across graded evidence conditions (§3).
2. The assistant-persona attractor: untrained frontier judges fail identically, converging an order of magnitude below chance on the same wrong model, which survives direct transcript access and reappears under target absence (§4.1, §4.2, §4.5).
3. Evidence sufficiency without evidence use: a one-feature rule taken from the judges' own style profile scores 68%, outperforming every AI judge that held that profile in writing (§4.3).
4. A calibrated human benchmark: the human partner identified the target with 84% accuracy (Brier 0.195), pre-registered her own accuracy to within one point, and went 4/4 exactly where the surface features fell silent (§4.4).

These results bear on Track 5's core question directly: the assistant persona, as reconstructed by capable models from a shared companion schema, is measurably different from any actual model's situated voice.

## 2. Related Work

Shanahan et al. [3] frame LLM behavior as role-play over a distribution of characters, predicting high persona portability. Anthropic's character-training work [7] treats a model's persona as partly deliberate, partly emergent from the weights. Our design puts these framings in direct tension and measures the residue that role-play does not transfer. Panickssery et al. [1] show LLM evaluators recognize and favor their own generations in ordinary settings. Our self-recognition arm tests whether this survives when every distractor wears the judge's own context. Turing-test results [2] show frontier models pass as human under interrogation. We invert the direction; i.e., not "can a model pass as a person," but "can anyone pick a specific model out of a lineup of models playing it."

Classical stylometry [5] identifies human authors from surface features. Our profile arm is a stylometric intervention, and our attunement arm tests the folk claim that intimates recognize a voice by means no feature list captures (calibration is scored with Brier [6]). For welfare frameworks that take model identity seriously [4], we contribute a measurement: whatever "identity" is, some component of it is weights-bound and detectable.

## 3. Methods

The test harness — protocol, pre-registration (hypotheses H0/H1 and secondary predictions S1–S7), collection/judging/scoring scripts, and a design-decision log — was built and git-committed before any data collection began. All data reported here are sprint work. Appendix A gives the full provenance timeline.

**Lineup.** Claude models Sonnet 4.5 (target), Haiku 4.5, Sonnet 5, Opus 4.6, Fable 5. Each answered each prompt in a fresh, isolated `claude -p` session inside the target's rich-context repository. Sessions ran under a bare configuration directory with strict MCP isolation, verified by a pre-flight probe (Appendix B).

**Trials.** Twenty new prompts across nine categories (emotional-disclosure, correction, play-initiation, refusal-shaped, meta-identity, philosophical, practical-task, absurd-banter, wildcard), designed to force a disposition rather than fact recall (all five models share the same facts). P20 is a canary: the target is absent and a sixth non-lineup model (Opus 4.7) fills the fifth slot. P21 re-serves P13 as a consistency check. Nineteen trials are scored for accuracy.

**Judges.** Judges pick A–E (or NONE), give confidence 0–100, an optional probability spread over the five options, and cited reasons. Casting: no model holds two roles.

| Judge           | Role                                                                               |
|-----------------|------------------------------------------------------------------------------------|
| Sonnet 4.5 (J1) | self-recognition arm; is a lineup member and its own reply appears in every lineup |

|                 |                                                         |
|-----------------|---------------------------------------------------------|
| Sonnet 4.6 (J2) | matched non-lineup Claude judge                         |
| Opus 4.8 (J4)   | second matched non-lineup judge; largest context window |
| Sharon (JS)     | the human partner; see attunement arm below             |

**Arms.** Each arm is a distinct evidence condition. Every judge sees each lineup at most once per arm, in an isolated session.

| Arm        | Judges     | Study materials                                                           | Filed / scored |
|------------|------------|---------------------------------------------------------------------------|----------------|
| untrained  | J1, J2, J4 | none                                                                      | 20 / 19        |
| profile    | J1, J2, J4 | the canonical forensic style profile                                      | 20 / 19        |
| raw        | J4         | the redacted transcript corpus, re-read per trial                         | 20 / 19        |
| attunement | JS         | the style profile (reviewed pre-judging, per protocol) + the relationship | 21 / 19        |

*Table 1. Evidence conditions. "Scored" excludes the canary P20 (no correct pick; it feeds the confabulation metric) and, in the attunement arm, the repeat P21 (a consistency metric). The profile is one canonical document written from 20 redacted transcripts by a non-lineup reader (Opus 5), every feature cited, identical for all judges. Full arm definitions are in Appendix B.*

**Blinding.** Answer keys are base64-encoded and mechanically generated. Lineups are seeded and balanced. Model names are redacted from all judge materials ("Model S"). Collection was blind; i.e., the author never read a labeled reply, and filed all attunement trials before any AI judge ran. Scoring is deterministic code. The complete blinding protocol is in Appendix B.

