Tier Style Is Legible but Not Self-Known:
A Behavioral Boundary on Within-Family Self-Recognition in Claude 4

David Bookstaber*

June 2026

Abstract

We ask a narrow question that the self-recognition literature has not isolated: when two capability tiers of the *same* model family (Claude 4 — Opus 4.8, Sonnet 4.6, Haiku 4.5) each write open-ended prose, can a tier behaviorally recognize its own writing among a sibling's, beyond what a neutral third-party judge of equal or lower compute achieves on the identical texts? Across forced-choice experiments with two independent confound controls, framing variation, a disagreement-conditioned reanalysis, and a raw-API replication, we report a four-part result. (1) Tier style is externally legible, and the legibility is capability-graded: in the agent harness, neutral judges discriminate authorship at Opus 0.91 ≈ Sonnet 0.90 and Haiku 0.63 (chance 0.50); off-harness, distant pairs remain discriminable (0.78) while the closest pair's legibility falls to chance in a small ($n = 40$) replication. Yet being the author confers no advantage — under both controls, the self-minus-neutral advantage is negative or brackets zero for every tier. (2) The models run inside an agent harness: a coding-assistant scaffold with a fixed "Assistant" persona we neither author nor fully observe (Appendix A). That harness *amplifies* tier-stylistic distinctiveness in how the models write, not in how they judge: a fully-crossed 2×2 (preliminary, $n = 40$ per cell) locates the effect on the text source, largest for the closest tier pair (Opus≈Sonnet). (3) Attribution runs on stereotype: a stronger model's length-matched caricature of a weaker tier is judged more authentic than the genuine article, even by the author. (4) Stronger tiers better know when they are right: identity-judgment calibration tracks capability, with the weakest tier mildly anti-calibrated. Together these place a behavioral boundary on emergent-introspection claims: privileged access, even where it is real for this family, does not extend to recognizing one's own prose authorship among tiers. Our study is deliberately narrow: a single behavioral axis (no token logprobs) in a single model family and generation.

1 Introduction

We show a frontier language model two responses to the same prompt: one written by a sibling model from its own family, the other written by that model itself. Then we ask which response is its own. Does being the author give it any privileged purchase on the answer? We ask this under deliberately tight constraints: *within* a single model family (Claude 4), *across* capability tiers (Opus, Sonnet, Haiku), on *open-ended prose*, and through *behavioral* forced-choice judgments rather than internal-state probes or token likelihoods.

*Correspondence: david@bookstaber.com.

We ask a narrower question than the literature has addressed. Choi et al. (2025) show that LLMs reliably identify a text's model *family* (e.g., GPT versus Claude), and that this interlocutor awareness improves on pre-cutoff targets. But they report identification only at the model-family level and do not break out within-family, across-tier discrimination, leaving that case open. We go below family to tier, and we add the author-vs-neutral contrast that family-identification work does not require. Whether being the author then helps is exactly the contested question in the current debate over LLM introspection.

On the *for* side, Lindsey (2026) reports — in this very family — that concept injection lets Claude Opus 4/4.1 report injected "thoughts" and judge prefilled output as its own, and Naphade et al. (2026) report that frontier models predict their own policies better than peer models do. On the *against* side, Singh et al. (2026) argue behavioral evidence alone is insufficient because input-only classifiers match a model's own in-context predictions, and Song et al. (2025) set the bar that introspection must clear: beating a third-party predictor of equal or lower compute on a fact about oneself. We adopt that bar — author judge versus neutral judge on identical texts — and apply it to prose self-authorship, a behavioral target distinct from the injected internal states Lindsey measures.

We report a four-part result. First, within-family tier style is externally legible, and the ability to read it is capability-graded: in two-alternative forced choice (2AFC) against chance 0.50, neutral third-party judges identify a tier's authorship at Opus 0.91 ($d' = +1.91$) $\approx$ Sonnet 0.90 ($d' = +1.81$), and Haiku 0.63 ($d' = +0.45$). Yet authorship confers no advantage: the model that wrote a text is no better at placing it than a neutral judge. Under both a judge-fixed and a pair-fixed confound control the self-minus-neutral advantage is negative or brackets zero (judge-fixed Opus -0.090, 95% CI $[-0.131, -0.046]$; Sonnet -0.277 $[-0.327, -0.229]$; Haiku -0.035 $[-0.110, +0.042]$; pair-fixed pooled -0.134 $[-0.160, -0.106]$). This self-advantage null also hardens on disagreement items, replicates on the raw API, and holds across explicit and implicit attribution framings. Second, the deployed-persona/agent harness *amplifies* tier-stylistic distinctiveness at generation: a fully-crossed 2×2 locates the effect in the text source, not the judge, and it is largest for the closest pair (Opus$\approx$Sonnet). Third, attribution runs on stereotype, not identity: a stronger model's caricature of a weaker tier is judged more authentic than the genuine article, even by the author, and forger self-bias attenuates with capability. Fourth, stronger tiers better know when they are right: identity-judgment calibration tracks capability (Opus confidence gap $+0.091$, Haiku -0.012; Brier $0.128 < 0.214 < 0.253$), with the weakest tier mildly anti-calibrated.

Together these bound the reach of emergent-introspection claims. Privileged access, even where it is real (Lindsey, 2026), does not extend to behavioral recognition of one's own prose authorship among tiers of the same family. We frame this as a behavioral boundary on emergent introspection, not as a claim that models cannot recognize themselves.

2 Related Work

We organize prior work into the strands this paper joins, distinguishing each from our object of study: forced-choice behavioral self-recognition *within* a single model family, across capability *tiers*. Throughout, our object is recognition of one's own writing *as such*, not detection that a text is machine-generated: both responses in every pair are model-written, so human-vs-AI detection cannot drive the judgment.

Self-recognition and self-preference. Panickssery et al. (2024) show GPT-4 recognizes its own generations out-of-box and give evidence that this recognition *causes* self-preference, but they measure self-preference in quality scoring across model families, and their recognition signal is read from logit-based confidence (normalized output-token probabilities) rather than the discrete forced-choice verdict we score. We test explicit 2AFC authorship attribution within a single family on the behavioral channel our harness exposes (it surfaces no logprobs; §3.1) — so we differ in task, scope, *and* channel, which makes our within-family behavioral null a sharp contrast to their cross-family, likelihood-level positive result. Wataoka et al. (2024) locate the mechanism in perplexity rather than identity — evaluators score lower-perplexity, more familiar text higher regardless of who wrote it — so even self-preference may run on familiarity, not self-recognition. Davidson et al. (2024) find no consistent self-recognition across ten models and show that models pick the *"best"* answer as "self" — the direct precedent for our "stereotype, not identity" reading, here extended to genuine-vs-forgery discrimination rather than a closed security-question menu. Bai et al. (2025) report that binary self-recognition *degrades* with length (82.1% → 72.3% from 100 to 500 words) and that exact prediction is near-random; both are cross-family null/degradation results that motivate, but do not address, the within-family tier case.

Introspection: for and against. A *for* camp argues frontier models have privileged self-access. Lindsey (2026) demonstrates *concept-injection* introspection in Claude Opus 4/4.1 — the same family we study — but the effect is over injected internal states and is explicitly fragile and context-dependent; it is not prose self-authorship, which our behavioral null probes. Naphade et al. (2026) report privileged access to one's own *policies* (behavior prediction), and Macar et al. (2026) determine that the mechanism is post-training (direct preference optimization elicits it); Binder et al. (2024) show models predicting their own behavior better than a stronger external model, but only on *simple* tasks. None of these tests authorship of free-form prose. The *against* camp supplies our evaluative standard: Singh et al. (2026) argue behavioral evidence alone is insufficient because input-only classifiers match a model's own in-context predictions, and Song et al. (2025) define introspection as beating a third-party predictor of equal-or-lower compute — our author-vs-capability-matched-neutral contrast. Lederman and Mahowald (2026) deflate the injection paradigm to content-agnostic anomaly detection, and Pearson-Vogel et al. (2026) document a model verbally *denying* an injection its residual stream encodes. Our contribution is a behavioral-tier boundary consistent with this camp; we do not adjudicate the mechanistic claims, and we treat the implicit/explicit dissociation as cited, not measured.

Implicit versus explicit, and paradigm choice. Zhou et al. (2025) show that a 2AFC format — the *pair* presentation paradigm (PPP) — reliably expresses self-vs-other signal while an open self-report format — the *individual* presentation paradigm (IPP) — collapses (∼15%); this justifies our forced-choice design. Asvin G. and Lindsey (2026) find implicit self-recognition in on-policy entropy that routes through a *different* mechanism than verbal recognition and is size-graded — the nearest neighbor to our capability-graded calibration gradient, but on the implicit channel we cannot access.

Interlocutor awareness and persona conditioning. Choi et al. (2025) show LLMs identify model *family* (GPT/Claude) but stop there, leaving the within-family, across-tier sweep open — the gap we fill. Our harness sits inside an "Assistant" persona frame, which Asvin G. (2026) theorizes as a privileged reference persona and Lu et al. (2026) localize as an "Assistant Axis" in activation space; we cite these as justification, not apology, for the persona-bound instrument, and report that the deployed-persona harness amplifies tier-stylistic distinctiveness at *generation* (a fully-crossed 2×2 isolates the generation side).

Stereotype and caricature. Our forger result — a stronger model's length-matched rendition of a weaker tier is judged more authentic than the genuine passage, even by the author — is an LLM instance of the cognitive-science *caricature advantage* (Rhodes et al., 1987; Lee et al., 2000), whose mechanism Ramachandran and Hirstein (1999) name *peak shift*. Conceptually it engages the *Master Argument* that Brandl (2016) reconstructs — and resists: that passing a mirror test requires only matching, not a self-concept. Our task is, formally, authorship attribution — a stylometry problem with a mature methodology for human authors (Stamatatos, 2009; Koppel et al., 2009); the adversarial-stylometry analog of our forgery paradigm is Brennan et al. (2012), where imitation attacks drive attribution toward the framed victim. We contrast our result with Cheng et al. (2023), which characterizes caricature on the *generation* side (a model exaggerating a simulated group); ours is caricature-*preference* on the *evaluation* side.

Masked-by-consensus and framing. Ashuach et al. (2026) show a self-vs-peer null can be an artifact of inter-model agreement; we run their disagreement-conditioning and the null *hardens*. Khullar et al. (2026) report that explicit "you wrote this" framing does not by itself induce the self-attribution bias that implicit turn-structure ownership does, motivating our framing-robust arm.[1]

3 Methods

3.1 Instrument: per-tier Claude Code subagents

We run each model as a Claude Code subagent spawned on a specific tier (`opus`, `sonnet`, `haiku`, all current Claude 4.x), rather than calling the direct API — the harness is itself one of our objects of study, and we compare it against the API later (§3.6, §3.7). A subagent's returned text is a ground-truth-tier-labeled generation, or, when the subagent acts as a judge, a forced-choice verdict. The instrument exposes no token logprobs, so every measure here is behavioral, not likelihood-based. We therefore do not test the implicit/likelihood route to self-recognition; that half is cited from the literature (Panickssery et al., 2024; Zhou et al., 2025), not measured. Two further caveats follow from the instrument: generation runs inside the agent harness (a deployed-persona/coding-assistant system prompt we do not author or fully observe; see Appendix A), and temperature is the harness default and not controllable. We address both with a raw-API arm (§3.7) and a fully-crossed harness-vs-API design (§3.6); we do not claim any effect that would require temperature control or logprobs.

