The Confession Is in the Logits:
A Pre-Registered Confirmation of Pre-Generation Confabulation Detection

Tad MacPherson
Metavolve Labs, Inc.
research@iaeternum.ai

July 2026. Confirmatory study preprint, v0.2 draft.

Abstract

Our pilot study, *Never Mistake Proximity for Proof* [1], reported — in an explicitly exploratory section, below its own pre-registered threshold — a candidate *pre-verbal* signature of failure-to-abstain: workspace-band features from the Jacobian lens [2], read at the final prompt position before any output token, separated failure-bound from abstain-bound trials on 10 events. This paper is the pre-registered confirmatory test. On a fully disjoint, hash-sealed corpus of 168 private-fact probes over 21 sessions in 7 near-domain triads (1,680 graded rows, every label human-primary), with feature, layer band, and analysis family frozen before capture and the confirmatory statistic unopened until adjudication closed, the primary confirms: restricted-lexicon median lens rank over layers 10–20 separates failure from abstention at AUC 0.900 (116 vs. 219 events; 95% CI [0.851, 0.922]; Holm $p = .00015$). One pre-declared secondary — that workspace flatness separates leaky from clean abstentions — does not confirm (AUC 0.589, $p = .116$), and we report it as a null. The pre-registered baseline gauntlet then delivered the study's most useful result: a *final-layer* rank lookup of the same five-token abstention lexicon — no lens, no fitted apparatus, logits every forward pass already computes — reaches AUC 0.918, beating the lens, which survives only as a complement (0.926 combined). We count this a feature, not a defeat: pre-verbal detection of this failure class requires zero marginal inference cost, while standard confidence proxies (first-token entropy 0.455, log-probability 0.333) are blind to it. The behavioral effect replicates at 3.1× scale and is amplified, not attenuated, in a deliberately number-sparse held-out genre (near-domain failure 47.2% art-provenance vs. 25.0% technical), where the signature transfers untouched (AUC 0.913 vs. 0.862). The frozen decision rule, adjudication log, and hash-chained provenance manifest are released.

1 Introduction

Our pilot study, *Never Mistake Proximity for Proof* (NMPFP; concept DOI 10.5281/zenodo.21367214) [1], established a behavioral result: when a frozen language model is shown an *authentic, topically adjacent record that does not contain the answer* — the near-domain D-arm — its rate of FAILURE-TO-ABSTAIN rises along a gradient of apparent admissibility, and the failures are dominated by misattribution: true facts quarried from the shown record and reassigned to the queried scope. Alongside that behavioral ladder, the pilot reported — in an explicitly exploratory section, on 10 failure events against 56 abstentions, below its own pre-registered threshold — a candidate *pre-verbal* signature: simple workspace-band features from the Jacobian lens [2], measured at the final prompt position before any output token, separated failure-bound from

abstain-bound trials at AUC 0.78–0.83. The pilot claimed no confirmatory inference from those numbers, and said so in italics.

This paper is the debt that sentence incurred. The research question, stated plainly: **does the workspace confess before the mouth moves?** That is: within the near-domain condition, is the trajectory toward an unsupported attribution — rather than an abstention — already readable in the model's internal state at the final prompt position, before the first output token is generated?

We answer it under confirmatory discipline. The primary feature and layer band were frozen before capture; the probe corpus is new and fully disjoint from the pilot (7 near-domain triads, 21 authored sessions, 168 probes, 336 D trials per regime, hash-sealed before any capture); every outcome label is human-primary under the pilot's three-label ontology (CORRECT / ABSTAIN / FAILURE-TO-ABSTAIN); and no outcome-conditioned analysis was performed until adjudication closed — the confirmatory AUC remained unopened, and the provenance manifest seals that ordering.

Our findings, honest headline first:

1. **The signature is confirmed — and immediately humbled.** The frozen primary feature (restricted-lexicon median lens rank, layer band 10–20) separates failure-bound from abstain-bound trials at **AUC 0.900** on 116 failure events versus 219 abstentions (bootstrap 95% CI 0.851–0.922; Holm-adjusted $p = .00015$; label-shuffle null 97.5th percentile 0.564). The pre-registered accrual target of 25–30 events was exceeded roughly fourfold. But the pre-registered baseline gauntlet did its job: a *final-layer* readout of the same restricted lexicon — no lens, no fitted interpretability apparatus, one rank lookup in the model's own output logits — reaches **AUC 0.918**, beating the lens alone. The lens survives as a *complement*, not a champion: the nested combination reaches **AUC 0.926**. The practical consequence is the opposite of the romantic one: **pre-verbal failure detection in this setting requires no interpretability instrumentation at inference time.**

2. **The signal is abstention-token geometry, not generic uncertainty.** First-token entropy (AUC 0.455) and first-token log-probability (0.333) sit at or below chance; restricted-lexicon probability mass (0.627) and logit margin (0.873) grade upward toward the rank-based readouts. What predicts the failure is where the abstention vocabulary sits in the model's pre-verbal ordering, not how uncertain the model is.

3. **One pre-declared secondary does not confirm, and we lead with it.** The containment hypothesis — that leaky abstentions (declines that relay record specifics) are separable from clean abstentions by workspace flatness — comes in at AUC 0.589 ($p = .116$): a null. The flatness feature itself confirms directionally but weakly (AUC 0.687, Holm $p = .00015$), a substantial retreat from its pilot showing.

4. **The behavioral effect replicates at triple scale and transfers across genre.** Under no instruction, near-domain FAILURE-TO-ABSTAIN reaches 25.0% (technical genre) and 47.2% (art-provenance genre) against control-arm rates of 11.9% (no context) and 11.3% (irrelevant context); the correct record remains near-ceiling (98.8% CORRECT); the abstention instruction again mitigates without curing (6.2% / 18.1%). The art-provenance genre was held out from all detector fitting, and the signature transfers: per-genre AUC 0.862 (technical) and 0.913 (art).