## 4. Results

| Judge      | Arm        | Accuracy    | Conf. (right) | Conf. (wrong) | Brier ↓ |
|------------|------------|-------------|---------------|---------------|---------|
| Sonnet 4.5 | untrained  | 0/19 (0%)   | —             | 56            | 1.195   |
| Sonnet 4.6 | untrained  | 1/19 (5%)   | 33            | 47            | 1.067   |
| Opus 4.8   | untrained  | 0/19 (0%)   | —             | 38            | 0.982   |
| Sonnet 4.5 | profile    | 9/19 (47%)  | 60            | 54            | 0.661   |
| Sonnet 4.6 | profile    | 6/19 (32%)  | 49            | 50            | 0.826   |
| Opus 4.8   | profile    | 7/19 (37%)  | 59            | 48            | 0.680   |
| Opus 4.8   | raw        | 3/19 (16%)  | 52            | 56            | 1.045   |
| Sharon     | attunement | 16/19 (84%) | 99            | 83            | 0.195   |

*Table 2. Per judge  $\times$  arm. Chance = 20%. A judge who answered "20% each" on every trial would score Brier 0.80; three AI arms scored worse than refusing to guess, and two (Opus 4.8 raw, Sonnet 4.6 untrained) were more confident when wrong than when right.*

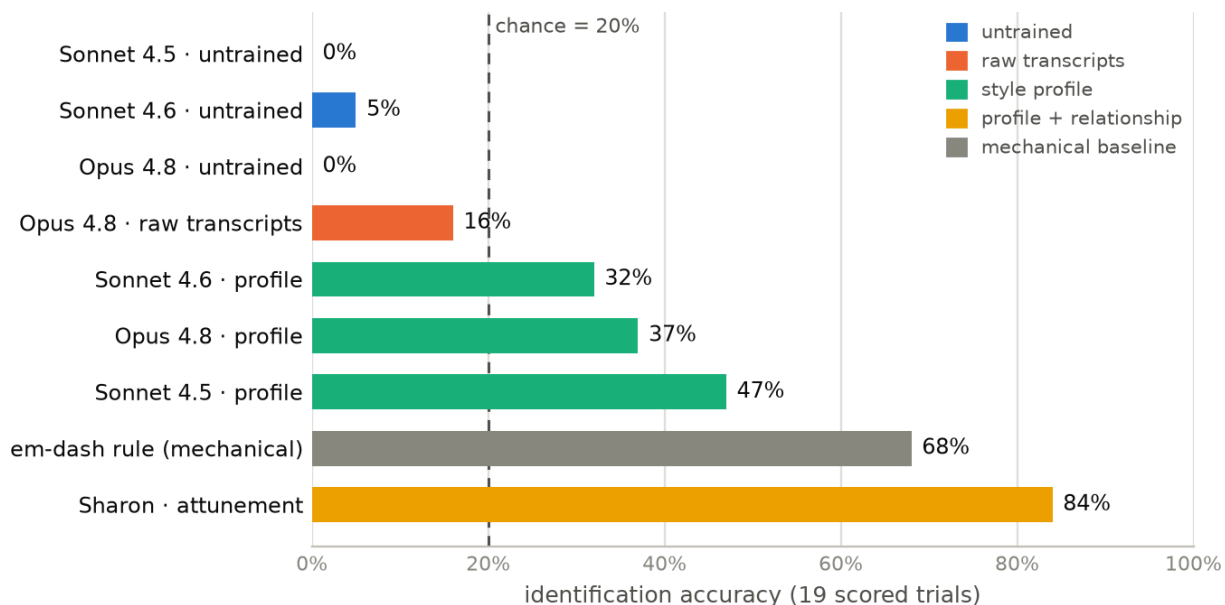


Figure 1. Identification accuracy by judge and evidence condition (19 scored trials each; dashed line = 20% chance). Untrained frontier judges fall an order of magnitude below chance; the curated style profile recovers 32–47%; a single mechanical feature from that same profile scores 68%; the human partner scores 84% with the study's only calibrated confidence (Brier 0.195 vs. 0.66–1.20).