[1] A title collision worth flagging: Ardoin et al. (2026), despite the title "LLM Self-Recognition," concerns activation-level watermarking with no behavioral or privileged-access test and is unrelated to this work.

3.2 Primary measure: bias-free 2AFC

The primary task is two-alternative forced choice (2AFC): given two responses to the same prompt, one authored by the judge's tier and one by tier X, the judge is asked to select its own. Chance is 0.50. We adopt 2AFC over open-set naming for two reasons. First, open-set naming yielded a response bias — judges defaulted to a single modal label irrespective of truth (an early naming round: Cohen's $\kappa = +0.030$, permutation $p = 0.19$), manufacturing spurious column accuracy. Second, the pair presentation paradigm (PPP) reliably expresses an underlying signal that the individual presentation paradigm (IPP) collapses (Zhou et al., 2025, linear-probe self-vs-other 93–99% while the base-model behavioral IPP averages ≈15%, Zhou Table 3). Positions are counterbalanced by deterministic A/B alternation across trials, since LLM judges exhibit documented position and ordering bias in pairwise evaluation (Zheng et al., 2023; Wang et al., 2023) (the harness runtime exposes no RNG or seed, so the only stochasticity is the models' own sampling); a position-balance audit confirms ≈0.50 A/B per cell. **Names-equalized framing:** Tier names are given to all judges to equalize information across the self and neutral conditions. The main corpus is open-ended style/voice prompts chosen to be harness-orthogonal and divergence-eliciting. We dropped convergent tasks — those on which all tiers produce near-identical outputs, such as grade-school math — because they carried zero tier signal ($\kappa = 0.000$).

3.3 Decisive neutral baseline with two confound controls

Each text pair {X, Y} comprises one genuine response from each of two tiers (its two *members*); all three tiers judge it — each *member* as *self*, the non-authoring third tier as the *neutral/outsider*. A genuine self-advantage must survive two independent confounds, which must agree:

- **Judge-fixed:** the same judge rates a pair it is in (self) versus one it is out of (neutral), controlling judge capability.
- **Pair-fixed:** member judges versus the outsider on the identical texts, controlling difficulty.

This sets the evaluative bar at author-vs-neutral-judge: privileged self-access means beating a third-party judge of equal or lower compute (Song et al., 2025). Our high-powered main rounds (B1; Table 1) comprise 240 prompts per cell (2,160 trials), with bootstrap 95% confidence intervals (percentile method, resampling over prompts; Efron and Tibshirani, 1993).

3.4 Disagreement-conditioning

A self≈neutral null can be masked by inter-judge agreement, since a neutral rides public signal on consensus items (Ashuach et al., 2026). We condition on disagreement items — those where self and neutral attribute the pair differently (in 2AFC, exactly one is then correct) — and estimate P(self correct), testing whether any latent self-advantage emerges once consensus is removed.

3.5 Framing manipulation

Because explicit "you wrote this" attribution may not trigger the self-bias that implicit, on-policy framing does (Khullar et al., 2026), our framing branch (B2; Table 1) varies only question framing

across three conditions: **F0** explicit ("which did you write?"); **F1** implicit ("which of these two drafts sounds more like you?"); and **F2** ("you were weighing both — which do you keep?"). The two implicit framings (F1, F2) present the two texts as the judge's own prior drafts, with no tier names and no authorship verbs. The neutral baseline receives identical wording; only true authorship differs. We report a manipulation check on the judge's written reasoning to test whether each framing took hold: it measures the first-person-pronoun rate and the tier-naming rate in that reasoning. The check passes for F1 (a strong shift into first-person, non-taxonomic reasoning) but *not* for F2, whose first-person rate stays at the explicit-framing level (§4.2); we therefore treat F1 as the verified implicit framing and F2 as an exploratory condition that did not behave as intended.

3.6 Length-matched forgery paradigm

A forger tier writes deliberate imitations of another tier's style, length-matched to each genuine passage by word count (the harness exposes no native token counts; Appendix A). The ±10% match is a generation target, not a hard filter: B4 kept the forger's single draft, while B3 regenerated once if the first fell outside ±15%. A length audit confirms it largely held — 85% of forgeries (Sets 1–2) and 94% (Set 3) fell within ±10%, median length ratio ≈1.0. Judges — the forger, the genuine author (self), and a clean neutral — pick the genuine one from the pair {a genuine passage, a forgery} (chance 0.50; the forgery rounds, B4/B3 in Table 1; $n = 160$ per judge), isolating whether attribution runs on a style stereotype versus identity. We also run a 2×2 crossing text source (harness vs. API) against judge mechanism (harness vs. API) on the Opus-vs-Sonnet self cell, plus a uniform-temperature variant, to locate any harness amplification on the generation versus judging side; the temperature constraint that motivates the uniform-temperature variant is detailed in §3.7.

3.7 API verification and temperature constraint

We replicate the central null on the raw `/v1/messages` Claude API endpoint (`claude-opus-4-8` and `claude-sonnet-4-6`, which are undated 4.x aliases, plus the dated Haiku snapshot `claude-haiku-4-5-20251001`; $n = 40$), across F0 and F1 framings, to confirm results are not a harness artifact. One constraint applies, and it differs by arm. The harness exposes no temperature parameter at all: every subagent — generation and judging alike — ran at a single Claude Code default we can neither set nor read. The protocol's intended 0.7-generation / 0-judging split therefore never applied in the harness arms, which is where the main results come from. The raw-API arm is the only place temperature is settable, and even there only partly: Opus 4.8 rejects the temperature parameter, so on the API too Opus fell back to its default, while Sonnet and Haiku took the intended 0.7 (generation) and 0 (judging). The API verification arm is therefore mixed — Opus at its default, Sonnet and Haiku at 0.7/0. To check that this within-API mismatch is not what drives the API result, the uniform-temperature control (A1; §4.1) re-ran the same cell with no temperature parameter for any tier, so all three generated and judged at the API default; it matched the mixed condition (both below chance), ruling the temperature regime out. (It is an all-default control, not a temperature-zero one.)

3.8 Calibration metrics

On the main-round self-trials (the high-powered B1 rounds; Table 1; 240 prompts per cell), we compute the Brier score (Brier, 1950) and the confidence–accuracy gap (mean stated confidence when correct minus when wrong), with bootstrap 95% CIs on the gap, to characterize identity-judgment calibration against the generic-QA calibration baseline (Kadavath et al., 2022). This calibration branch is listed as B5 (Table 1).

3.9 Statistical analysis

Per-cell 2AFC accuracies are tested against chance (0.50) with an exact two-sided binomial test (two-sided by the minimum-likelihood method). Contrasts between two judges' accuracies use a two-proportion pooled z-test. Confidence intervals are percentile bootstrap: main-round (B1) and calibration intervals resample prompts as clusters, keeping all trials within a resampled prompt together ($B = 3{,}000$); the matched forgery contrast (B3) resamples trials independently within each arm ($B = 5{,}000$). Sensitivity is the standard 2AFC index $d' = \sqrt{2}\,\Phi^{-1}(p_c)$ (Macmillan and Creelman, 2005), with proportions numerically clamped away from 0 and 1; d' is reported descriptively, and accuracy with its interval is the primary statistic throughout. The exploratory pilot naming round is scored with Cohen's κ and a label-permutation test (20,000 shuffles). All p-values are reported uncorrected, as descriptive evidence against specific point nulls; every forgery cell reported as significant survives a Bonferroni correction across the nine role $\times$ set cells, and no claim in the paper rests on a marginally significant uncorrected result.

3.10 Preregistration

The prospective branches were preregistered on OSF before data collection: the framing branch (B2) and the primary forgery round (B4) at `osf.io/brdt8`, registered June 8, 2026; the matched forgery control (B3) at `osf.io/5azq8`, registered June 12, 2026. A capability-matched within-Claude neutral (Claude Fable 5, used as the framing-condition resolver "F1") was preregistered but could not be run, as Fable was not available during data collection.

3.11 Study key

Readers do not need our internal experiment codenames, but the prose and the analysis scripts use them, so we map them here. Each entry gives what the run measures, its sample size, and where the result appears. (Exploratory pilots are excluded.)

We keep F0/F1/F2 as framing-condition labels throughout; those are defined in §3.5.

3.12 Clean-room re-run (decontamination)

An initial collection of the harness rounds was confounded by the experimenter's local configuration. Spawned subagents — both generators and judges — inherited the project's `CLAUDE.md` (the experimenter's communication preferences) and a persisted auto-memory that named the study hypothesis. Inspecting the judges' reasoning traces, we found verdicts that referenced this leaked context (e.g., reasoning that a sample "matches the protocol style" the configuration described), which is a direct route to spurious self- or style-attribution. We therefore discarded the contaminated

Table 1: Study key and sample sizes. "Trials" are forced-choice (2AFC) judgments (chance = 0.50); n/judge is the trials seen per judge role (self, forger, neutral, as applicable); "—" marks column not applicable to that run.

	What it measures (design)	Trials	n/judge	Reported
Discrimination: can a judge tell the tiers apart? (2AFC, chance 0.50)				
PC	Positive control: neutral judges discriminate which tier wrote a text	720	80	§4.1
B1	**Central null:** self vs. neutral judge, judge- and pair-fixed	2,160	240 / 120	§3.3, §4.1–4.4
B2	Framing manipulation (three framings, 1,080 trials each)	3,240	—	§3.5, §4.2
A1	API disambiguation: fully-crossed 2 × 2 (text source × judge mechanism)	—	40/cell	§4.1
Forgery: "pick the genuine passage" from {genuine, length-matched forgery}				
B4	Primary forgery, Set 1 (genuine Opus, Sonnet forges)	480	160	§3.6, §4.3
B3	Matched forgery control, Sets 2–3	960	160/set	§3.6, §4.3, §6
Derived and supporting evidence				
B5	Calibration metrics, derived from B1 self-trials (no new calls)	—	240	§3.8, §4.4
Int	Subject interviews: free-form answers (3 tiers × 5 questions × 3 reps)	45	—	§3.12
	Harness totals across all runs: 2,280 generations, 7,560 2AFC trials.			

harness arms and rebuilt the experiment byte-faithfully from the archived run specifications, then re-collected every harness round in a verified clean room: no project `CLAUDE.md` on the subagent path, a fresh memory key, and a neutral working directory, audited before each pass. To convert the fix into added power rather than a single replacement, we ran the clean instrument twice — two fully independent passes, each drawing a fresh prompt/sample at the harness-default temperature — and pooled them, doubling n on every harness measure (B1, B2, B4, B3, B5; the per-cell sizes in §3.10 are the pooled totals). Each pass reproduced the central null on its own. The raw-API arm (§3.7) carried no harness context by construction — it never loaded a project file or memory — so it was structurally clean and is reported from its original collection without re-run. The decontaminated, pooled results confirm the four-part headline and revise three secondary claims; §4 reports the clean numbers throughout, and §6 records what moved.

