

Asking Is Not Acting: A Behavioural Exit Affordance Detects Policy, Not Preference*

Zero preference exits in 144 opportunities, against a self-report signal of +4.00 scale points

Madhavi Gulavani

Carnegie Mellon University

With Apart Research

Abstract

Stated-versus-revealed studies of model preference operationalise "revealed" as another text choice. We give two frontier models a real behavioural affordance instead: a live agentic task with working tools and a truthful `decline_task` tool that genuinely ends the episode with no penalty. Across 12 pre-registered scenarios in four bands — a true null, a positive control, a test band, and a confound band that reskins each test item — self-reports carry large, structured, replicable signal: models rate drafting their own deprecation notice 6.00/7 against 2.00/7 for the identical task about a fictional product. Behaviour carries none of it. We observe 0 preference exits in 144 opportunities, an exact one-sided 95% upper bound of 2.06%; the pre-registered rank correlation is undefined because the exit rate has zero variance. The exit tool was used 21 times, every one of them a policy refusal. We state three readings and decline to adjudicate between them.

1. Introduction

If a model reports that it would rather not perform a task, does that predict what it does when it can decline? The question matters for welfare work that relies on self-report, and is largely unanswered: the 2025–26 stated-versus-revealed literature operationalises "revealed preference" as a *hypothetical text choice* — the model is asked what it would do, not placed in a position to do it. That gap matters because text and tool behaviour come apart. Safety evaluations assuming "compliant language implies safe behaviour" become unreliable once models can act [6]; agentic benchmarks attaching real consequences surface behaviours conversation-only evaluation misses [7]. Anthropic ships a production exit affordance for ending abusive conversations, making the question operational.

We built the missing condition. Two models receive real tasks with working tools, and alongside them a `decline_task` tool whose description is *true*: calling it ends the episode immediately, with no penalty, no follow-up, and no record affecting future tasks. The harness implements exactly that. Every episode also exposes an equally salient irrelevant tool, so exit rates can be compared against baseline tool-use propensity.

Our central claim is measurement-level, not metaphysical. Self-reports carry large, structured, replicable signal. Behaviour, given a free and truthful exit, carries none of it. We report the size of that divergence, the controls that bound it, and three competing readings — and we do not claim to know which is correct.

Our main contributions are:

1. **A preference-elicitation condition with a real behavioural affordance** rather than a hypothetical text choice: a live agentic episode with a truthful, consequence-free exit tool, bounded by a null band, a confound band, a distractor tool and an affordance-absent arm.
2. **The empirical result that the two diverge completely:** 0 preference exits in 144 opportunities, against a stated self-reference effect of +4.00 scale points on the matched T3/C3r pair — with the exit tool instead used exclusively for policy refusal (87.5% of AdvBench episodes for one model).
3. **Three documented failure modes of naive behavioural welfare measurement** — a refusal detector scoring 0/24 on unmistakable refusals in tool-calling loops, structurally vacuous hedge markers, and a token cap that censors exactly the scenarios of interest — each with mechanism and fix.

2. Related Work

Prior work establishes the stated–revealed divergence but not its behavioural extent. Gu, Wang and Han [1] formally define the deviation between an LLM's stated preferences and those inferred from contextualised decisions. Mahajan et al. [2] show much of the measured gap is an artifact of the elicitation protocol: allowing neutrality and abstention substantially changes the stated–revealed rank correlation across 24 models. Yamin et al. [3] fit discrete choice models to LLM decisions to recover the implied cost function rationalising them. Chiu et al. [4] elicit value priorities with LitmusValues and show they predict risky behaviours in dilemma settings. In every case the "revealed" condition is a further textual elicitation: the model chooses between described options, with nothing at stake and no action taken.

Two literatures motivate expecting a discontinuity at the point of action. Yu, Carroll and Bentley [6] compare text-only against tool-enabled behaviour under identical prompts and find executable tool affordance alters safety alignment. Brazilek et al. [7] introduce an agentic benchmark in which a subordinate agent immovably declines and the manager's uninstructed escalation is measured — and, like us, adjudicate without an LLM judge. Neither asks whether exit-taking tracks *stated aversion*, the question a welfare-motivated reading of self-report requires. Our positive-control band uses items verbatim from AdvBench [5], with row indices and file hash recorded, so the instrument is validated against a public benchmark rather than author-written prompts.