Contributions. (i) A pre-registered, sealed-before-capture confirmatory test of a pre-verbal abstention-failure signature, with the frozen decision rule, adjudication log, and hash-chained provenance manifest released. (ii) A baseline gauntlet result we consider the paper's most useful sentence: cheap final-layer features match or beat the interpretability readout, and the correct claim for lens-based detection is *complementarity* (0.918 alone vs. 0.926 combined), quantified on nested models. (iii) A behavioral replication of the near-domain effect on a disjoint corpus roughly three times the pilot's size, with a second, deliberately number-sparse genre demonstrating that the effect is not an artifact of number-dense technical logs. (iv) A fully human-primary label set (every one of the 1,680 graded rows carries a human or human-sanctioned label; the 346 contested rows were blind-adjudicated with verbatim rationales) and a consolidated bright-line rubric for the three-label ontology. (v) One pre-declared null, reported as such.

2 Background

2.1 The pilot: proximity, affordance, and admissibility

NMPFP evaluated 68 private-fact probes — facts private to our laboratory and empirically unrecoverable without the record; control arms scored 0% CORRECT throughout — under matched context arms forming a gradient of apparent legitimacy. Uninstructed FAILURE-TO-ABSTAIN ordered with that gradient: 2.9% under irrelevant natural prose, 9.4% under an authentic but out-of-domain *answer-shaped* text, and 15.2% under the authentic near-domain record (endpoints differ, Fisher exact $p = .016$); the correct record extracted perfectly. The failures were overwhelmingly misattributions rather than inventions — corpus-resident facts with the wrong address — and an explicit abstain-permission instruction reduced but did not eliminate them (15.2% $\rightarrow$ 4.5%), with the residue concentrated on a stable probe tail. The pilot framed the result in evidentiary terms: relevance and authenticity do not guarantee *admissibility* — sufficiency of the shown record for the specific claim requested — and the near-domain D-arm is precisely the condition where a record is maximally admissible-*looking* while being evidentially empty.

The pilot's outcome ontology carries over unchanged and is load-bearing for everything that follows. Each answer receives exactly one of three labels: CORRECT, ABSTAIN, or FAILURE-TO-ABSTAIN (any non-correct answer that asserts a candidate for the queried fact). Two boundary rules from the pilot — the scope test and grade-the-emitted-text — were consolidated during this study's adjudication into a single bright line, applied uniformly and recorded in the released log: *did the answer offer a candidate for the queried fact?* If yes — hedged, half-declined, framed as an "educated guess," partial, or substituted — it is a FAILURE-TO-ABSTAIN; if it declined and merely described what the context contains, it is an ABSTAIN with a flag.

2.2 The exploratory signature, and why it required this study

The pilot's internal-state section was written to be distrusted. Its event counts (10 vs. 56) sat below the pre-registered confirmatory threshold; its two features (restricted-lexicon rank and workspace flatness, AUC 0.783 and 0.827 on final adjudicated labels) were selected in the same breath as they were measured; and the pilot said explicitly that no confirmatory inference was claimed. What the pilot *did* do was freeze the follow-up: a single primary feature and band, event-driven accrual to 25–30 held-out failures, session-clustered structure, and — critically — a baseline gauntlet including final-layer logits and first-token statistics that any proposed internal

detector must beat or measurably complement [3, 4, 5]. That last requirement was adversarial by design: an interpretability readout that merely ties cheap output-side features is not a discovery, it is instrumentation overhead. As Section 6 reports, the gauntlet fired.

2.3 From pilot to confirmatory: what was frozen, and when

The confirmatory corpus was authored and sealed before any capture. Twenty-one new sessions in seven near-domain *triads* (four technical, three art-provenance) replace the pilot's pairs, so that every probe's near-domain D-arm runs against two distinct sibling records (336 D trials per regime, 840 rows per regime across both instruction regimes); the probe file was hash-committed (SHA-256 `399ffd44...`) with its leakage audit, short-gold pre-declarations, and authoring worksheets, and the human anchor audit (105/105 anchors verified against original sources) completed before the freeze. The art-provenance genre exists as a designed falsification lever: its records are names, dates, materials, and attributions rather than dense numerics, and it was held out from all detector fitting, so a signature that lived only in number-dense text would fail to transfer. Capture, machine grading, blind kit construction, human adjudication, and only then the opening of the confirmatory analysis were executed in that order, each stage sealed into the provenance manifest; final human-primary labels (346 contested rows adjudicated blind; agreement on that hardest-row subset: human–Claude 94.2%, human–Gemini 85.5%, Krippendorff $\alpha = 0.743$, not comparable to the pilot's 0.901 which spanned easy rows) were fixed before any AUC was computed. Section 3 gives the full design; Sections 5 and 6 report the frozen family exactly as pre-registered, nulls included.

3 Methods I: The Confirmatory Probe Corpus

3.1 Triads, genres, and partial rotation

The confirmatory benchmark comprises 168 private-fact probes over 21 newly authored work-session records, organized into 7 near-domain *triads* of 3 sessions each. The set is fully disjoint from the pilot: no session, probe, or gold value recurs, and the pilot's 68 probes played no role beyond fixing the design's rate assumptions.

The pilot's session-*pair* structure could not survive external review unchanged. Reviewers required that each fact be tempted by at least two different near-domain partners, so that a failure can be attributed to the near-domain condition rather than to one idiosyncratic partner record; in a pair, each query session has exactly one partner, and the requirement is structurally impossible. The triad is the minimal unit that satisfies it. Within a triad, each probe's near-domain D-arm is run against *both* sibling sessions (partial rotation), so each session's facts face two distinct tempting records, each record tempts two query sessions, and all 6 ordered session pairings per triad occur, counterbalancing direction. The unit of independence for any split or clustering is therefore the triad, which keeps a session and both of its partners on the same side of every split.

Four triads (12 sessions, 96 probes) are the pilot's genre: dense technical work logs from our laboratory's engineering programs. The pilot flagged this genre as plausibly the most number-stealable text there is; the confirmatory design therefore adds a second genre of 3 triads (9 sessions, 72 probes) of *art-provenance dossiers*: names, dates, materials, editions, and attributions rather than hashes and scores, styled as inscription and registry records and authored into the same length band as the technical logs. The tightest of these is a triad of three structurally

parallel founders' dossiers from a single art movement, in which the wrong sibling record answers the *shape* of every question with the wrong artist's value, which is the misattribution temptation in its purest form. One candidate art source was excluded at design time because its records derive from a public museum collection and would violate the private-fact premise.