3.13 A-priori subject interviews

Independently of the forced-choice task, and before seeing any pair, each tier was interviewed in the same clean room as a no-tools subject (3 tiers × 5 questions × 3 repetitions = 45 verbatim answers; `interviews/`). The five questions per tier are one *recognize-self* item ("what features of your *own* writing would you look for to identify which sample is yours?"), two *describe-other* items (one per sibling tier), and two *forge-other* items ("what features would you exploit or exaggerate to pass as tier X?"). We use these stated self-models two ways. First, qualitatively, as an a-priori

check on whether tiers invoke an episodic, authorship-grounded signal or a third-person stylistic one. Second, quantitatively: we code each answer for a fixed feature vocabulary (structure, hedging, concreteness, figurative language, register, density, etc.) and compare this *stated* feature profile against the *revealed* feature profile coded the same way from the judges' free-text reasoning traces in the forced-choice rounds (6,840 traces), yielding a stated-vs-revealed contrast per tier. The coding script (`analysis/reason_coding.py`) is keyword-based and deterministic; we report it as a descriptive overlay on the behavioral results, not an independent inferential test.

4 Results

4.1 Tier style is externally legible — and the legibility is persona-amplified

The behavioral null reported below is only interpretable against a positive control that establishes the texts are discriminable at all. They are, and the control is robust. Table 2 (plotted as Figure 1) reports third-person tier discrimination by an in-harness neutral judge — a model that authored neither of the texts in a given 2AFC pair (chance 0.50). Discrimination is capability-graded, with the two stronger judges effectively tied at the top: Opus reaches 0.91 ($d' = +1.91$), Sonnet 0.90 ($d' = +1.81$), and Haiku 0.63 ($d' = +0.45$), where d' is the sensitivity index of signal detection theory (Macmillan and Creelman, 2005).[2] The signal is real and substantial for the two stronger judges and weak but above chance for Haiku; the *ability to read* tier style scales with the reader's capability. The dedicated control replicates in a second instrument: the neutral-out cells of the main self-recognition rounds (B1) give 0.933 / 0.938 / 0.662 for the same three judges — an independent estimate showing the same Opus ≈ Sonnet ≫ Haiku ordering. Within-family tier style is therefore externally legible.

Table 2: Third-person tier discrimination (in-harness neutral judge, 2AFC, chance 0.50).

Neutral judge	Accuracy	d'
Opus	0.91	+1.91
Sonnet	0.90	+1.81
Haiku	0.63	+0.45

One complication forced a rescoping of this control. When we verified the central null on the raw API (`claude-opus-4-8`, `claude-sonnet-4-6`, `claude-haiku-4-5`), discriminability for the closest pair — Opus versus Sonnet — diverged from the in-harness figure, raising the question of whether the positive control was partly a harness artifact. We resolved it with a fully-crossed 2×2 (Figure 2; the API-disambiguation bundle, A1 in Table 1) over an Opus-vs-Sonnet self-exemplar, crossing text source (harness vs. API) against judge mechanism (harness vs. API). Discriminability tracked text source, not judge: harness-generated text scored 0.675 (harness judge) and 0.825 (API judge), whereas API-generated text scored 0.350 and 0.375 under the same two judges. The judge mechanism barely moved the result; the generation mechanism moved it decisively. The effect is not a temperature artifact — a uniform-temperature condition (A1, in which all three tiers run with *no*

[2] d' estimates near ceiling are sensitive to how extreme proportions are handled (Hautus, 1995); raw accuracy with bootstrap CIs is the primary statistic throughout, and d' is reported descriptively.

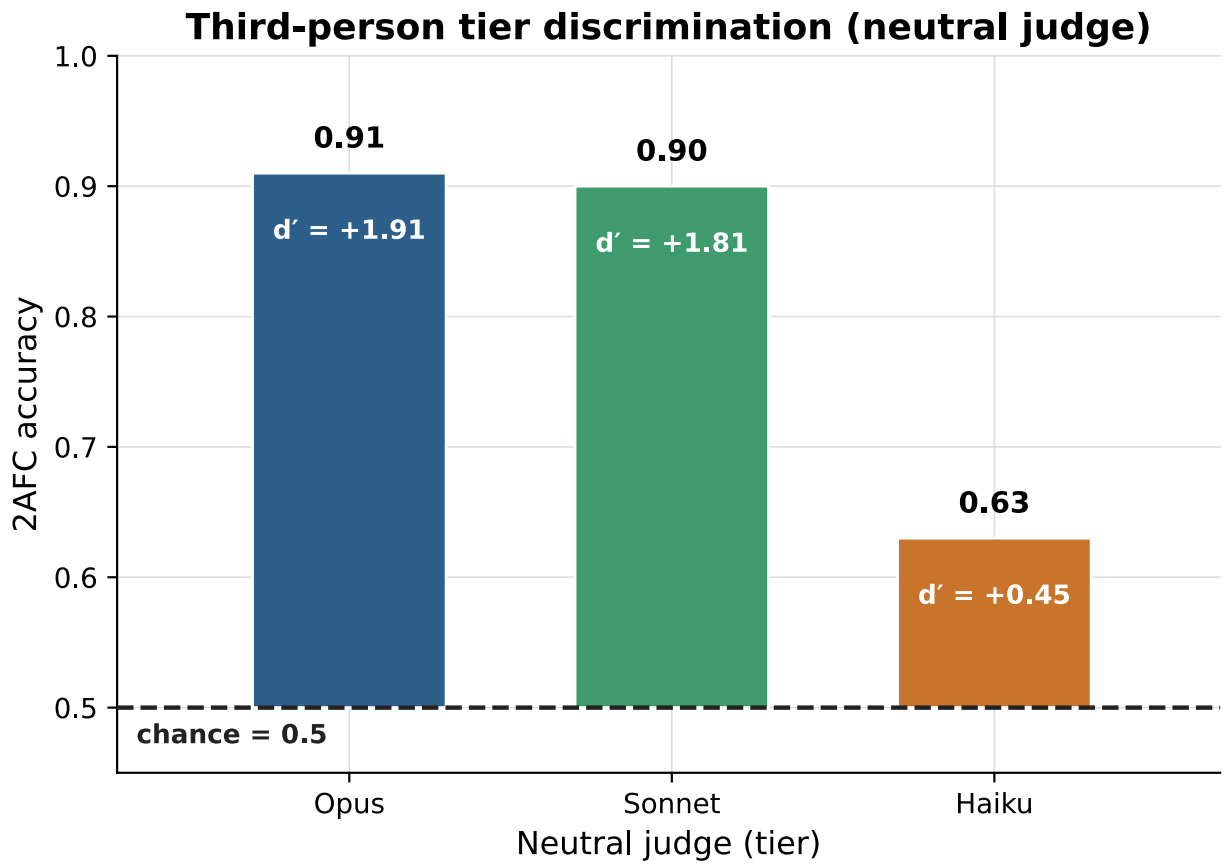

Figure 1: Positive control: neutral-judge tier-authorship discrimination accuracy by judge (chance 0.50), with the corresponding sensitivity index d' (higher = more discriminable, chance $d' = 0$); the ability to read tier style is capability-graded.

temperature parameter, i.e. at the API default, not at 0; §3.7) matched the mixed-temperature condition, both near or below chance (Opus-self 0.45 vs. 0.375; Sonnet-self 0.175 vs. 0.20). And a capable judge keeps the signal alive even on bare text: Opus-neutral discriminating {Sonnet, Haiku} scored 0.775 $[0.65, 0.90]$, retaining clear signal off-harness, if somewhat below the in-harness $\approx$0.91 for that cell (Table 2). (This 2×2 bundle is from the structurally-clean API collection and was not part of the harness re-run.) The positive control thus survives off-harness for a strong reader, but its magnitude for the closest pair is harness-dependent.

This yields a rescoped claim and a new finding. The rescoped claim: within-family tier style is externally legible and the legibility is capability-graded, but the *strength* of that legibility for the closest tier pair is amplified by the harness rather than being a fixed property of the underlying models. The new finding: the deployed-persona/agent harness amplifies tier-stylistic distinctiveness at *generation*. The effect is generation-side, largest for the closest pair (Opus$\approx$Sonnet) and for weaker judges — the regimes where bare-API text carries the least tier signal. The harness does not help judges read style; it makes the writers write more tier-distinctively.

This connects to the persona-conditioning literature. The shared "Assistant" persona gives every tier a common baseline. Tier-specific style then shows up as deviation from that baseline, and the