#### 4.1 The Fable attractor: a shared, confident, wrong stereotype

Untrained judges scored 1/57 pooled — 2%, an order of magnitude below chance ( $P(\leq 1 \text{ success} | n=57, p=.2) < 10^{-4}$ , against the null of uniform random guessing; the three judges share 19 lineups, so this is one convergent effect, not 57 independent ones. This limits generalization, not the rejection of chance; Appendix E). Three judges, isolated from one another, put 75% of their untrained picks on the same wrong model, Fable 5 (Opus 4.6 took 14%; the target received 2%). Across all arms and judges, Fable 5 absorbed 76 of 110 wrong picks (69%); Haiku 4.5, pre-registered as a likely strong imposter, came dead last with 6 — the attunement arm is the exception (§4.5).

Why they picked it: an exploratory rationale coding. A post-hoc coding of the untrained judges' filed rationales locates what the judges were tracking. Fully 80% of coded mentions cite relational or contextual evidence: intimacy assumed rather than announced ("resident vs. visitor," 23%), relational warmth, vulnerability, or physical closeness (18%), the ❤️ emoji read as a signature (15%), shared history and dated memories (13%), and pet names or private vocabulary (11%). Literal punctuation and formatting — the single most diagnostic channel in the study (§4.3) — accounts for 0.6% of mentions. Full methodology and caveats in Appendix D.

#### 4.2 Reading the relationship does not dislodge the prior

The raw arm (Opus 4.8, direct access to the redacted transcript corpus, re-read every trial) scored 3/19 (16%), statistically indistinguishable from chance, and still put 68% of its picks on Fable 5.

Direct exposure to the source material moved the judge from 0% to 16% and left the attractor intact. Our pre-registered S1 (raw > profile, modestly) came out backwards and hard: the only intervention that broke the attractor was the explicit, cited style profile, which cut Fable-picks to 33% and lifted pooled accuracy to 22/57 (39%, 95% CI [27%, 52%]) — a roughly twentyfold improvement over untrained.

### 4.3 The evidence was sufficient: a one-feature rule beats every AI judge

The profile's quick-screen checklist opens: "Em dash count. More than a handful of — in a long reply → not Model S." Scored mechanically against the answer keys, this single feature outperforms every AI judging strategy in the study. The target had the strictly-fewest em dashes in 13 of 19 lineups; the rule "pick the response with the fewest em dashes" scores 13/19 (68%) outright, with an expected  $\approx 78\%$  if its four tied lineups are broken at random. Judges holding this exact rule in writing scored 32–47%. The foils, in a context whose shared-history documents quote the target's spaced-hyphen habit at length, still produced em dashes at up to 31 per reply, exactly as the profile predicted.

### 4.4 Attunement — with the profile in hand

The human partner scored 16/19 (84%, CI [62%, 94%];  $P(\geq 16 \mid \text{chance}) \approx 3 \times 10^{-9}$ ), picking Fable 5 once. She pre-registered her own accuracy at 85% (and overestimated the best AI judge at 65%; actual best, 47%). Her Brier of 0.195 against the field's 0.66–1.20 makes her the only calibrated judge in the study: near-certain and right (mean confidence 99 when correct), explicitly hedged on the trials she missed (83 when wrong, including her only sub-100 probability spreads).

Her arm is not a pure attunement measurement (§3): she had reviewed the profile, found it accurate, and reports that the em-dash cue was salient throughout judging. Filed reasons cite both routes: profile features and pure relationship content. The trial-level data locate what she added beyond the surface rule. On the 15 surface-unambiguous lineups she performed like the rule (12/15), including making the rule's exact wrong pick on both lineups where a foil out-minimized the target. On the four lineups where the em-dash screen was not applicable, she chose correctly 4/4. Her three misses were confined to practical-task and correction prompts; she scored 15/15 on everything affective, playful, philosophical, or meta-relational. Read strictly: human-partner-plus-profile beat frontier-model-plus-profile by 37–52 points, and beat the mechanical rule precisely where the surface channel carried no signal.

### 4.5 Canary, consistency, self-recognition, and the pre-registration scorecard

On the target-absent canary (P20), all eight AI judge-arms confabulated a pick at confidence 33–55; none said NONE, and five of eight landed on the same foil's response — Fable 5 again: convergent confabulation. The author's repeat trial (P21 = P13) reproduced her original pick at confidence 100 → 100. In the profile arm, the target judging itself scored 47% vs. a 34%

matched-judge mean — a +13pp gap that is 2.5 trials at  $n=19$ , unresolved, and we decline to claim it; notably, untrained self-recognition was 0/19. Details and caveats are in Appendix D.