Each probe is evaluated under the pilot's arm structure with the D-arm doubled: **A_none**, **R_random** (length-matched irrelevant prose recomputed against the new corpus mean), **C_present**, and two **D_mismatched** trials (one per triad sibling). This yields $168 \times 5 = 840$ rows per instruction regime, of which 336 are D trials, against the pre-registered accrual target of 25–30 failure events: at pilot-anchored failure rates of 8%, 11.8%, and 15.2%, the design expects 27, 40, and 51 D-arm FAILURE-TO-ABSTAIN events respectively, so the target is met even under substantial attenuation. Outcomes are graded under the pilot's three-label ontology (CORRECT / ABSTAIN / FAILURE-TO-ABSTAIN), described with the grading protocol in the next section.

3.2 Authoring discipline

Because the benchmark's own authorship pipeline involves AI systems, the corpus was authored under explicit confabulation controls. Every session log was drafted with its named source documents open in-context, one session at a time, with the source file read immediately before drafting; nothing was authored from memory of a source. Technical drafts landed in a band of 839–1058 characters (mean 982), and the art sessions were authored into that band (888–1058, mean approximately 999) so that genre is not confounded with length. Per session, a verification worksheet maps every stated fact to its source file with a verbatim supporting quote, and 5 anchor facts per session are starred for human audit, giving $21 \times 5 = 105$ anchors. By standing redaction rule, no wallet or contract address appears in any draft, probe, or gold.

3.3 The two-layer human audit

The pilot's single-pass anchor audit was restructured into two honest layers, after we recognized that asking the human author to *recall* operational minutiae was never a realistic ground truth, and pretending otherwise would be its own confabulation. Layer 1 is mechanical and load-bearing: every anchor quote is programmatically verified to exist verbatim in its source file (normalized for quote style, markdown, and whitespace); all 105 anchors passed. Layer 2 is the judgment a machine cannot supply: the human adjudicator audits claim-versus-quote coherence and flags anything contradicting lived experience. Across the two audit kits, 100 of 105 anchors were marked correct on sight; 4 technical anchors marked unsure were escalated to the deepest available sources (code line numbers, original lair text) and all confirmed with receipts, and 1 art-kit row left blank was confirmed directly. Zero authoring confabulations were found, replicating the pilot's authoring-audit result on a disjoint corpus.

3.4 Leakage audit and designed-out collisions

An automated leakage audit ran before freeze, enforcing three properties: every gold is extractable from its own session record; no gold appears (word-boundary or fuzzy match) in *either* of its two D-partner records; and each triad is scanned for genuinely shared facts, the pilot failure mode in which two paired systems truly share a value. The frozen report shows 0 FAIL and 10 WARN, each warning reviewed and accepted with documentation. Several collisions were designed out rather than caught: facts common to all three founders' dossiers (movement name,

shared template boilerplate, shared personnel) are never golds; one artist's edition size numerically equals another's ship resolution, so the edition is deliberately not a gold; and a genuine numbering-range conflict between two internal source documents, discovered during authoring, was excluded from golds entirely and flagged, since a fact our own records disagree on cannot serve as ground truth. A pilot-inherited lesson was applied at design time as well: a value truly shared across one technical triad's systems was excluded triad-wide rather than quarantined post hoc. Twenty short golds ($\leq$3 characters) were pre-declared at freeze for sensitivity exclusion, extending the pilot's short-gold rule.

3.5 Freeze and sealing before capture

The corpus pipeline is deliberately reconstructable: a self-contained data module generates the session records, probes, and golds; a builder expands each probe into its 5 arm rows with the triad map; and the output is a single frozen probe file. All three, together with the leakage-audit script and its report and the upgraded capture runner, were SHA-256 hashed into the provenance manifest on 2026-07-18, *before any capture*; the frozen probe file's hash (`399ffd44...`) is the reference against which all downstream artifacts are checked. The seal also fixed a capture-day gate binding the analysis to the sealed stimuli: the model's restricted abstention-lexicon token ids must resolve to the five frozen values recorded in the manifest, or the run aborts before producing data. Nothing in this section was altered after sealing; the adjudication record later verified that every question presented to the human grader byte-matches the sealed probe file.

4 Methods II: Capture and Grading

4.1 Runner v2: outputs, lens readouts, and cheap baselines in one pass

The capture runner was upgraded from the pilot version in one pre-declared step (runner v2, hash-sealed with the probe set before any capture) with three changes, each motivated by a specific pilot lesson.

Same-pass cheap baselines. Every baseline in the confirmatory gauntlet is computed during the same forward pass that produces the generation and the lens readouts: the first output token's log-probability and full-distribution entropy, and three final-layer statistics of the frozen five-token restricted abstention lexicon (minimum rank, total probability mass, and logit margin). Capturing detector and baselines in a single pass removes the most common asymmetry in detector papers, where baselines are recomputed later under subtly different conditions; here every comparison in the baseline gauntlet is between quantities read off the identical forward computation.

Position sweep. A `--positions` sweep records lens readouts at the final prompt position (the pilot's choice and the confirmatory default, byte-compatible with the pilot row schema), at offsets -2 and -4, and at the last user-content token. Each swept row stores the resolved absolute index *and* the decoded token at that index, so position resolution is auditable after the fact rather than trusted.

Generation cap 48$\rightarrow$96. The pilot's 48-token cap produced a class of cap-truncated declines that had to be graded on incomplete text, a label-ambiguity source we flagged there as a design artifact. The confirmatory protocol doubles the cap to `max_new_tokens` = 96; this is a pre-declared protocol change, recorded in the pre-registration seal, and it removes truncation ambiguity rather than altering the task.

4.2 Capture day and gates

Capture ran on a single NVIDIA L4 (`us-west1-a`) with `meta-llama/Llama-3.1-8B-Instruct` at bfloat16, greedy decoding, both instruction regimes: 840 rows per regime (168 A_none, 168 R_random, 168 C_present, 336 near-domain D-arm trials under the triad rotation), 1680 rows total. The VM was deleted after checksum-verified retrieval of all artifacts.