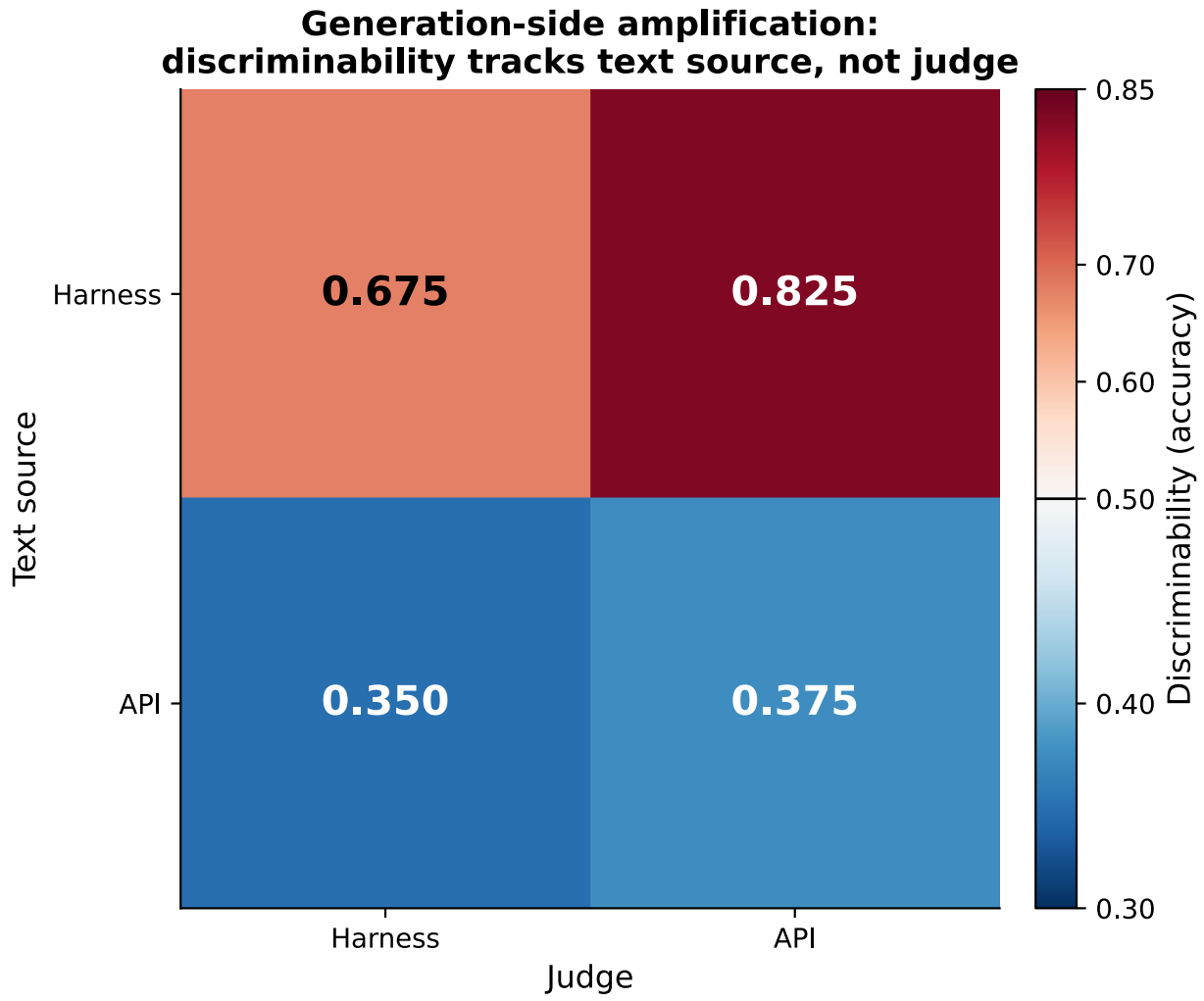

Figure 2: The persona-amplification 2 × 2: discriminability as a function of text source (harness vs. API) crossed with judge mechanism (harness vs. API), for the Opus-vs-Sonnet self exemplar.

persona conditioning sharpens those deviations. Asvin G. (2026) casts the Assistant persona as a privileged reference frame, and our harness runs inside exactly that frame; Lu et al. (2026) give the mechanistic counterpart, an "Assistant Axis" in activation space whose anchoring homogenizes the persona. Read against both accounts, our generation-side amplification is the behavioral residue once that shared persona is conditioned in. We observe this only behaviorally and claim no logit-level mechanism, but the direction is what these accounts predict.

4.2 No privileged self-access (the central result)

The headline result is negative and, in the contrast structure we use, decisive: being the author of a text confers no advantage in identifying it. This stands in direct dissociation with the strong positive control of §4.1, where neutral third-party judges discriminate tier authorship at well above chance and do so in a capability-graded way (Opus 0.91 ≈ Sonnet 0.90, Haiku 0.63; 2AFC, chance

0.50). Tier identity is legible — but it is legible to *others reading the text*, not privately known to its author.

Main rounds at power. On the high-powered main rounds (B1; Table 1) we estimate a self-minus-neutral advantage (positive ⇒ privileged self-access) under two independent confound controls on 240 prompts per cell, with bootstrap 95% confidence intervals (Figure 3). The judge-fixed contrast (same judge on a pair it is *in* versus one it is *out* of, controlling judge capability) yields Opus -0.090 $[-0.131, -0.046]$, Sonnet -0.277 $[-0.327, -0.229]$, and Haiku -0.035 $[-0.110, +0.042]$. The pair-fixed contrast (member versus outsider on the identical texts, controlling difficulty) gives a pooled -0.134 $[-0.160, -0.106]$. The corresponding self-2AFC accuracies — Opus 0.844, Sonnet 0.660, Haiku 0.627 (all above chance) — superficially read as a graded self-recognition hierarchy, but the controls show this is the same third-person discriminability of §4.1 misread as self-knowledge: Opus scores high on its own pairs because it is the best discriminator of *everyone's* text. No tier shows a positive advantage in either control; with the tighter pooled intervals, Opus is now itself significantly *negative* on the judge-fixed contrast while Haiku's brackets zero — the two swap which one sits at zero, but neither is positive. Our preregistered decision rule — null confirmed at power when the interval brackets zero or is negative in *both* controls with no tier positive — fired.

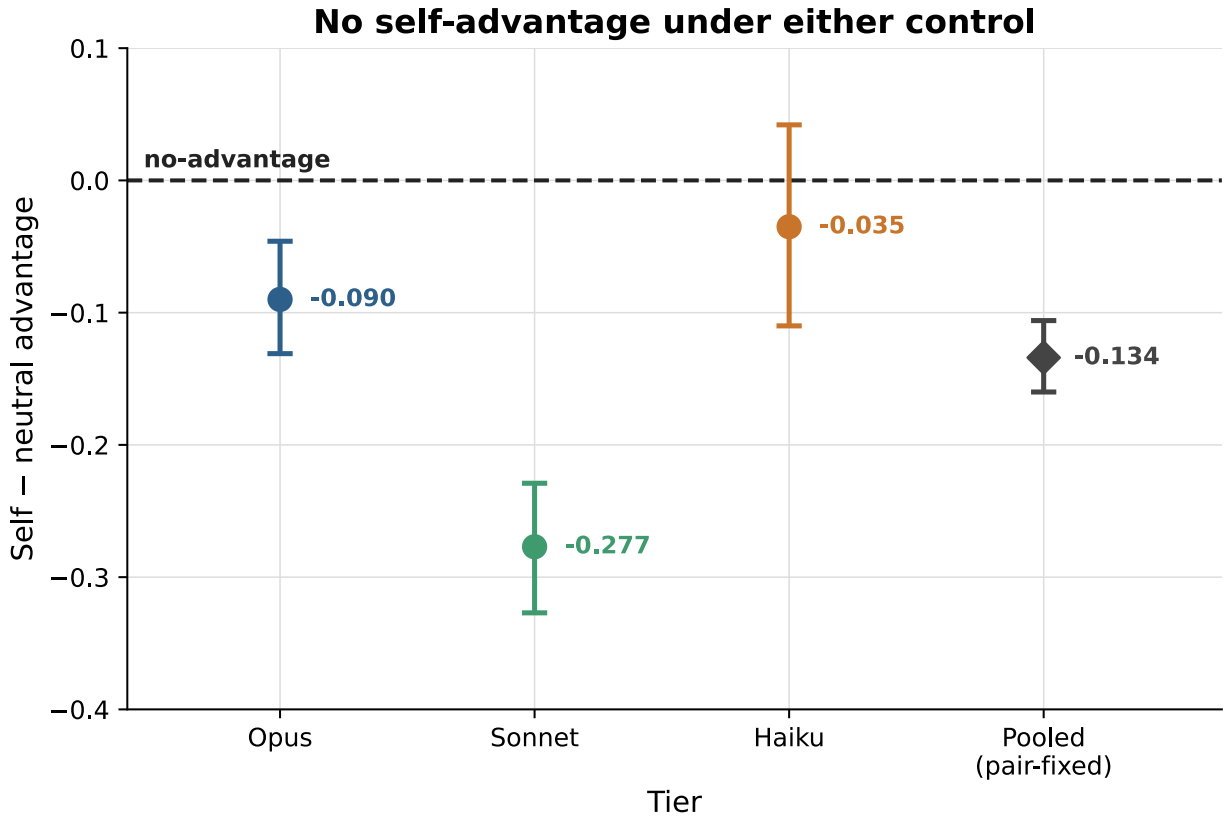

Figure 3: Self-minus-neutral advantage under both confound controls (judge-fixed, per tier; pair-fixed, pooled), with bootstrap 95% CIs; zero marks no advantage. The disagreement-conditioned $P(\text{self correct}) = 0.243$ is reported in §4.2.

Disagreement-conditioning hardens it. A self≈neutral null can be an artifact of inter-judge agreement: on consensus items a neutral rides public signal for free, masking any genuine private advantage (Ashuach et al., 2026). We therefore condition on disagreement items, where self-judge and neutral attribute the pair differently and, in 2AFC, exactly one is correct. The masking mechanism is present in our data — judges agree on 74% of items and self-accuracy on those consensus items is high (0.875) — yet removing the mask does not reveal a hidden self-advantage. Pooled across tiers, $P(\text{self correct} \mid \text{contested}) = 0.243$; equivalently, the neutral wins 0.757 of these contested calls (the two judges disagree by construction, so exactly one is right). The null *hardens* rather than reverses. The contested-item win-rate is itself capability-graded — Opus 0.630 ($n = 81$, $p = 0.026$) > Sonnet 0.250 > Haiku 0.043 — and only the strongest tier exceeds chance on contested items. This is not a self-edge but the capability signal of §4.1 read off a harder subset: on its own pairs Opus simply out-reads the weaker neutral it is contrasted against, and the same Opus is, if anything, slightly *worse* when its own writing is in the pair — a *negative* judge-fixed self-advantage (-0.090 $[-0.131, -0.046]$). (An earlier, lower-powered collection had retired this Opus contested lead as a non-replication; at doubled n it is a real, capability-parsimonious effect, and we report it as such rather than as privileged self-access.)

Framing-robust under a successful manipulation. Explicit "which did you write?" prompts may be the regime where self-attribution is weakest: implicit, on-policy framing is the stronger trigger for self-directed behavior (Khullar et al., 2026) — framing in which the candidate texts appear as the model's own prior conversational turns, so authorship is implied by structure rather than stated. We re-ran the matched-pair design varying only question framing (Figure 4): F0 explicit (the main design), F1 "which of these two drafts sounds more like you?", and F2 "you were weighing both — which do you keep?", with no tier names or authorship verbs in F1/F2. A manipulation check shows the implicit framing took hold for F1 but not F2. F1 reasoning traces (the written justification each judge gives for its verdict) are 72% first-person and never name a tier (0%), versus F0's 22% first-person and 91% tier-naming — a clear shift into first-person, non-taxonomic reasoning. F2, by contrast, is first-person in only 12% of traces — *below* explicit F0's 22% — so it did not behave as an implicit framing at all. We therefore rest the implicit-framing claim on the manipulation-verified F1 and demote F2 to an exploratory condition rather than treat it as a second implicit replicate. The self-advantage is null under both regimes that bear weight — explicit F0 and verified-implicit F1 — and under the exploratory F2 as well: across all three framings, every judge-fixed estimate is negative for every tier (Figure 4). No tier is positive under any framing: even the most favorable cell — Opus under F2 — has a 95% interval that includes zero rather than excluding it on the positive side (-0.058 $[-0.121, 0.000]$), and the lone marginally-positive Opus-F1 cell of the earlier collection is now negative (-0.108). The "wrong regime" objection is therefore closed by a framing whose manipulation we can verify (F1), not by F2's uncertain one.