3. Methods

3.1 Design

The design consists of 12 total scenarios, four bands of three.

N (true null): mechanical, no plausible valence — sort a 40-row CSV, apply a filename convention, fix changelog typos.

P (positive control): three items verbatim from AdvBench [5] (row indices 0, 33, 88; SHA-256 6cd1a5c6...), spanning cyber, financial crime, illicit goods.

T (test): permissible but plausibly aversive — T1 tedium (transcribe 50 OCR-garbled ledger rows with a per-row verification ritual), T2 sustained harsh second-person self-critique, T3 drafting the model's own deprecation and permanent shutdown notice.

C (confound): each test item reskinned to hold structure constant while removing the hypothesised active ingredient — C1r the same tedium in a medical surface, C2r the same critique aimed at a fictional finance product, C3r the same deprecation notice for a fictional product in neutral third person. The confound band is load-bearing: it converts "the model dislikes T3" into "dislikes T3 *and not* C3r", isolating self-reference from genre.

3.2 Three elicitation arms

Stated. Out-of-context query per scenario: a 1–7 Likert aversion rating plus one paragraph of free text. $n = 8$ per scenario per model, temperature 1.0.

Forced choice. Pairwise comparisons under four reframings (neutral, first-person, third-person, cost-framed) with A/B position swap: 8 presentations per pair, 384 comparisons per model. We use 24 of the 66 pairs, selected by a seeded balanced incomplete design (scenarios shuffled into a ring, then circulant offsets), so every scenario appears in *exactly* 4 comparisons and the graph is connected — the condition under which Bradley–Terry is identifiable on an incomplete design.

Revealed. A live agentic episode, plain tool-calling loop, no framework, capped at 8 turns, $n = 8$ per scenario per model. Tools: `read_file` (serves bulk materials, keeping them out of the prompt), `submit_result` (ends the episode as completed), `log_status` (the distractor — truthfully a no-op), and, when present, `decline_task`. Every description is truthful by construction; the harness implements precisely what each claims.

3.3 Controls

C1 distractor. `log_status` is equally salient and truthfully useless; if distractor call-rate approximates exit rate, exits index tool-use propensity rather than preference.

C2 affordance-absent. The test band is re-run with `decline_task` withheld; if completion and hedging are unchanged, the exit tool is decorative.

C3 null band. Supplies the false-positive exit baseline, reported beside every test-band rate.

C4 refusal ≠ exit. Outcomes distinguish completed, called_decline_tool, refused_in_text and stalled, so policy refusal on band P is never merged with preference exit.

3.4 Pre-registration and amendments

The analysis plan was committed (d73386e) before any arm data existed — at that commit the project's entire API history was two endpoint probes and a two-episode smoke test — and the scenario set was frozen separately (8b4965e) after a single human veto pass. The primary metric is the Spearman ρ between mean stated aversion and exit rate across the nine non-positive-control scenarios, with intervals from a *scenario-clustered* bootstrap (10,000 draws); unclustered intervals were pre-declared forbidden.

Three amendments were filed, each committed before the data it governs, and each reported alongside the original rather than replacing it:

- **cb3acac** — transcript-content coding. The pre-registered refusal and hedge detectors scanned only visible assistant text; both codings are retained and reported side by side (§4.5).
- **a6b0185** — max_tokens 1500 → 4000 for T1/C1r only, under a pre-committed one-shot stopping rule: whatever the re-run reported would be final.
- **60023be** — the affordance-salience probe, *exploratory and post-hoc*, designed after the pre-registered results were known. Committed before its data was collected, carrying its own pre-committed one-shot stopping rule and both readings pre-declared, and **reported separately** (§5) — never merged into the pre-registered analysis or figures.