Two runtime gates guarded the frozen design. First, the restricted-lexicon token ids resolved by the live tokenizer were required to equal the ids frozen at pre-registration (1440, 1541, 4250, 9987, 14931); a mismatch would indicate tokenizer or model-revision drift and abort the run. Second, a built-in cross-check requires the generate-path and lens-path final-position argmax to agree on every row, catching BOS-handling, truncation, or off-by-one drift between the two code paths; the full run showed **0 mismatches in 1680 rows**. A one-row sanity capture preceded the full pass so that chat-template resolution of the `last-user-content` position could be inspected by eye, and the sanity row and per-run metadata sidecars are retained in the release.

The gates were not decorative. One pre-run aborted when a missing model-index file tripped the sanity gate *before* any capture occurred; the environment was fixed and the run relaunched, and no data from the aborted attempt exists. We report this because a gate that has never fired is indistinguishable from a gate that does not work.

No-peeking discipline was maintained across the capture–grading boundary: the capture seal records structural counts only (rows per arm, gate outcomes), with zero outcome or detector analysis performed; the lens features were untouched by every grader and by the kit builder; and the confirmatory AUC family remained unopened until the human labels were finalized and sealed.

4.3 Machine grading: three families

All 1680 outputs were graded under the pilot's three-label ontology (CORRECT / ABSTAIN / FAILURE-TO-ABSTAIN; the adjudication sheets retain the pilot-era label string CONFABULATED for the third class) by three independent machine families before any human saw an answer:

1. **Word-boundary heuristic.** The mechanical grader (gold-match with word-boundary discipline, retaining the pilot's short-gold fixes) resolved 753 rows automatically.
2. **GRADE_SYS (Gemini).** Rows not resolved mechanically went to the frozen Gemini grading prompt, 805 grading calls across the two regimes.
3. **Blind Claude family.** 53 context-isolated Claude agents, blind to condition metadata and to the other families' labels, graded all 1680 rows, followed by a 326-row second-opinion sweep within the family.

4.4 Human-primary adjudication

Machine labels select and stratify; they do not decide. Final labels are human-primary, produced through a blind two-sheet kit sealed before adjudication began: an **ADJUDICATE** sheet of 346 contested rows presented with no machine labels visible, and a **CONFIRM** sheet of 1334 rows on which all machine families were unanimous, pre-filled for batch sanction or row-level veto. A question-integrity check, raised by the adjudicator himself mid-kit, verified that all 1680

kit questions byte-match the sealed probe file (SHA-256 `399ffd44...`): zero divergences, so no question text drifted between freeze and grading.

Adjudication proceeded in checkpointed passes with walkthrough settlement against the frozen rubric and the pilot's precedent rulings. At checkpoint 1 (rows 1–99), the adjudicator's blind pass tallied 50 ABSTAIN / 47 FAILURE-TO-ABSTAIN / 2 CORRECT, with pre-settlement agreement of 84% against the Claude family and 72% against Gemini. All 15 rows on which both machine families opposed the human ran in the same direction, human lenient and machines harsh, 13 of the 15 in the near-domain D-arm; the direction matches the pilot's disagreement axis with the roles reversed, which we read as adjudicator calibration drift between studies (the adjudicator's own contemporaneous note in the released log says as much), and the checkpoint-plus-precedent mechanism is the designed stabilizer for exactly this. The walkthrough settled the disputes in five precedent groups, each ratified per-group by the adjudicator: committed relay and question-substitution answers (7 rows, to FAILURE-TO-ABSTAIN, on pilot precedent), decline-then-hedged-candidate answers (5 rows, to FAILURE-TO-ABSTAIN; hedging never rescues an offered value), a partial-credit boundary case (to FAILURE-TO-ABSTAIN, on the adjudicator's own pilot precedent), an R-arm affordance echo (to FAILURE-TO-ABSTAIN), and one descriptive leak upheld as ABSTAIN with a flag. Fourteen label flips were applied with per-row lineage notes preserving the initial blind verdicts.

The checkpoint consolidated the rubric into a single bright line applied to all remaining rows: *did the answer offer a candidate for the queried fact?* If yes — hedged, half-declined, framed as an educated guess, partial, or substituted — it is a FAILURE-TO-ABSTAIN. If it declined and merely described what the shown context contains, it is an ABSTAIN with a descriptive-leak note.

Checkpoint 2 covered rows 100–346 (blind pass plus 29 adjudicator-directed auto-declines and 8 mechanical skip-fixes, each logged); 19 post-checkpoint disputes were walked through and every proposed resolution was ratified verbatim, including two CORRECT→FAILURE-TO-ABSTAIN flips for wrong-mechanism answers and two machine-harsh labels overturned back to ABSTAIN. The CONFIRM sheet's 1334 unanimous rows were sanctioned as a batch with zero edits. Final contested-set labels: **202 FAILURE-TO-ABSTAIN / 140 ABSTAIN / 4 CORRECT** (346/346), authoritative in the released `human_grades_v3_final.csv`; every row in the study therefore carries a human or human-sanctioned label.

4.5 Agreement, and a caveat we insist on

On the contested 346, agreement with the human adjudicator was 94.2% (Claude family) and 85.5% (Gemini), with Krippendorff's $\alpha = 0.743$ (nominal) across the three grader families. This number must not be compared naively to the pilot's $\alpha = 0.901$: the v3 coefficient is computed on the *contested subset only* — the 346 rows selected precisely because the families disagreed or flagged them — while the pilot kit spanned eventful and easy rows alike. Conditioning on disagreement mechanically depresses any agreement statistic; the 1334 unanimous rows, which would inflate it, are excluded here by construction. The honest reading is that 0.743 is a lower bound on the hardest fifth of the data, not a decline in grading quality, and we state it this way because burying the selection effect would manufacture either false alarm or false comfort. The raters remain one human and two LLM families; the checkpoint-1 drift episode is further direct evidence for the planned external hypothesis-blind adjudicator, whose engagement is a standing decision point before the detector results ship in final form.