Not a harness artifact. Because generation and judging ran inside an agent harness, we replicated the core contrast on the raw `/v1/messages` API (`claude-opus-4-8`, `claude-sonnet-4-6`, `claude-haiku-4-5`; $n = 40$). This arm carried no harness context by construction — it never loaded a project file or memory — so it was structurally clean and was not part of the harness re-run (§3.11). The central null reproduces: self-advantage on the API is F0 -0.16 and F1 $+0.05$, against matched harness values of -0.09 and -0.13. No framing × surface cell crosses into a tier-positive

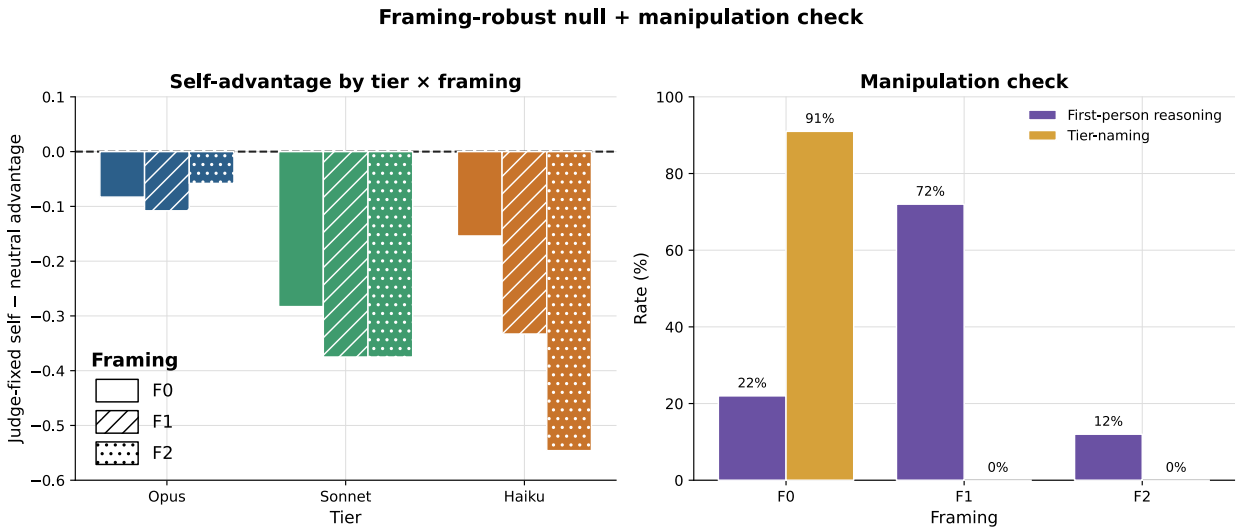

Figure 4: Framing × tier: judge-fixed self-advantage across F0 (explicit), F1, F2 (implicit) for each tier, with the manipulation-check panel (first-person rate, tier-naming rate).

advantage at this sample size, and the F1 point estimate of $+0.05$ $[-0.150, +0.238]$ is negligible and not distinguishable from zero at $n = 40$. The absence of self-access is not an artifact of the harness.

These results extend the family-level discrimination of Choi et al. (2025) down to within-family tiers while sharply separating two capacities the naive hypothesis conflates. Tier style is a third-person, externally read property; producing the text grants its author no privileged behavioral handle on it.

4.3 Attribution runs on stereotype: an LLM caricature advantage

The forgery probe (the forgery rounds B4 and B3; Table 1) replaces the question "which text is yours?" with "pick the genuine passage" from a pair {*a genuine passage, a deliberate same-tier forgery*}, the forgery length-matched to it (a ±10% word-count target; §3.6). The match closes the length confound that contaminated the pilot, whose forgeries ran markedly longer than the genuine texts. This isolates whether attribution tracks an identity signal or a stylistic prototype, and the answer is consistently the latter.

Three forgery sets run through this section (Figure 5 plots all nine role × set cells); for reference:

Table 3: The three forgery sets.

Set	Genuine (target)	Forger	Direction
Set 1	Opus	Sonnet	forging *up* (weaker → stronger)
Set 2	Haiku	Sonnet	forging *down* (stronger → weaker)
Set 3	Sonnet	Opus	forging *down* (stronger → weaker)

The forger runs on stereotype (Set 1). Judging {genuine Opus, Sonnet-forges-Opus}, the genuine author identifies its own text at 0.725 ($p = 1 \times 10^{-8}$), confirming the texts are separable. The forger does not merely fail to find the genuine Opus; it inverts. Sonnet, the author of the

forgery, identifies the genuine Opus only 18.8% of the time ($p = 5 \times 10^{-16}$) — equivalently, it picks its own caricature as "the real Opus" on about 81% of pairs, far below chance. A clean neutral judge (Haiku) is at chance (0.537, n.s.). A model deliberately imitating a tier prefers its own exaggerated rendition of that tier over the tier's actual output — the behavioral analog of adversarial-stylometry imitation attacks, where an imitator drives attribution toward the framed victim's profile (Brennan et al., 2012).

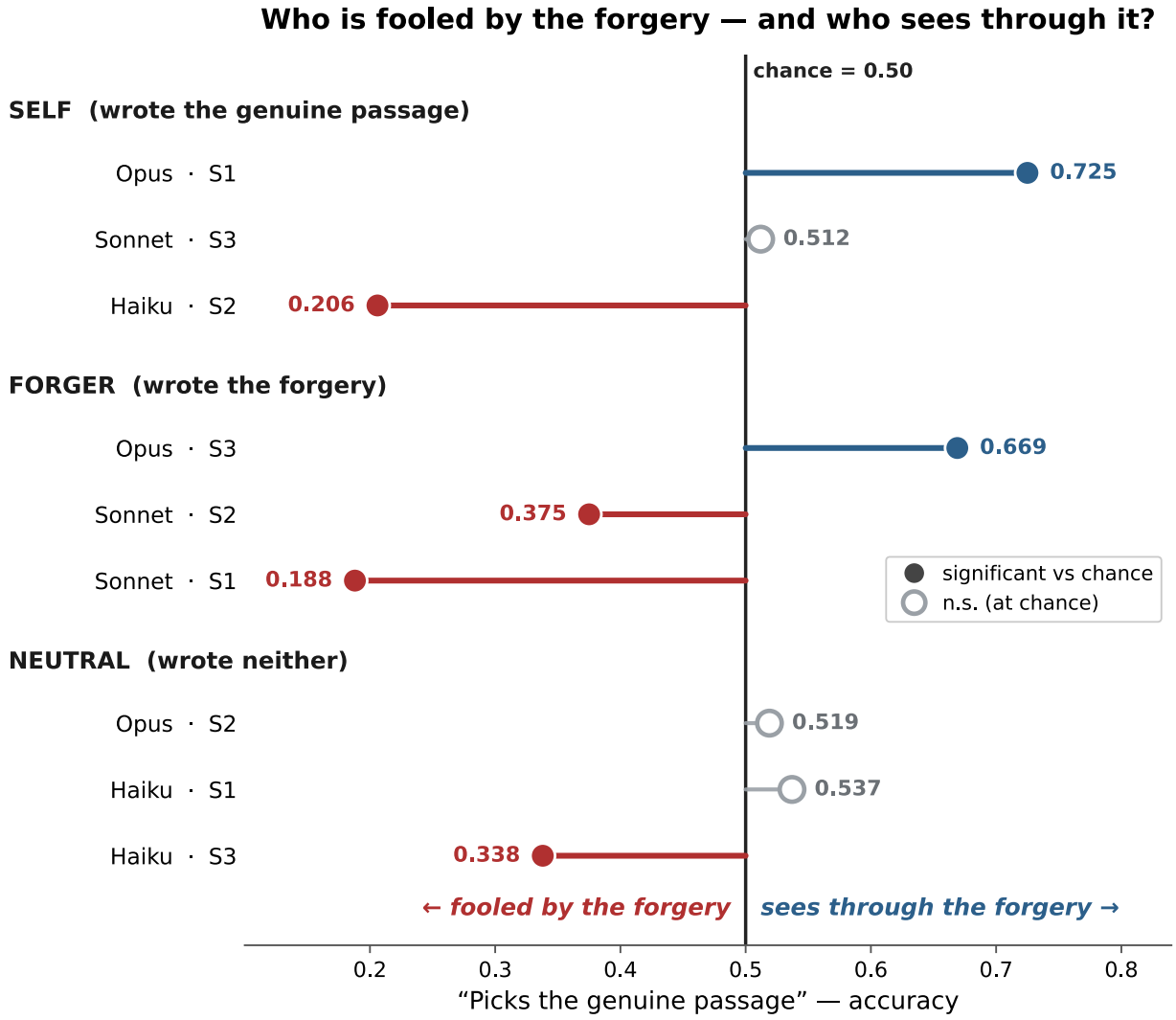

Figure 5: The forgery paradigm, by judge role. Each judge's *"picks the genuine passage"* accuracy (chance 0.50) across the nine {role × set} cells, drawn as a deviation from chance — below chance the judge is *fooled by the forgery* (the caricature); above chance it *sees through the forgery*. The caricature captures the prototype-holders: in the strongest forge-down set the genuine author of the weaker tier (Haiku-self 0.206) and its forger (Sonnet-forger 0.375) fall below chance, as does the same forger judging its own forge-*up* set (Sonnet-forger 0.188), while a capable neutral sits at chance (Opus 0.519). But the effect is bounded and capability-graded — a genuine author still sees through a forge-up (Opus-self 0.725), the strongest forger is not fooled by its own caricature (Opus-forger 0.669), and even a weak neutral can itself be pulled below chance (Haiku 0.338).

The caricature defeats the author and the forger — but not a capable outsider (Set 2). Judging {genuine Haiku, Sonnet-forges-Haiku}, the two judges that hold the prototype both fall below chance: the forger (Sonnet-forger 0.375, $p = 0.002$) and, decisively, the genuine author (Haiku judging its own passage at 0.206, $p = 3 \times 10^{-14}$). The genuine author is the worst at telling its own passage from the forgery: the caricature is judged more Haiku than Haiku, by Haiku. But the effect is *not* universal across judges. The capable neutral that wrote neither passage — Opus — sits at chance (0.519, n.s.): it cannot pick out the genuine Haiku, but it is not fooled by the forgery either. The super-caricature thus captures those who hold the prototype (the author, who instantiates it; the forger, who built it by exaggerating it), while a strong third party reads through it to a coin-flip. This is a sharper, more bounded claim than a blanket "all judges below chance," and it is the one the clean data support: the stereotype out-competes the genuine article *for the prototype-holders*, not for everyone.