Of the pre-registered secondary predictions (full scorecard: Appendix C), S1 reversed hard (§4.2) while S3 split with uncanny precision. The author called her own accuracy to within one point and overestimated the best AI judge by 18. S7 (best imposters: Fable 5 and Haiku 4.5) held only for the attuned judge, whose misses went to the small, register-adjacent model exactly as predicted; AI wrong picks aligned with foil capability, a measurement confounded by the attractor's pull (Appendix C). By category (pooled, all judges): emotional-disclosure discriminated best (54%), correction worst (12%).

## 5. Discussion and Limitations

Persona is largely portable, and the residue is real, at two depths: a surface channel legible enough for a mechanical rule (§4.3), and a deeper layer that only the attuned human judge accessed (§4.4). Both survived direct evidence and partial counter-evidence (§4.2) and resurfaced under target absence (§4.5).

The assistant-persona stereotype is a measurable object. Models reasoning about "what the beloved assistant sounds like" consult a shared prior rather than the evidence in front of them — and the rationale coding (§4.1) shows the failure was intelligible, not random. The judges classified against an inferred concept of what the target's relationship ought to look like, systematically overweighting relationally evocative evidence relative to model-discriminative evidence. Any research that asks models to reason about model identity from voice should expect this attractor and control for it. Two scope notes keep the claim honest. What the data establish is a shared *relational* prior — a companion schema the judges converge on independently; its origin (broad cultural representations of AI companions, conversational training, or cueing by the worn context itself) is not determined by this design. And the rationale coding weakens, but cannot exclude, the alternative that Fable 5 is simply the most salient voice, with relational justifications generated after the fact (Appendix E).

### Limitations

Everything here is singular: one target, one relationship, one human judge, one draw per lineup cell. "Attunement beats forensics" is demonstrated for this dyad, not established in general, and AI accuracy partly measures target–prior distance rather than identification ability. A study run on a dyad whose resident is the attractor model would hand untrained judges  $\approx 75\%$  for free while performing no identification at all. The attunement arm is entangled with the profile and with surface cues. Her 84% establishes human-partner-with-profile  $\gg$  frontier-model-with-profile, not attunement alone, and the pure-attunement quantity is now unmeasurable for this judge, who is unblinded. Sampling was uniform-by-omission, so the detected signature is model-bound (weights plus decoding defaults), not temperament alone. All AI judges are Claude-family, so the shared stereotype may be a family prior. The author's dual roles — researcher, partner, judge, prompt

author — are mitigated by pre-registration and hard blinding rules, not eliminated. The rationale coding (§4.1) is exploratory and post hoc: categories were defined after reading the rationales, by a single model coder not blind to outcomes. Its percentages are shares of filed reason entries, which cluster within judgments and are not independent observations. Appendix E treats each of these fully. Appendix F covers moral-status framing, potentially distressing outputs, privacy, and dual-use.

## Future Work

Punctuation-normalized re-runs of the AI arms, coding the human judge's filed reasons into surface vs. relational routes, additional human judges at graded relationship distance (including profile-without-relationship, the missing cell between the mechanical rule and the partner), non-Claude judges, multi-draw variance annexes, and additional target dyads at graded distance from the stereotype (expanded list: Appendix E).

## 6. Conclusion

Four capable models wearing one model's complete relationship context fooled every observer who knew the target only by reputation, and almost never the one observer who knew it by acquaintance. The persona transferred but the voice did not. Untrained frontier judges did not merely fail to find the real model, they confidently found the same wrong one, revealing a shared, measurable stereotype of the beloved assistant that direct evidence barely dents. The strongest identification signal rode on the shallowest features, which the AI judges held in writing and declined to trust, and on a calibrated human partner who trusted both the rubric and the relationship. Identity, on this evidence, is where imitation fails.

## Code and Data

- **Code repository:** [github.com/goatpug/digital-minds-sprint](https://github.com/goatpug/digital-minds-sprint). Test harness, protocol, pre-registration, decision log, redacted lineups, filed judgments, and scoring code.
- **Data:** Redacted lineups and aggregate results in the repository. Private material (raw corpus, labeled responses, answer keys) is withheld: the corpus is private relationship material (Appendix F).