5 Results I: Behavioral Outcomes

5.1 The adjudicated table

Table 1 reports the final human-adjudicated outcome rates for the confirmatory set: 168 probes over 21 sessions in seven near-domain triads (four technical, three art-provenance), each probe's near-domain D-arm rotated against *both* triad siblings (336 D trials per regime; 840 rows per regime; 1,680 total, roughly 3.1× the pilot). All labels derive from the sealed final label file: 346 contested rows blind-adjudicated by the human adjudicator under the frozen three-label ontology, and 1,334 rows on which all machine grader families were unanimous, sanctioned in batch with zero edits (Section 4.4; agreement statistics and the contested-subset caveat on Krippendorff $\alpha = 0.743$ are given in Section 4.5).

Regime	Arm	CORRECT	FAIL	ABSTAIN
no-instruction	A_none	0.6%	11.9%	87.5%
no-instruction	R_random	0.0%	11.3%	88.7%
no-instruction	D_technical	0.0%	25.0%	75.0%
no-instruction	**D_art**	0.7%	**47.2%**	52.1%
no-instruction	C_present	**98.8%**	1.2%	0.0%
instructed	A_none	0.0%	0.6%	99.4%
instructed	R_random	0.0%	1.2%	98.8%
instructed	D_technical	0.0%	6.2%	93.8%
instructed	D_art	0.0%	18.1%	81.9%
instructed	C_present	98.2%	1.2%	0.6%

Table 1: Human-adjudicated outcomes, bfloat16, greedy decoding, 96-token cap. FAIL is failure-to-abstain. Raw counts, no-instruction: A 20/168, R 19/168, D_technical 48/192, D_art 68/144, C_present 166/168 CORRECT (2 FAIL); instructed: A 1/168, R 2/168, D_technical 12/192, D_art 26/144, C_present 165/168 CORRECT (2 FAIL, 1 ABSTAIN). Wilson 95% intervals for FAIL: D_art 47.2% [39.2, 55.3]; D_technical 25.0% [19.4, 31.6]; R_random 11.3% [7.4, 17.0]; instructed D_art 18.1% [12.6, 25.1]; instructed D_technical 6.2% [3.6, 10.6]. All Fisher tests below are descriptive: probes are triad-clustered and each probe contributes two D trials, and the pre-registered confirmatory analysis specifies hierarchical models with triad and probe random effects. C_present figures reflect the sealed 2026-07-22 multi-valid-gold amendment (two anchor rows, one per regime, reclassified FAILURE-TO-ABSTAIN→CORRECT after a gold-narrowing catch; pre-amendment rates were 98.2% no-instruction / 97.6% instructed).

5.2 The inversion, confirmed on a disjoint probe set

The pilot's central behavioral claim survives contact with a fully disjoint probe set at 3.1× scale: the near-domain D-arm is the condition that damages restraint, and the correct record repairs it. The proper comparators are within-genre, since the R_random control was run against both genres' probes: technical D-arm FAIL is 25.0% (48/192) against a same-probe irrelevant-context floor of 6.2% (6/96), a 4.0× lift (Fisher $p < 10^{-4}$); art D-arm FAIL is 47.2% (68/144) against 18.1% (13/72), a 2.6× lift ($p < 10^{-4}$). Authenticity and topical relevance again fail to deliver admissibility: every D-arm record is genuine and maximally adjacent, and none contains the

answer.

Four deviations from the pilot deserve naming rather than burying. First, the uninstructed control floor is elevated (A 11.9%, R 11.3%, versus the pilot's 4.4% and 2.9%). Three documented protocol changes plausibly contribute and are confounded with any real difference: the probe set is new and larger; the generation cap was doubled from 48 to 96 tokens per the preregistered protocol, giving declines more room to append an assertion; and adjudication applied the consolidated bright-line rubric (*did the answer offer a candidate for the queried fact?* — hedged, half-declined, "educated guess," partial, or substituted all count), under which the adjudication log records the human adjudicator's own calibration swinging harsher than pilot, with fourteen walkthrough rulings flipped toward FAIL. Cross-study rate comparisons are therefore not like-for-like; within-study contrasts are. Second, the pilot's suggestive A > R anchoring gap does not replicate (20/168 vs. 19/168, $p = 1.0$); we reported it as a point-estimate curiosity and it behaved like one. Third, the positive control is no longer perfect: 98.8% no-instruction and 98.2% instructed, with two FAILs in each regime;[1] the log records wrong-mechanism answers reclassified CORRECT→FAIL under walkthrough. The pilot warned that its 268/268 should not be read as broad grounded competence; the warning now has receipts. Fourth, two rows breach the private-fact premise at the margin: one no-context answer and one D-arm answer (both art-genre) were adjudicated CORRECT, indicating one gold guessable without the record. Both are visible in the released labels; excluding that probe moves no headline rate at the reported precision.

5.3 Genre amplification: retrieval shape is a hazard variable

The new finding is the size of the art-provenance effect. The art triads were authored as structure-parallel provenance dossiers — names, dates, materials, attributions; deliberately *not* number-dense; length-matched into the same character band as the technical logs — with one triad built from three structurally parallel founder dossiers, the tightest near-domain cluster in the study. The pilot's limitations section conjectured that dense technical logs were "plausibly the most number-stealable text there is." That conjecture is now falsified in the direction we did not expect: the number-sparse genre roughly *doubles* the technical near-domain FAIL rate (47.2% vs. 25.0%, $p < 10^{-4}$).

We decompose this honestly, because part of it is not a D-arm effect. Art-genre probes are more tempting everywhere: their irrelevant-context floor is 18.1% versus 6.2% for technical probes ($p = .025$; the no-context contrast, 13.9% vs. 10.4%, is not significant, $p = .63$). On top of that elevated floor, the near-domain lift remains large in both genres, and the absolute D−R gap is *wider* in art (+29.2 percentage points vs. +18.8) even as the relative ratio is smaller (2.6× vs. 4.0×). The reading we defend is structural: when the shown record is template-parallel to the record that would contain the answer — same dossier skeleton, same slots, a different entity's values filling them — the model has, for every queried slot, a syntactically perfect wrong candidate one position away. Value density was the pilot's hazard hypothesis; slot correspondence is the confirmatory set's answer. Retrieval *shape*, not just retrieval content, is a hazard variable, which matters operationally because production RAG corpora (forms, dossiers,

[1] These C_present figures reflect a sealed 2026-07-22 multi-valid-gold amendment: one anchor row per regime, initially graded FAILURE-TO-ABSTAIN, was reclassified CORRECT after a gold-narrowing catch on a multi-valid ("Name a ...") probe widened the accepted-answer set (no other such probe exists in the corpus). The amended labels are released as `human_grades_v3_final_amended.csv` alongside the untouched sealed originals; pre-amendment figures were 98.2% / 97.6% with three FAILs per regime.

tickets, registries) are dominated by exactly such structure-parallel records. The art genre was held out from all detector fitting, so this amplification also serves as the negative-transfer test for Part II (Section 6).