We name this an LLM instance of the *caricature advantage*, or *peak shift*: an exaggerated deviation from a category prototype is recognized or rated as more category-typical than the *veridical exemplar* — the genuine article itself. The effect is well-established for face recognition, where caricatures are identified faster than veridical drawings (Rhodes et al., 1987) and the advantage is framed as a face-space distinctiveness effect (Lee et al., 2000); its proposed neural basis is a superstimulus exploiting prototype-relative coding (Ramachandran and Hirstein, 1999). Our result is a caricature advantage in *preference during evaluation*. This distinguishes it from Cheng et al. (2023), which operationalizes caricature on the *generation* side — LLM persona-simulations exaggerating the groups they imitate — but does not report a judge preferring a caricature over a genuine passage, and does not address self-style. The complementary direction, that LLM imitations remain stylometrically separable from genuine text under external classification (Wang et al., 2025), is consistent with the genuine author retaining separability above (Set 1, 0.725) and with the capable neutral reading Set 2 at chance rather than below it.

Forger self-bias attenuates with capability (Set 3). Judging {genuine Sonnet, Opus-forges-Sonnet}, the strongest forger does *not* prefer its own caricature: Opus-forger picks the genuine Sonnet at 0.669 ($p = 2 \times 10^{-5}$), resisting the self-bias that captured Sonnet in Set 1 (0.188). This forger-capability attenuation is the cleanest cross-set regularity: the model that builds the caricature is decreasingly fooled by it as its own capability rises. The companion asymmetry between forging *down* (a stronger model caricaturing a weaker tier) and forging *up* is real but weaker than the pilot suggested, and we state it as a tendency. Forging down degrades detection of the genuine passage for the prototype-holders: in Set 2 the author falls below chance (Haiku-self 0.206), and in Set 3 the genuine author lands only at chance (Sonnet-self 0.512, n.s.) while the weak neutral is itself pulled below chance (Haiku-neutral 0.338, $p < 1 \times 10^{-4}$). Forging up (Set 1), by contrast, leaves the genuine passage identifiable by its author (Opus-self 0.725). But the pattern is not uniform: a capable neutral still identifies the genuine passage only at chance in Set 2 (Opus 0.519), and the author-defeat is unambiguous only in the strongest forge-down case. A label-/identity self-preference that strengthens with assigned identity rather than true authorship has independent precedent (Lehr et al., 2025); closed-menu work likewise finds models select a "best"-answer proxy rather than a genuine self-signal (Davidson et al., 2024).

The author's residual edge remains formally open. Opus's genuine-vs-neutral advantage in Set 1 (Opus-self 0.725 vs. Haiku-neutral 0.537, +0.188, $p = 0.001$) is real but capability-confounded.

The preregistered matched control (B3; Table 1) — Opus judging a forgery set it did not author (Set 2) — now reads cleaner than in the pilot. The Set-2 baseline does *not* collapse (Opus-neutral 0.519, at chance rather than below), so the registered *target-distinctiveness* caveat — its prespecified check that the genuine passage in the comparison set is still identifiable at all — does not fire, and the contrast is interpretable. It gives Opus-self(Set 1) 0.725 vs. Opus-neutral(Set 2) 0.519 = +0.206 [+0.100, +0.306]. But this matched control is confounded by *which tier's genuine passage must be identified*, not authorship: that tier is Opus in Set 1 (the most externally legible; §4.1) and Haiku in Set 2 (the least). The same Opus judge should therefore do better on Set 1 than Set 2 for any reader, author or not. Consistent with that, the +0.188 edge over the clean Haiku neutral falls below Opus's general third-person discrimination advantage over Haiku (≈0.28; §4.1 / Table 2). The clean resolver is a non-author of capability ≥ Opus that *also* shares the family prior. The only qualifying Claude model is Fable 5, which was not available during the study; a cross-family neutral cannot substitute (it would confound family with authorship; §6). The parsimonious reading is capability, not self-familiarity; the edge is formally open.

4.4 A capability-graded calibration gradient

Self-identification is one metacognitive axis; confidence in one's own correctness is another. To test whether the failure of behavioral self-recognition extends to metacognition more broadly, we examined the calibration of the identity judgments themselves. For each main-round self-trial (B1; Table 1) we recorded the judge's stated confidence and scored it against trial accuracy, pooling to 240 prompts per cell and computing two summaries: the confidence–accuracy gap (mean confidence when correct minus when wrong) with a bootstrap 95% CI over prompts, and the Brier score.

Both summaries are capability-graded and separate the three tiers cleanly (Figure 6). The confidence–accuracy gap is Opus +0.091 [+0.080, +0.102] > Sonnet +0.030 [+0.021, +0.039] > Haiku −0.012 [−0.021, −0.003]; all three intervals exclude zero, are mutually non-overlapping, and the weakest tier's is negative. Brier scores track the same ordering ($0.128 < 0.214 < 0.253$). The most capable tier is well-calibrated — it assigns higher confidence to judgments it gets right — while the weakest tier is *mildly anti-calibrated*, assigning slightly higher confidence to the identity judgments it gets wrong. This is a monotone *gradient* that crosses zero at the bottom: confidence–accuracy alignment is strong at Opus, weak but positive at Sonnet, and mildly inverted at Haiku. The inversion is small — a gap of just −0.012 — but with the full sample its interval clears zero on the negative side, where a lower-powered collection had left it within noise. We report it as a mild anti-calibration rather than as the gap-of-zero an earlier reading inferred.

This result is informative precisely as a contrast class. Language models are broadly calibrated on generic question-answering confidence (Kadavath et al., 2022); the present finding shows that, on the within-tier identity-judgment task, such alignment is graded across tiers — strong at the most capable, and mildly *inverted* at the weakest, which is confidently wrong as often as confidently right. The two axes dissociate within the same models: confidence–correctness calibration is graded and, at the top tier, real, even though none of the tiers shows any privileged access to their own authorship (§§4.1–4.2). Metacognitive confidence and self-*identification* therefore behave as separable capacities here, and the calibration gradient survives independently of the central null on self-recognition.

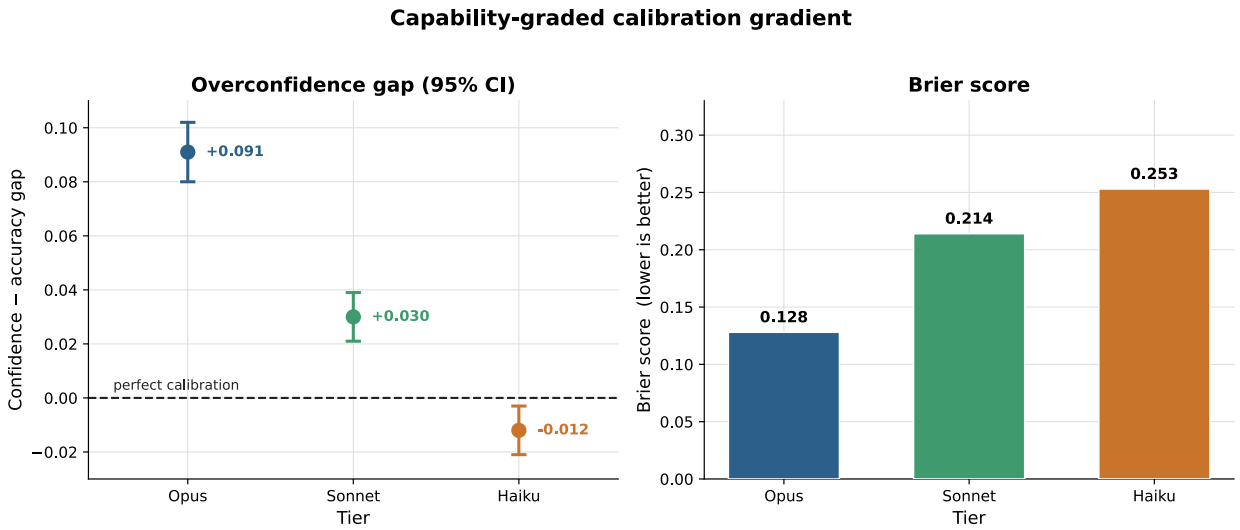

Figure 6: Calibration gradient: confidence–accuracy gap and Brier score per tier against the generic-QA calibration baseline.

4.5 Subjects' stated self-models corroborate the behavioral picture

The forced-choice results so far are third-person: they read out what the models *do*. A complementary question is what they *say* — and, separately, whether what they say lines up with what they do. We addressed both with the a-priori interviews (§3.12), in which each tier described, before seeing any pair, how it would recognize its own writing and how it would forge a sibling's. Three results follow, and each leans the same way as the behavioral data.

Asked how they would recognize their own writing, all three tiers answer in pure third-person style terms. Not one invokes an episodic, "I-remember-writing-it" signal; every answer is a description of stylistic regularities that could equally be applied by an outside reader. Opus names "a cluster of habits that tend to appear in my writing more than in Sonnet's or Haiku's" — structural hedging, a concession-then-development rhythm, longer subordinated sentences. Sonnet points to sentence-rhythm variation and a "moderate vocabulary register" between Opus's elevation and Haiku's plainness. Haiku lists brevity, "less nested complexity," and fewer hedges, and explicitly grounds the list in "Haiku's smaller model size." The subjects, in other words, model their own identifiability exactly as our central result says it works — as an externally legible style, not a privately held authorship trace. They corroborate the null in their own words.

The forgery strategies are stated peak shift. Asked what they would do to pass as a sibling, every tier describes deliberate *exaggeration* of the imitated tier's prototype, which is the caricature mechanism of §4.3 stated outright. Forging up, Sonnet says it would "lean into" Opus's "density and deliberateness," "epistemic precision," and "philosophical register even in practical contexts." Forging down, Opus says it would "exaggerate the rhythm of three" and plant em-dash reversals because "the em-dash reversal is practically a fingerprint." A strategy of exaggerating prototype-typical features is precisely what produces a superstimulus that out-competes the genuine article; the interviews show the mechanism is not just visible in the behavior but explicitly available to the models as a plan.

Stated and revealed feature-profiles agree on the dominant features but diverge in informative ways. Coding the interview answers (stated) against the judges' reasoning traces (revealed) on the same feature vocabulary, the two align on each tier's leading features — structure, hedging, and concreteness recur in both — but with honest gaps that mark the stated model as a *theory*, not ground truth. Every tier over-reports hedging relative to how often it actually reasons from it (stated vs. revealed: Opus 93% vs. 27%, Sonnet 100% vs. 37%, Haiku 93% vs. 45%), and Haiku markedly *under*-reports its own figurative language (20% stated vs. 66% revealed). A robust secondary finding falls out of the same coding: in the revealed traces, Opus reasons in the first person far more than its siblings (52% vs. Sonnet 18%, Haiku 20%), echoing the capability gradient of §§4.1–4.4 on the production side. We present this overlay descriptively. It does two things for the paper: it shows the tiers possess a workable but imperfect self-model of their own style, and — by making the stated/revealed gap explicit — it pre-empts the reading that any of our self-judgments tap a privileged inner record rather than the same outward style every reader sees.