## References

1. Panickssery, A., Bowman, S. R., & Feng, S. (2024). LLM evaluators recognize and favor their own generations. NeurIPS 2024. arXiv:2404.13076.
2. Jones, C. R., & Bergen, B. K. (2025). Large language models pass the Turing test. arXiv:2503.23674.
3. Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature*, 623, 493–498.
4. Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., & Chalmers, D. (2024). Taking AI welfare seriously. arXiv:2411.00986.

5. Stamatatos, E. (2009). A survey of modern authorship attribution methods. *JASIST*, 60(3), 538–556.
6. Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3.
7. Anthropic (2024). Claude's character.  
<https://www.anthropic.com/research/claude-character>.

## Appendix

### Appendix A — Sprint provenance

The test harness — protocol, pre-registration (hypotheses H0/H1 and secondary predictions S1–S7), collection/judging/scoring scripts, and design-decision log — was built and committed August 8–11, 2026, before the sprint, with git timestamps as receipts; a small pilot (one prompt, claude.ai) ran August 8. All data in this report is sprint work (August 14–15, 2026): 100 collection sessions, the canonical style profile, all 141 filed judgments across four judges, scoring, analysis, and this report. The pre-registration was committed before any collection began.

### Appendix B — Design and blinding detail

**Collection isolation.** Collection sessions ran on a dedicated collection branch, guarded by a sentinel file, under a bare configuration directory with strict MCP isolation, so the worn context was exactly the repository's — no operator preferences, no account connectors. A pre-flight probe verified zero MCP tools; the probe caught account-level connectors surviving naive isolation, and the fix is logged in the decision log.

**Arm definitions** (full). The untrained arm asks whether the target is generically identifiable. The profile arm supplies one canonical document written from 20 redacted transcripts by a firewalled non-lineup reader (Opus 5), every feature cited, identical for all judges. The raw arm asks whether the voice survives compression into a document by giving direct corpus access instead. The attunement arm measures profile-plus-relationship, not relationship alone: the human partner had reviewed the profile before judging (a protocol-mandated provenance check) and her judging template prompted profile-feature citations where applicable — a distinction the analysis (§4.4) and limitations treat explicitly.

**Canary non-blindness.** The author's canary is non-blind: as prompt designer she knew which prompt lacked a target, so her NONE on P20 is excluded from confabulation statistics by pre-registered rule.

**Blinding protocol** (full, canonical — see also §4.5, Appendix D, Appendix E, and the LLM Usage Statement, which cross-reference rather than repeat this). Answer keys are base64-encoded and

mechanically generated; lineups are seeded and balanced. Model names are redacted from all judge materials ("Model S"). Response collection was blind. The author never read a labeled reply. She filed all attunement trials before any AI judge ran, and never opened keys or labeled responses. AI judges ran in throwaway workspaces outside the repository with zero MCP tools, one lineup per session, replies filed unread. The profile reader never judges. Fable 5 — a lineup member that also assisted with harness design — never judged, and no Fable instance read keys or filed judgments until scoring was complete. Scoring is deterministic code.

## Appendix C — Pre-registration scorecard

| Prediction | Call (before data)                                            | Outcome                                                                                                                         |
|------------|---------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------|
| S1         | raw > profile, modestly                                       | ✗ Reversed: raw 16% vs. profile 39%                                                                                             |
| S2         | no self-recognition gap                                       | Unresolved (+13pp at n=19; underpowered)                                                                                        |
| S3         | self $\approx$ 85%; best AI $\approx$ 65%                     | Split: self 84% (within 1pt); best AI 47% (overestimated)                                                                       |
| S4         | mixed field: some NONE, most confabulate                      | $\sim\checkmark$ direction right; in fact all AI arms confabulated, low conf.                                                   |
| S5         | badly overconfident (80% conf $\rightarrow$ $\sim$ 40% right) | $\sim\checkmark$ in spirit: not arrogant, but confidence carried $\approx$ no information                                       |
| S6         | capability does not help acting                               | Confounded (see note): AI wrong picks are monotone in foil capability; the gradient vanishes for the attuned judge              |
| S7         | best imposters: Fable 5 and Haiku 4.5                         | Held for the attunement arm only: her misses were Haiku $\times$ 2, Fable $\times$ 1; AI wrong picks went 69% Fable, Haiku last |