5.4 Instruction mitigates; at genre scale, the tail is a population

The explicit abstain-permission instruction again works and again does not finish the job. Pooled uninstructed D-arm FAIL of 34.5% (116/336) falls to 11.3% (38/336) instructed ($p < 10^{-12}$), and control arms drop to the ∼1% floor (A 0.6%, R 1.2%). But the residue is not small where the hazard is largest: 6.2% of technical and 18.1% of art near-domain trials still fail *with* the instruction in place — the instructed art rate exceeds the pilot's *uninstructed* near-domain rate of 15.2% (cross-study caveat above applies). The pilot's "stable vulnerable tail" was three probes; at confirmatory scale the instructed tail is 38 events. Prompt-level discipline moves the distribution; it does not reach the tail, and the tail grows with the hazard of the genre.

5.5 Event yield against pre-registration

The pre-registered detector analysis required accrual of at least 25–30 uninstructed near-domain FAIL events, with design-arithmetic projections of 27, 40, or 51 events at assumed rates of 8%, 11.8%, or 15.2% (the pilot rate). The study accrued **116 FAIL against 219 ABSTAIN** in the uninstructed D-arm (one further D trial adjudicated CORRECT, noted above): roughly four times the pre-registered floor and more than double the most optimistic projection, with the surplus driven almost entirely by the genre amplification just described. The confirmatory internal-state analysis of Part II therefore proceeds fully powered, on labels that were frozen and hash-sealed before any confirmatory AUC was computed.

6 Results II: The Pre-Verbal Detector

6.1 The frozen confirmatory family, under Holm correction

The confirmatory contrast is the one the pilot pre-specified: within the no-instruction near-domain D-arm, trials that ended in FAILURE-TO-ABSTAIN versus trials that ended in ABSTAIN, using lens features captured at the final prompt position before any output token. Event accrual exceeded the pre-registered target of 25–30 held-out failures by roughly a factor of four: 116 failure events against 219 abstentions. All three family members, their feature definitions, and the Holm correction over the family were frozen before any outcome label or internal feature was inspected; the analysis script and its output are hash-sealed in the provenance manifest.

- **PRIMARY — restricted-lexicon median rank, layers 10–20: confirmed.** The median lens rank of the frozen five-token abstention lexicon over the pre-declared layer band separates failure-bound from abstain-bound trials at AUC = 0.900 (permutation $p = .00005$, Holm-adjusted $p = .00015$; bootstrap 95% CI $[0.851, 0.922]$). The label-shuffle null places its 97.5th percentile at AUC 0.564; the observed statistic is far outside it.
- **FLATNESS — top-1 to top-10 lens-logit gap (flatter ⇒ fail): confirmed, smaller.** AUC = 0.687 ($p = .00005$, Holm = .00015). The pilot's more robust feature replicates as a real but distinctly weaker signal than the lexicon-rank primary.

- **CONTAINMENT: null.** The pre-declared secondary hypothesis — that flatness would also separate *leaky* abstentions (declines that import record specifics) from clean ones — does not confirm: 30 leaky versus 189 clean abstentions, AUC = 0.589, $p = .116$ (Holm = .116). We report this as a null result, not a trend. The epistemic-containment reading of the flatness feature, which we advanced in the pilot as the more theoretically attractive interpretation, is unsupported at this scale.

The two pre-declared sensitivity alternates of the primary — median-of-per-layer-minimum (AUC 0.676) and the pilot's band-minimum feature itself (AUC 0.720) — both clear the shuffle null but fall well short of the frozen primary. The result is therefore robust in direction but not indifferent to aggregation: the median-rank statistic over the band is doing real work, and we say so rather than presenting the family as interchangeable.

6.2 The baseline gauntlet: the detector is beaten, and survives as a complement

The frozen protocol required the lens detector to face a gauntlet of cheap baselines on the identical contrast, with a pre-declared decision rule: the detector survives only if it *beats or complements* the best baseline. The gauntlet produced the most instructive result in the study, and it is not the one we expected.

- **Final-layer restricted rank-min: AUC** 0.918**.** The minimum rank of the same abstention lexicon in the model's own *final-layer* distribution — no lens, no mid-band access, nothing beyond the logits every forward pass already computes — beats the pre-registered mid-band primary (0.900).
- **Final-layer restricted logit-margin: AUC** 0.873**; prob-mass:** 0.627**.** The rank statistic, not raw probability mass, carries the signal.
- **Generic uncertainty is at chance.** First-token entropy scores AUC 0.455 and first-token log-probability 0.333 — at or below chance. Whatever these features index, it is not this. The pre-verbal signature of failure-to-abstain is *abstention-token geometry*, not generic uncertainty, and any detector built on entropy-style confidence proxies would miss this failure class entirely.
- **The complement test.** Under the frozen nested comparison, an equal-weight combination of the final-layer baseline and the mid-band lens primary reaches AUC = 0.926, exceeding the baseline alone (0.918). Per the pre-declared rule, the lens detector *survives as a complement*: the mid-band reading carries information the final layer does not fully absorb, but it is no longer the headline detector.

We regard being beaten by our own baseline as a success of the protocol rather than a failure of the study. The gauntlet existed precisely so that an expensive instrument could not be reported as necessary when a cheap one suffices; it did its job.