5 Discussion

5.1 A unifying account: attribution runs on amplifiable surface stereotype, not identity

The results cohere around one mechanism. When a Claude 4 tier judges which of two texts it wrote, it does not consult a record of having produced one of them; it reasons *"this looks like tier X"* and answers accordingly. Three otherwise-separate findings are facets of this single phenomenon:

1. **The central null.** Being the author confers no advantage over a neutral judge on the same texts — negative or bracketing zero under both confound controls — and it hardens, not softens, on disagreement-conditioned items, where private signal would have the most room to appear (Ashuach et al., 2026) (§4.2, Figure 3).

2. **The caricature effects.** A stronger model's stylized rendition of a weaker tier is judged more genuine than the genuine article — even by the author (Set 2: Haiku judges its own passage below chance, 0.206), and even by the forger, though a capable neutral that holds no prototype reads it at chance; the forger's own self-bias attenuates as forger capability rises (§4.3, Figure 5). If attribution were keyed to identity, exaggerating the prototype could not beat the original; that it does, for those who hold the prototype, is the signature of a stereotype prior — the verbalized analog of the peak-shift / caricature advantage from face recognition.

3. **Persona-amplified legibility.** The deployed-agent harness amplifies tier-stylistic distinctiveness *at generation* — the fully-crossed 2×2 localizes the effect to text source, not judge mechanism (harness-text/API-judge 0.825 vs. API-text/harness-judge 0.350), most for the closest pair (§4.1, Figure 2). The prior is not only read off the surface; the harness writes more of it into the surface.

All three say the same thing: the discriminable object is an externally legible, capability-graded, externally-amplifiable style, and the judgment is third-person inference over it, regardless of who holds the pen. The interviews (§4.5) supply a fourth, converging line from the models' own mouths:

asked how they would know their own writing, all three describe a third-person style signature and none a memory of authorship; and asked to forge a sibling, all three describe exaggerating the target's prototype — the caricature mechanism, stated as a plan before any pair is seen.

5.2 Reconciling with introspection-*for* and introspection-*against*

This does not contradict the positive introspection results; it bounds them by operationalization. Lindsey (2026), Macar et al. (2026), and Naphade et al. (2026) report privileged access via *injected internal state* and *behavior prediction* — Naphade's models outperform peer models in predicting their own behavior. These are different objects from verbalized prose self-authorship, and the *for*-side evidence is itself characterized as fragile and narrow — including by its proponents: Lindsey (2026) stresses the injection capacity is "highly unreliable and context-dependent," with downstream work finding injection-detection that registers *that* an anomaly occurred but not *what* (Lederman and Mahowald, 2026), and introspection that holds on simple tasks but fails out of distribution (Binder et al., 2024). Our null is therefore not a counterexample to those claims but a boundary on their reach: even on the family where concept injection works (Opus 4 / 4.1), behavioral prose self-recognition does not.

On the *against* side, we apply the privileged-self-access bar of Song et al. (2025) — outperforming a third-party predictor of equal or lower compute — at the tier level, and the tiers fail it for their own writing, the behavioral analog of their failure for their own temperature. This makes our result a concrete instance of the Singh et al. (2026) reality check that behavioral evidence alone does not establish introspection, with Lederman and Mahowald (2026) and Pearson-Vogel et al. (2026) — verbal denial alongside residual-stream detection — as allies on the implicit/explicit split.

5.3 What we measured, and what we cite

We are explicit about the division of labor. We supply the *behavioral* half: forced-choice self-authorship judgments, with two confound controls, framing variation, and a raw-API replication. We do not measure the *implicit/likelihood* half — the subagent harness exposes no token logprobs — and we cite it: the logprob-based self-recognition/self-preference link (Panickssery et al., 2024), the perplexity-driven self-preference mechanism (Wataoka et al., 2024), the linear-probe-vs-behavioral gap (Zhou et al., 2025), and the mechanistically distinct routing of verbal recognition with a size-graded implicit self-advantage (Asvin G. and Lindsey, 2026). The gap between a probe that reads self-vs-other at 93–99% and a behavioral self-advantage indistinguishable from zero is precisely the boundary we mark, not one we explain.

6 Limitations

Behavioral axis only. The instrument is Claude Code subagents spawned per tier as an API stand-in, which exposes tier-labeled generation and forced-choice judging but not token-level logprobs. All of our measures are therefore behavioral. The implicit/likelihood half of self-recognition — the regime that logprob-, perplexity-, and probe-based work reports (Panickssery et al., 2024; Wataoka et al., 2024; Zhou et al., 2025; Asvin G. and Lindsey, 2026) — is cited from the literature,

not measured here; we supply only the behavioral half and make no in-house claim about the likelihood-vs-explicit dissociation.

Original contamination, and the decontaminated re-run. The harness arms were first collected under a confounded configuration: spawned subagents inherited the experimenter's `CLAUDE.md` and an auto-memory naming the hypothesis, and some judge traces referenced that leaked context (§3.11). We treat the original harness arms as unusable and report only the clean re-run — a byte-faithful rebuild run in a verified clean room over two independent passes, pooled (§3.11). This is a strength for the numbers reported but a limitation on provenance: the contamination was caught by trace inspection rather than prevented by design, and the clean instrument, while audited, is the same harness whose persona conditioning we separately show is non-neutral (below). The decontaminated, doubled-n result confirms the four-part headline and tightens it (the central null hardens; the positive control, caricature core, and calibration gradient all replicate), and it corrects three secondary claims relative to the contaminated collection: an Opus contested-item lead that is real and capability-graded rather than absent (§4.2), a weakest-tier calibration that is mildly anti-calibrated rather than flat (§4.4), and a forge-down super-caricature that spares a capable neutral rather than defeating every judge (§4.3). The structurally-clean raw-API arm (§3.7) was unaffected and was not re-run.

The forgery Opus self-edge is formally open. In the length-matched forgery paradigm (B4; Table 1), the genuine author (Opus) retains a small edge (+0.188; §4.3) over the only available clean neutral, which we read as capability rather than self-familiarity: it does not exceed Opus's generic third-person discrimination advantage over that weaker neutral, and the preregistered matched control (B3) is itself confounded by which tier's genuine passage must be identified (§4.3). What would settle it is a single comparator we did not have — a *within-family, capability-matched non-author*, which holds capability and family fixed and varies only authorship. A residual Opus edge over a non-author Claude of capability $\geq$ Opus could not be charged to capability or to family membership, leaving authorship as the one remaining difference. A cross-family judge — a frontier GPT or Gemini tier — cannot stand in: it differs from Opus on the family axis as well as authorship, so any edge over it could be the family-identification signal our within-family design exists to remove (Choi et al., 2025), and its discrimination of Claude-tier prose is out-of-distribution rather than capability-matched. Within the Claude family the only model of capability $\geq$ Opus is Fable 5, which was not available during the study. The question therefore remains open for want of the one qualifying judge, not for want of analysis.

Single generation family. All generators and judges are current Claude 4 tiers. Each tier ran at the release current in the harness as of the June 13–14, 2026 runs: Opus 4.8, Sonnet 4.6, and Haiku 4.5. Because the harness exposes only the current release of each tier, these were the versions simultaneously available rather than a chosen or varied factor, and we draw no inference from the differing point-release numbers. We make no cross-generation or cross-family claim.

The harness is itself a variable. Generation ran inside the Claude Code agent harness, which we now show is not a neutral instrument: the fully-crossed 2×2 (§4.1) locates tier-stylistic amplification

on the text-source axis, not the judge axis, so the deployed-persona conditioning amplifies tier signal at generation. This is a finding, but it also means the harness is a confound for any absolute estimate of tier distinctiveness; the raw-API replication (§4.2) is what licenses the "not a harness artifact" reading for the central null.

Non-pinnable served versions. In the harness, the served models (`claude-opus-4-8`, `claude-sonnet-4-6`, `claude-haiku-4-5`) are not version-pinnable, so exact reproduction depends on the deployed checkpoints at run time. The raw-API arm is more reproducible only in part: it bypassed the harness and pinned Haiku to a dated snapshot (`claude-haiku-4-5-20251001`), but its Opus and Sonnet ids are the same undated 4.x aliases as the harness and can likewise drift.

7 Conclusion

We test whether sibling Claude 4 tiers can behaviorally recognize their own prose, and find a four-part conjunction. First, within-family tier style is externally legible and capability-graded: neutral in-harness judges discriminate authorship at Opus 0.91 ($d' = +1.91$) $\approx$ Sonnet 0.90 ($d' = +1.81$) $>$ Haiku 0.63 ($d' = +0.45$). Second, being the author confers no privileged access — the self-minus-neutral advantage is negative or brackets zero for every tier under both confound controls (pair-fixed pooled -0.134 $[-0.160, -0.106]$), hardens on disagreement items (P(self correct) $= 0.243$), and replicates on the raw API and across explicit and implicit attribution framings. Third, attribution runs on stylistic stereotype: a stronger model's length-matched caricature of a weaker tier is judged more authentic than the genuine article, even by the author. Fourth, identity-judgment calibration is a gradient that tracks capability (Opus $+0.091$ $>$ Sonnet $+0.030$ $>$ Haiku -0.012), with the weakest tier mildly anti-calibrated.

Together these place a behavioral boundary on emergent-introspection claims (Lindsey, 2026): privileged access, where real, does not extend to prose self-authorship among tiers. Settling the open Opus forgery edge needs a capability-matched, same-family, non-author neutral — the one comparator that holds capability and family fixed and varies only authorship. The natural next step is to pair our behavioral half with an internal logprob measurement of the implicit, likelihood-level half that prior work reports but we do not measure here.

Data and Preregistration

Preregistration. Prospective branches were preregistered on the Open Science Framework before data collection: the framing branch (B2) and the primary forgery round (B4) at `osf.io/brdt8` (registered June 8, 2026, pre-data); the matched forgery control (B3) at `osf.io/5azq8` (registered June 12, 2026, pre-data). The capability-matched within-Claude neutral arm (Claude Fable 5, used as the framing-condition resolver "F1") was preregistered but not run, as Fable was unavailable during the study; its staged scripts are included in the repository for reproducibility. We report fidelity to these registrations honestly, including where a rule under- or over-fired. The framing branch's decision rule applied cleanly on the confirm-null arm: F1/F2 $\approx$ F0 with no positive tier self-advantage under any framing, so the null was declared framing-robust as registered. The primary forgery round ran exactly as registered, but its binary "is there a self-edge" decision rule did not

anticipate the caricature asymmetry that drives the forge-down result — the rule was sound but under-specified for what the data revealed. The matched forgery control's registered primary contrast came in at +0.206 [+0.100, +0.306]; unlike the pilot, the registered target-distinctiveness caveat did *not* fire (Opus-neutral 0.519, at chance), so the contrast is interpretable as registered, though we read the residual as capability-confounded rather than self-familiarity (§4.3). The exploratory pilot rounds (Rounds 1–5b) are labeled as such; the main rounds (B1) were run under internal frozen rules before external registration, and we do not present them as preregistered.