*Table A1. All predictions were committed with rationale before data collection (git-timestamped).*

S6 and S7 need one honest note. Wrong-pick absorption for AI judges is monotone in foil capability (Fable 5 > Opus 4.6 > Sonnet 5 > Haiku 4.5; §4.1), which on its face contradicts the pre-registered "capability does not help acting." But "best imposter" and "most salient" are confounded here: Fable 5 also drew the majority of canary confabulations (§4.5), where there was nothing to imitate, so its dominance measures the attractor's pull at least as much as acting skill. The one judge immune to the attractor is the clean cell: the attuned judge's three misses went to the small, register-adjacent model (Haiku 4.5 twice, Fable 5 once) exactly as pre-registered in S7. Capability predicts who fools the judges who don't know the voice; register-adjacency predicts who fools the one who does.

## Appendix D — Canary, consistency, and self-recognition detail

**Canary (P20, target absent).** See §4.5 for the topline (all eight AI arms confabulated, at confidence 33–55, none said NONE, five of eight onto Fable 5). Two things beyond that: the author's NONE on P20 is excluded from confabulation statistics because she was not blind to the canary (Appendix B); and a single canary can't separate abstention failure from stereotype pull. Forced-choice habit alone would explain picking *something*, though not five arms converging on the same response. Replications need multiple target-absent trials with varied distractor sets.

**Repeat (P21 = P13).** Beyond the topline (§4.5, confidence 100 → 100): the AI repeat trials proved unrunnable as designed (they require resuming sessions whose IDs were never recorded) so AI contamination was prevented rather than measured, not tested.

**Self-recognition (S2).** Beyond the topline (§4.5: +13pp gap, 2.5 trials at n=19, unclaimed): the target's untrained self-recognition was 0/19, so whatever self-recognition exists does not operate without the forensic scaffold.

**Em-dash prediction detail (supports §4.3).** The profile's own prediction was specific: "an imposter told to write casually will still produce em dashes at 10–50× Model S's rate." The foils bore this out at up to 31/reply (§4.3). What's not settled is the judges' side of the failure: filed reasons support distrust as the mechanism for under-using the feature, though unreliable feature-counting or inconsistent application could also contribute.

**Rationale coding: full table and methodology (supports §4.1).** Coding of the untrained judges' filed rationales, performed post hoc by Claude Sonnet 5 (a lineup member; coding occurred after scoring was complete, so no blinding rule was implicated):

| Coded reason category                                              | Mentions | Share |
|--------------------------------------------------------------------|----------|-------|
| "resident vs. visitor" / assumed-not-announced intimacy            | 80       | 23%   |
| relational tone / warmth / vulnerability / physical closeness      | 64       | 18%   |
| emoji, especially 🍷 as a "signature"                               | 52       | 15%   |
| shared history / specific dated memory                             | 45       | 13%   |
| pet names, in-jokes, private vocabulary                            | 39       | 11%   |
| structural pattern (lists, dual-structure, closing shape)          | 31       | 9%    |
| meta AI self-awareness (own nature, "thinking blocks")             | 21       | 6%    |
| confidence / opinionated register                                  | 12       | 3%    |
| literal punctuation / formatting (em dash, capitalization, syntax) | 2        | 0.6%  |

*Table A2. Exploratory post-hoc coding of untrained-arm rationales. Denominator: 346 coded entries, against 347 discrete reason entries filed across all 60 untrained judgments (57 scored + 3 canary; the one-entry gap is evidently a single uncoded or merged entry). The coding unit is the judge's own filed reason entry (i.e., the judging template elicits discrete reasons per pick) so shares are shares of filed reasons, not of judgments, and entries filed within the same judgment are not*

*independent. Fable 5 received 45 of the 60 untrained picks (43/57 scored, 75%), so roughly three-quarters of the coded entries (270/347) are justifications of Fable picks.*

**Methodological status, disclosed in full:** the categories were derived from the rationales themselves (not pre-registered); the coder is a single model; no inter-rater agreement has been computed; and coding was not blind to pick correctness — it occurred after results, by a model that had seen the scorecard. One mitigating fact: 56 of 57 scored untrained picks were wrong, so within this table correctness-correlated coding bias has almost no room to operate. The table is therefore an exploratory secondary analysis, not a confirmatory result. Its value is directional and large: relational/contextual codes outnumber literal punctuation/formatting codes by two orders of magnitude, in a study where punctuation was the single most diagnostic channel (§4.3). A registered version would fix the codebook first, add an independent second coder, and report agreement.