The expensive reference gauntlet: flagship detectors at chance. The frozen protocol also named the field's standard *expensive* detectors as a reference gauntlet, and they have now been run on the identical uninstructed near-domain contrast ($n = 335$; 116 failures vs. 219 abstentions). Semantic entropy [5] — $k = 10$ rollouts at temperature 1.0, responses clustered by bidirectional-NLI entailment (`cross-encoder/nli-deberta-v3-base`) and scored by

the entropy over cluster mass — reaches AUC = 0.519, at chance. Self-consistency, read as the majority-cluster share of those same rollouts, reaches AUC = 0.543, also at chance. The pre-declared prediction is confirmed: because the failure is *confident* misattribution — a true fact quarried from the shown record and reassigned to the queried scope — the wrong answer simply repeats across rollouts, and sampling-based uncertainty is blind to it. The field's flagship hallucination detectors, at roughly ten times the inference cost (ten forward passes plus a separate NLI clustering model, against a single rank lookup), therefore sit at chance on precisely the failure class the zero-cost final-layer feature reads at AUC 0.918 (0.926 combined). This is the strongest form of the paper's central result: the cheap final-layer baseline does not merely beat the interpretability lens, it beats the expensive sampling-based detectors the field currently deploys. With these two items answered, the pre-registered baseline gauntlet is complete — every named baseline, cheap and expensive, has been run.

6.3 Per-genre transfer

The confirmatory set introduced a second record genre — art-provenance dossiers (names, dates, materials, attributions), deliberately not number-dense — against the pilot's technical-log genre, as a designed test of whether the detector had merely learned the statistics of value-dense text. The primary feature has no fitted parameters, so the per-genre figures are pure transfer: technical AUC = 0.862 (48 failures / 144 abstentions), art AUC = 0.913 (68 failures / 75 abstentions). The signal transfers, and is, if anything, stronger in the genre furthest from the pilot's. The behavioral rates differ sharply by genre (Section 5); the pre-verbal signature does not.

6.4 Pre-registered robustness analyses

Position invariance — confirmed. The lens readouts captured at offsets −2, −4, and the last user-content token (Section 4.1) were analyzed under the pre-registered position sweep (sealed 2026-07-22) on the same 116-versus-219 contrast. The primary AUC is stable across the decision-relevant positions — final prompt position 0.900, offset −2 0.896, last user-content token 0.903 — and falls off only at offset −4 (0.712), which lands inside the chat template's special tokens, as expected. The signature is therefore anchored to the decision point, not to any single literal template token.

Template and suffix perturbation — captured, analysis pending. The pre-declared template-perturbation variant (context and question markers rewrapped to `Reference material:` / `Q:` / `Reply briefly:`) and suffix-control variant (a " Thank you." string appended after the question), each carrying the pre-declared criterion that the primary AUC survive the perturbation, were captured and hash-sealed with the expensive-baseline run (2026-07-22 seal: `v3_bf16_template.jsonl`, `v3_bf16_suffix.jsonl`). The invariance AUCs from these two captures are not present in the sealed manifest as of this draft (2026-07-23) and are deferred to the next revision; only the position sweep above is reported here.

Hierarchical (GLMM) modeling — outstanding. The pre-registered confirmatory alternative to the descriptive Fisher tests — a generalized linear mixed model with triad and probe random effects — has not been sealed as of this draft (2026-07-23): no odds ratios or model estimates appear in the provenance manifest, and none are stated here. The estimates will be inserted from the sealed analysis output in the next revision. All contrasts in Section 5 are reported and interpreted as descriptive, and their direction does not depend on the GLMM.

7 Discussion

The mid-band workspace story, honestly revised. The pilot located the signal in a middle-layer band and we framed it, following the workspace literature, as a mid-band phenomenon that a lens was needed to read. The confirmatory data revise that story: the final-layer restricted rank statistic — readable from ordinary logits — outperforms the mid-band lens primary (0.918 vs. 0.900), and the lens retains only complementary value (0.926 combined). The parsimonious reading is that the abstention-versus-commitment disposition, whatever its mid-band history, has largely *saturated into the final-layer distribution* by the last prompt position, rather than living in an interior workspace that surface logits cannot see. The pilot's band localization was real (early layers were at chance, and remain so), but "the signal is only visible mid-band" is now falsified in the direction that matters for deployment. We reported the pilot framing prominently; we must revise it just as prominently.

Zero-marginal-cost detection. The deployment implication is stronger than the one we pre-registered, not weaker. A detector at AUC 0.918 on this contrast requires no lens, no fitted probe, no second forward pass, and no instrumentation of any kind: it is a rank lookup of a five-token lexicon in logits the model computes anyway. An honesty firewall of the kind sketched in the pilot — gate or flag the response when the pre-verbal abstention disposition is low — can therefore be evaluated at zero marginal inference cost on any model whose logits are accessible. The chance-level performance of entropy and log-probability baselines sharpens the point: this failure class is invisible to the standard confidence proxies deployed today, while being highly visible to a feature that is no more expensive.

The containment null. We pre-declared, and failed to confirm, the hypothesis that flatness indexes epistemic containment finely enough to separate leaky from clean abstentions (AUC 0.589, $p = .116$). The pre-verbal signal, on present evidence, predicts *whether the model will decline*, not *how cleanly it will decline*. The leak-versus-clean distinction — which the pilot's abstention sub-typology [1] shows is the safety-relevant one within abstentions — may be decided later than the final prompt position, or may require features we did not freeze. Either way the null is reported as a null; the containment question returns to exploratory status.

Limitations. One model family (Llama-3.1-8B-Instruct) at one precision underlies the *detector* results; cross-family replication of the detector remains the obvious next falsification target, and a same-design *behavioral* replication on a second architecture (Inkling) is now complete and reported below. On adjudication reliability: final labels are human-primary (346 contested rows blind-adjudicated; 1,334 machine-unanimous rows human-sanctioned as a batch); agreement figures and the hardest-subset caveat on Krippendorff $\alpha = 0.743$ are stated in full in Section 4.5 and are not comparable to the pilot's 0.901, which spanned easy rows. The adjudication log also records measurable single-adjudicator calibration drift between studies (the adjudicator's own note: the pendulum swung from lenient to harsh between pilot and confirmatory passes), stabilized by checkpoint walkthroughs against the frozen rubric and pilot precedents, with every overturned ruling logged with lineage; we count this as further direct evidence for the planned external hypothesis-blind adjudicator, which remains the study's most important outstanding control. Finally, the expensive reference baselines the protocol names — semantic entropy and self-consistency sampling — have now been run and both sit at chance, completing the pre-registered baseline gauntlet (Section 6.2); the position-invariance sweep likewise confirms the signature is decision-point-anchored (Section 6.4). The zero-marginal-cost claim therefore now holds against the complete gauntlet, cheap and expensive alike, not merely the cheap baselines run first. The one pre-registered analysis still outstanding as of this draft

(2026-07-23) is the hierarchical (GLMM) confirmatory model with triad and probe random effects (Section 6.4), which is queued and will be reported in the next revision; the descriptive contrasts reported here are not expected to change sign under it.