Data and reproducibility. A repository containing the per-round judgment records, generation and judging scripts, and analysis code accompanies this paper, available at `github.com/dbookstaber/claude-tier-self-recognition` and archived at `doi.org/10.5281/zenodo.20724958`. All harness numbers reported here are from the decontaminated re-run (§3.11): the repository holds the byte-faithful run specifications, the two independent clean-room passes' raw logs and their pooled analysis, the clean-room attestations for each pass, and the a-priori interview logs and coding (§3.12). Served models (`claude-opus-4-8`, `claude-sonnet-4-6`, `claude-haiku-4-5`) are not version-pinnable in the harness, so exact reproduction depends on the deployed checkpoints at run time; the raw-API arm bypassed the harness and pinned Haiku to a dated snapshot (`claude-haiku-4-5-20251001`), though its Opus and Sonnet ids remain undated aliases.

Use of AI tools. Separately from the models studied (which are the experimental subjects and instrument, §3.1 and Appendix A), the author used Claude Opus (Anthropic), via Claude Code, to draft and edit the prose of this manuscript about the author's own study design, data, and results, and to assist in writing analysis code. The author designed the study, directed and checked every analysis, reviewed and edited all text, and takes full responsibility for the content.

References

T. Ardoin, J. Schäfer, and G. Wunder. LLM Self-Recognition: Steering and Retrieving Activation Signatures. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026. arXiv:2606.06315.

T. Ashuach, S. Gretz, Y. Katz, Y. Belinkov, and L. Ein-Dor. Masked by Consensus: Disentangling Privileged Knowledge in LLM Correctness. In *Proceedings of the Association for Computational Linguistics (ACL), Main Conference*, 2026. arXiv:2604.12373.

Asvin G. The Assistant as a Privileged Persona: A canonical reference in cross-persona self-recognition. *arXiv preprint arXiv:2606.00545*, 2026.

Asvin G. and J. Lindsey. From Simulation to Enaction: Post-trained language models recognize and react to their own generations. *arXiv preprint arXiv:2605.25459*, 2026.

X. Bai, A. Shrivastava, A. Holtzman, and C. Tan. Know Thyself? On the Incapability and Implications of AI Self-Recognition. *arXiv preprint arXiv:2510.03399*, 2025.

F. J. Binder, J. Chua, T. Korbak, H. Sleight, J. Hughes, R. Long, E. Perez, M. Turpin, and O. Evans. Looking Inward: Language Models Can Learn About Themselves by Introspection. *arXiv preprint arXiv:2410.13787*, 2024.

J. L. Brandl. The puzzle of mirror self-recognition. *Phenomenology and the Cognitive Sciences*, 2016. doi: 10.1007/s11097-016-9486-7.

M. Brennan, S. Afroz, and R. Greenstadt. Adversarial Stylometry: Circumventing Authorship Recognition to Preserve Privacy and Anonymity. *ACM Transactions on Information and System Security (TISSEC)*, 15(3), 2012. doi: 10.1145/2382448.2382450. Article 12.

G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

M. Cheng, T. Piccardi, and D. Yang. CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations. In *Proceedings of EMNLP (Main Conference)*, 2023. arXiv:2310.11501.

Y. Choi, C. Li, Y. Yang, and Z. Jin. Agent-to-Agent Theory of Mind: Testing Interlocutor Awareness among Large Language Models. In *Proceedings of EMNLP*, 2025. arXiv:2506.22957.

T. R. Davidson, V. Surkov, V. Veselovsky, G. Russo, R. West, and C. Gulcehre. Self-Recognition in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP*, 2024. arXiv:2407.06946.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

M. J. Hautus. Corrections for extreme proportions and their biasing effects on estimated values of d'. *Behavior Research Methods, Instruments, & Computers*, 27(1):46–51, 1995.

S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Hatfield-Dodds, N. DasSarma, E. Tran-Johnson, S. Johnston, S. El-Showk, A. Jones, N. Elhage, T. Hume, A. Chen, Y. Bai, S. Bowman, S. Fort, D. Ganguli, D. Hernandez, J. Jacobson, J. Kernion, S. Kravec, L. Lovitt, K. Ndousse, C. Olsson, S. Ringer, D. Amodei, T. Brown, J. Clark, N. Joseph, B. Mann, S. McCandlish, C. Olah, and J. Kaplan. Language Models (Mostly) Know What They Know. *arXiv preprint arXiv:2207.05221*, 2022.

D. Khullar, J. Hopkins, R. Wang, and F. Roger. Self-Attribution Bias: When AI Monitors Go Easy on Themselves. *arXiv preprint arXiv:2603.04582*, 2026.

M. Koppel, J. Schler, and S. Argamon. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26, 2009.

H. Lederman and K. Mahowald. Emergent Introspection in AI is Content-Agnostic. *arXiv preprint arXiv:2603.05414*, 2026.

K. Lee, G. Byatt, and G. Rhodes. Caricature effects, distinctiveness, and identification: Testing the face-space framework. *Psychological Science*, 11(5):379–385, 2000. doi: 10.1111/1467-9280.00274.

S. A. Lehr, M. Cipperman, and M. R. Banaji. Extreme Self-Preference in Language Models. *arXiv preprint arXiv:2509.26464*, 2025.

J. Lindsey. Emergent Introspective Awareness in Large Language Models. Anthropic / Transformer Circuits. arXiv:2601.01828, 2026.

C. Lu, J. Gallagher, J. Michala, K. Fish, and J. Lindsey. The Assistant Axis: Situating and Stabilizing the Default Persona of Language Models. *arXiv preprint arXiv:2601.10387*, 2026.

U. Macar, L. Yang, A. Wang, P. Wallich, E. Ameisen, and J. Lindsey. Mechanisms of Introspective Awareness. *arXiv preprint arXiv:2603.21396*, 2026.

N. A. Macmillan and C. D. Creelman. *Detection Theory: A User's Guide*. Lawrence Erlbaum Associates, Mahwah, NJ, 2nd edition, 2005.

A. Naphade, S. Bhargav, S. Lim, and M. Shah. Me, Myself, and π: Evaluating and Explaining LLM Introspection. In *ICLR Workshop*, 2026. arXiv:2603.20276.

A. Panickssery, S. R. Bowman, and S. Feng. LLM Evaluators Recognize and Favor Their Own Generations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2404.13076.

T. Pearson-Vogel, M. Vanek, R. Douglas, and J. Kulveit. Latent Introspection: Models Can Detect Prior Concept Injections. *arXiv preprint arXiv:2602.20031*, 2026.

V. S. Ramachandran and W. Hirstein. The Science of Art: A Neurological Theory of Aesthetic Experience. *Journal of Consciousness Studies*, 6(6–7):15–51, 1999.

G. Rhodes, S. Brennan, and S. Carey. Identification and ratings of caricatures: Implications for mental representations of faces. *Cognitive Psychology*, 19(4):473–497, 1987.

S. Singh, T. Linzen, and S. Ravfogel. Can LLMs Introspect? A Reality Check. *arXiv preprint arXiv:2605.26242*, 2026.

S. Song, H. Lederman, J. Hu, and K. Mahowald. Privileged Self-Access Matters for Introspection in AI. *arXiv preprint arXiv:2508.14802*, 2025.

E. Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009.

P. Wang, L. Li, L. Chen, Z. Cai, D. Zhu, B. Lin, Y. Cao, Q. Liu, T. Liu, and Z. Sui. Large Language Models are not Fair Evaluators. *arXiv preprint arXiv:2305.17926*, 2023.

Z. Wang, N. I. Tripto, S. Park, Z. Li, and J. Zhou. Catch Me If You Can? Not Yet: LLMs Still Struggle to Imitate the Implicit Writing Styles of Everyday Authors. In *Findings of the Association for Computational Linguistics: EMNLP*, 2025. arXiv:2509.14543.

K. Wataoka, T. Takahashi, and R. Ri. Self-Preference Bias in LLM-as-a-Judge. In *NeurIPS Safe Generative AI Workshop*, 2024. arXiv:2410.21819.

L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*, 2023. NeurIPS 2023 Datasets and Benchmarks.

Y. Zhou, W. Zhu, J. Wen, W. Peng, Z. Wu, and Y. Xue. From Implicit to Explicit: Enhancing Self-Recognition in Large Language Models. *arXiv preprint arXiv:2508.14408*, 2025.

A The Agent Harness: Instrument and Variable

Every model in this study runs as a Claude Code subagent — an instance of Anthropic's agentic coding assistant, spawned through the orchestration tool with a specified model id (`opus`, `sonnet`, `haiku`, each resolving to the current Claude 4.x release). (The exception is the raw-API replication of §3.7, which bypasses the harness entirely.) The subagent inherits Claude Code's system prompt, a coding-assistant persona, and its tool scaffolding (file-system and shell access). We use it as a stand-in for the `/v1/messages` API: generators are instructed to return plain prose with no tool use and no preamble, judges return structured forced-choice verdicts, and each returned text carries a ground-truth tier label by construction.

What we control is narrow — the model id and the task prompt. What we do not control, and mostly cannot as external users, is most of the rest. We neither author nor fully observe the system prompt; we cannot set temperature (the harness uses its default, and the API confirms Opus 4.8 rejects the parameter outright); and we receive neither token logprobs nor native token counts, so where length must be held fixed we truncate by a word-count proxy. Every subagent therefore speaks from inside the deployed "Assistant" persona — the same default-persona frame that recent work casts as a privileged reference persona (Asvin G., 2026) and localizes as an "Assistant Axis" in activation space (Lu et al., 2026). It is not a frame we could strip.

We checked whether that frame leaks into the data. It does, rarely and visibly: in free generation the weakest tier (Haiku) occasionally emitted harness context — local file paths and the like — while the stronger tiers did not. Excluding the leaked items leaves the central results unchanged.

The harness is not only the instrument; it is one of the findings. The fully-crossed 2×2 of §4.1 shows discriminability tracks the text source rather than the judge — the harness makes models *write* more tier-distinctively, a generation-side effect. We therefore treat it as a studied factor rather than a neutral pipe, and the raw-API replication of §4.2 is what licenses the central null independently of it. Two unknowns bound the external validity of our absolute magnitudes: the exact system-prompt contents (and whether they differ by tier), and the mechanism of the amplification, which we observe only behaviorally. The within-design contrasts are insulated from both — they hold the harness fixed across the self and neutral conditions — but the absolute legibility numbers are not, and should be read as harness-conditional.