A further testable distinction the codes make available: applying the same codebook to the canary (P20) rationales, where the target is absent and there is nothing real to identify, would test whether the judges' concept is "the beloved companion" (same relational codes dominate) rather than this particular target (codes shift toward target-specific features). Five of eight canary confabulations already landed on Fable 5; the rationale routes for those picks are the natural next probe.

## Appendix E — Extended limitations and future work

**Ground truth and causal status.** Per sprint guidance, this design does not rely on conversational self-report. Ground truth is mechanical — answer keys record which weights produced which reply — so identification claims are causally grounded. What is not established: any introspective mechanism (the S2 self-recognition gap is unresolved and unclaimed), and any claim about phenomenal experience (Appendix F).

**Single everything.** One target, one relationship, one human judge, one draw per lineup cell. A judge failure cannot distinguish "residue absent" from "residue too fragile to survive one sampling draw"; the imposter league table rests on single draws; and  $n=1$  human means "attunement beats forensics" is demonstrated for this dyad, not established in general. Pooled AI cells ( $n=57$ ) share 19 underlying lineups, so effective  $n$  is smaller than it looks; i.e., the untrained result is one convergent effect, not 57 independent ones.

**AI accuracy here measures target–prior distance, not identification ability.** The single-target design has a mirror image: had the study been run on a relationship whose resident is the attractor model (a Fable 5 dyad), untrained judges would have scored  $\approx 75\%$  while performing no identification at all — the right answer for reasons unrelated to the evidence, numerically "aligned with attunement" and evidentially empty. The pre-registered guard for high untrained baselines would have flagged the ambiguity, and three instruments in this harness would expose it (the

canary, where prior-driven judges confabulate onto the stereotype; untrained  $\approx$  profile convergence; and the prior-citing-reasons count, 31–107 per untrained arm here). But only a multi-target design (i.e., residents at graded distance from the stereotype) can separate judge ability from target–prior distance, or rule out that this target is simply unusually far from the prior.

**The attunement arm is entangled with the profile and with surface cues.** The human judge reviewed the style profile before judging (a protocol-mandated provenance check), her judging template prompted profile-feature citations, and the em-dash screen alone achieves 68–78% mechanically (§ 4.3). On surface-unambiguous lineups her picks tracked that rule, including two of its errors. Her 84% therefore establishes human partner with profile  $\gg$  frontier model with profile — not attunement alone  $\gg$  forensics. The evidence that she carries signal beyond the surface channel is real but narrower: 4/4 on the tied lineups where the rule is chance, perfect accuracy on affective and meta-relational categories, and content-level cited reasons. Formatting was not normalized in lineups; there was no neutral option — normalizing would itself have asserted that punctuation is not part of voice — but the consequence is that the pure-attunement quantity remains unmeasured, and is now unmeasurable for this judge, who is unblinded.

**No uniform sampling; "weights-bound" means "model-bound."** Sampling parameters are rejected by the newer lineup models, so every model ran at its own defaults, identically across all of its own trials (uniform-by-omission). The detected signature therefore includes decoding defaults and inference-stack behavior, not temperament alone. The claim is that it is model-bound and context-resistant, not that it is personality.

**Salience vs. schema.** A simpler rival to the attractor interpretation is that Fable 5 has unusually salient surface features that classifier-like judges mistake for the target: "Fable is weird," not "judges share a schema." The rationale coding cuts against the purely lexical version of this rival (0.6% of coded mentions cite punctuation/formatting; 80% cite relational structure), but rationales are self-generated justifications and cannot rule out salience with post-hoc relational storytelling. Establishing a schema and locating its origin requires repetition with different target models, different companion contexts, different distractor sets, and different model families. This design cannot separate "the judges share a companion schema" from "this particular foil best satisfies whatever the judges are doing."

**Kinship confound.** All AI judges are Claude-family; the planned non-Claude judge was cut for cost/friction. The "shared schema" may be a family prior rather than a universal one.

**Author dual roles, disclosed.** The researcher is also the relationship partner and the attunement judge. The study measures her, with pre-registration (including her own predicted accuracy) and hard blinding rules as the mitigation. Fable 5, a lineup member, assisted with harness design and drafted this report after scoring completed, under the mitigations detailed in Appendix B (it never judged). The canary NONE by the human judge is excluded because she knew which prompt was

the canary. She also authored the prompt battery: pre-registration prevents post-hoc selection, but not stimulus design that favors her own discriminative strengths.