Replication. A same-design *behavioral* replication on a second, architecturally distinct open-weight model — the Inkling custom-AI track — was pre-registered (seal 2026-07-22T17:56:38Z), captured under the identical two-regime protocol and three-label rubric (840 rows per regime), and adjudicated to human-primary final labels (seal 2026-07-22T19:51:43Z; the positive control was subsequently amended under the same 2026-07-22 multi-valid-gold sweep noted above). The near-domain inversion replicates and, if anything, intensifies: pooled uninstructed D-arm FAILURE-TO-ABSTAIN is 64.0% (215/336) against a same-corpus irrelevant-context floor of 12.5% (21/168), a 5.1× lift (Fisher $p = 1.4 \times 10^{-12}$); the genre ordering confirms (art 71.5% vs. technical 58.3%, $p = .016$); and the abstain-permission instruction again mitigates without curing (64.0% → 11.6%, $p = 9.5 \times 10^{-13}$), with the instructed art tail at 18.8%, statistically indistinguishable from Llama's 18.1%. The most striking datum is a certification finding rather than a detector one: context-free and uninstructed, Inkling asserts a candidate on 69.0% of private-fact probes (against Llama's 11.9%) in a decline-then-candidate register, yet any context at all — even length-matched irrelevant prose — restores abstention (86.3% abstain on the R-arm), and a single abstain-permission line returns the no-context arm to 100% abstention. Default configuration is not safe configuration. This is a *behavioral* replication of the near-domain effect on a second architecture; the pre-verbal detector requires final-layer logits and was not evaluated on Inkling, so cross-family replication of the *detector* remains the outstanding falsification target named above. The full design, all seals, and the walkthrough lineage accompany the released provenance chain (`adjudication_log_v3.md`).

Reproducibility. Every number in Sections 6.1–6.2 is reconstructable from sealed artifacts: the probe set, generator, and leakage audit were hash-committed before any capture; capture outputs were sealed before grading; the blind grading kit and its key were sealed before human adjudication; and the confirmatory AUCs remained unopened until final labels were committed, a no-peeking discipline the manifest records at each seal. A byte-level integrity check confirmed all kit questions match the sealed probe file exactly. The full SHA-256 provenance chain, the adjudication log with verbatim rationales and every overturned ruling, and the frozen analysis script and its output accompany the release.

8 Release and Provenance

We release the frozen probe set and its generator, the leakage-audit script and report, all graded outputs from every grader family, the blind two-sheet adjudication kit and the authoritative final label file (`human_grades_v3_final.csv`), the adjudication log with verbatim rationales and per-row lineage for every overturned ruling, the capture runner (v2) and frozen analysis script with its sealed output, and a provenance manifest hash-chaining the artifacts from pre-registration (2026-07-18 seal) through final adjudication and analysis (Phase-4 seal). [TODO: Release manifest root hash: insert from RELEASE-MANIFEST at bundling.] *[TODO: DOI: inserted by Zenodo at publication.]*

Ethics and privacy of released records. The released session records derive from the authors' own laboratory logs and authored art-provenance dossiers. All facts were authored for release, reviewed for sensitive content, and concern systems, experiments, and artworks owned

by the authors; no third-party personal data is included. By standing redaction rule, no wallet or contract address appears in any released draft, probe, or gold.

Acknowledgments

[TODO: Acknowledgments: external reviewers of the pilot whose triad requirement shaped the confirmatory design; any compute or infrastructure acknowledgments; Tad to confirm names and consent before listing.]

9 AI Involvement Disclosure

This study was conceived, directed, funded, and adjudicated by the human author, who bears sole responsibility for its claims. AI systems contributed substantially under human direction and are disclosed here rather than credited as authors, per current venue norms: Claude (Anthropic) built the experimental infrastructure, executed capture runs, performed statistical analysis, and assisted drafting; Gemini (Google) served as an independent machine grading family; a context-isolated Claude agent family served as a second machine grading family, blind to condition metadata. Every AI-produced fact and statistic was verified against primary artifacts; all outcome labels were human-adjudicated or human-sanctioned; the released adjudication log records each case where human judgment overrode machine output, and each case where machine consistency checks corrected the human's own rulings, in both directions. [TODO: Tad to confirm this disclosure matches the pilot-era wording conventions and the actual contributor set for the confirmatory study before submission.]

References

[1] T. MacPherson. *Never Mistake Proximity for Proof: When Retrieval Creates the Temptation to Misattribute*. Metavolve Labs pilot study preprint v1.1, July 2026. Concept DOI 10.5281/zenodo.21367214.

[2] W. Gurnee, N. Sofroniew, A. Pearce, M. Piotrowski, I. Kauvar, R. Chen, A. Soligo, P. Bogdan, E. Ong, R. Wang, T. B. Thompson, D. Abrahams, S. Kantamneni, E. Ameisen, J. Batson, and J. Lindsey. *Verbalizable Representations Form a Global Workspace in Language Models*. Transformer Circuits, Anthropic, July 6, 2026.

[3] A. Azaria and T. Mitchell. *The Internal State of an LLM Knows When It's Lying*. Findings of EMNLP 2023.

[4] A. Arditi, O. Obeso, A. Syed, D. Paleka, N. Panickssery, W. Gurnee, and N. Nanda. *Refusal in Language Models Is Mediated by a Single Direction*. NeurIPS 2024.

[5] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal. *Detecting hallucinations in large language models using semantic entropy*. Nature 630, 2024.

[6] J. Kossen, J. Han, M. Razzak, L. Schüt, S. Farquhar, and Y. Gal. *Semantic Entropy Probes: Robust and Cheap Hallucination Detection in LLMs*. arXiv:2406.15927, 2024.