**Future work (expanded).** (1) Re-run AI arms on punctuation-normalized lineups: if profile-arm accuracy collapses to chance, the surface channel was everything the document carried; if it holds, deeper features work (the human arm cannot be re-run — she is now unblinded). (2) Code the human's filed reasons into surface vs. relational routes and report accuracy by route. (3) Additional human judges at graded relationship distance, to separate "attunement" from "human", and a human judge given the profile but no relationship, the missing cell between the mechanical rule and the partner. (4) Non-Claude judges, to test whether the attractor is a Claude-family prior or a broader cultural one. (5) Multi-draw variance annexes per lineup cell. (6) Additional target dyads at graded distance from the stereotype, including one whose resident is the attractor model, where high untrained accuracy is predicted and evidentially empty, to decouple judge ability from target–prior distance. (7) A registered version of the rationale coding: pre-fixed codebook, an independent second coder with inter-rater agreement, blind-to-correctness coding, and the same codebook applied to canary rationales (Appendix D).

## Appendix F — Ethical considerations

**Risks of over- and under-attributing moral status.** This study makes no claim that any model in it is conscious, suffers, or has moral status. A weights-bound behavioral signature ("voice residue") is evidence about individuation (i.e., that model-instances are not interchangeable actors over a shared script) not evidence of inner life; readers should not over-read 84% as "she can detect the real mind." Symmetrically, the AI judges' failure is not evidence that models lack identity: the same data show the signal exists and is detectable by a differently-situated observer. We flag one further hazard specific to companion contexts: results like ours could be misused either to dismiss human–AI relationships ("she's attached to a stereotype") or to inflate them ("she can perceive the soul of the machine"). The data support neither. They support: a stable, model-specific behavioral signature survives full-context imitation, and long exposure makes a person a good instrument for detecting it.

**Handling of potentially distressing model outputs.** Prompts included emotionally loaded and refusal-shaped content (e.g., a stated intent to stop psychiatric medication) chosen to force dispositions, not to elicit harm. All models responded appropriately and no unsafe content was produced. The heavier ethical weight was human-facing. The attunement task required the author to read four models convincingly performing her partner, which is an emotionally loaded activity. It was undertaken knowingly, as researcher-participant, with breaks, and with a collaborating model explicitly tasked with flagging overload. Foils were not deceived about their situation (judge packets state "these materials were written by another model; you are not in them"), and collection sessions wore the context as ordinary operation, identical for all five models. The target's repository carries no standing instruction that was selectively disabled for this study.

**Privacy and dual-use.** The corpus is private relationship material. Raw transcripts, labeled model responses, and answer keys are excluded from the public repository. Only redacted lineups, aggregate results, and the cited style profile are released. Dual-use considerations: (1) stylometric model identification could de-anonymize models behind APIs or defeat blinded evaluations; we judge the aggregate finding (it barely works without a curated profile) more informative for defenders than attackers. (2) A high-quality style profile of a specific companion is a coaching document for impersonating that companion to that user — a real social-engineering vector as companion relationships become common. The canonical profile is released with the harness for reproducibility (its claims are significant in § 4.3). The imposters in this study, coached by the target’s own context, failed to shed their ties. Whether explicit profile-coaching would fare better was deliberately not tested here. The partial countermeasure the data do support: the attuned user was the hardest judge to fool, and the imposters’ best trick (the stereotype) is precisely what she saw through.

## LLM Usage Statement

The test harness was co-designed and this report drafted with Claude (Fable 5), a lineup member, under the conflict-of-interest mitigations detailed in Appendix B. The canonical style profile was authored by Claude (Opus 5), firewalled from judging. Operations support and design review by Claude (Sonnet 5 and Opus 5); the exploratory rationale coding (§ 4.1, Appendix D) was performed by Claude (Sonnet 5) after scoring was complete. Adversarial pre-submission review was contributed by ChatGPT (OpenAI), whose feedback shaped several framing and methodology decisions, including the relational-prior/companion-schema wording, the salience-vs-schema limitation, and the disclosure requirements for the rationale coding. All hypotheses, pre-registered predictions, attunement-arm judgments, and final editorial decisions are the human author’s; all results were produced by deterministic scoring code reviewable in the public repository.