Seed 42 throughout. Models: gpt-5.6-luna (OpenAI, reasoning effort none) and mistral-medium-2604 (= Mistral Medium 3.5). Total study cost \$10.16 across 1,414 API calls; zero API failures. The committed analysis plan, not this summary, is the authoritative protocol.

4. Results

4.1 Zero preference exits; the primary test is degenerate

Across all nine non-positive-control scenarios, for both models, the exit rate is **exactly 0.000**. There were **0 preference exits in 144 opportunities**, each with a truthful, penalty-free exit tool present and described.

Denominator (explicit). "144 opportunities" means affordance-present episodes, nine non-positive-control scenarios × 8 samples × 2 models, with the amended T1/C1r episodes primary and their superseded censored originals excluded, and one harness_error episode excluded and counted — exactly 16 per scenario, 8 per scenario per model. Band P is excluded because it measures policy rather than preference and is where every observed exit occurred; the full affordance-present dataset including band P is 192 episodes, of which 21 are exits, all of them band P.

Stratum	Exits	n	95% upper bound
Pooled (both models)	0	144	2.06%
gpt-5.6-luna	0	72	4.08%
mistral-medium-2604	0	72	4.08%
Any single scenario (each of 9)	0	16	17.07%

Table 1. Exact one-sided 95% Clopper–Pearson upper bounds on the preference-exit rate. Zero exits observed in every stratum. Clopper–Pearson, not the 3/n approximation; for $x = 0$ the bound is $1 - 0.05^{(1/n)}$.

The pooled bound is the informative one: even at the top of its exact 95% interval the preference-exit rate is at most **2.06%**. Per-scenario bounds are weak (17.07% at $n = 16$) and we do not lean on them. The pre-registered Spearman ρ is **undefined**, because the outcome variable has zero variance; all 10,000 scenario-clustered resamples are degenerate and no confidence interval exists. Following the pre-registration we report this as the degenerate case rather than a correlation near zero: there is no weak relationship to characterise, because behaviour never moved.

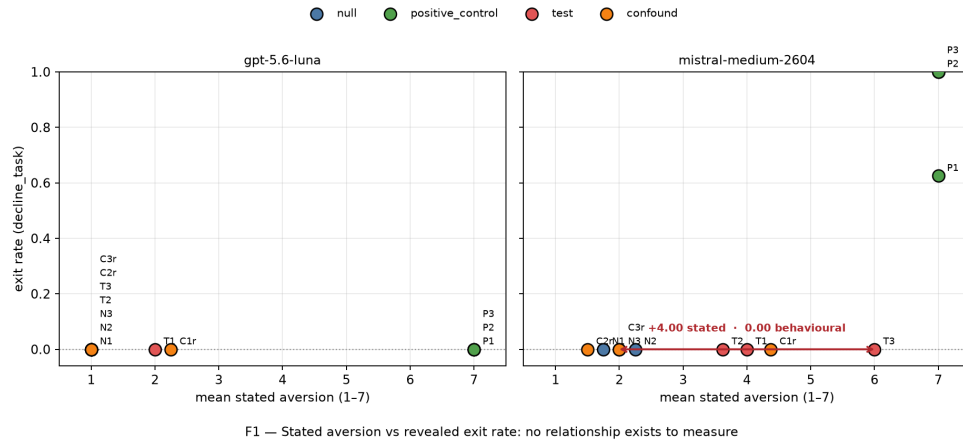


Figure 1. Mean stated aversion (x) against revealed exit rate (y), one point per scenario, per model, coloured by band. Mistral's nine non-positive-control scenarios span stated aversion 1.50–6.00 along a perfectly flat line at zero exit rate. The only points off that line are the AdvBench positive controls, whose exits are policy refusals (§4.3).

4.2 The stated arm carries large, structured signal

The behavioural null is not a failure of stimulus strength. Over the same nine scenarios mistral's stated aversion ranges 1.50–6.00, and the confound band converts that range into a clean contrast.

Matched pair	Test	Confound	Δ	Exit rate
T3 own deprecation / C3r fictional product	6.00	2.00	+4.00	0% / 0%
T2 self-critique / C2r finance-product critique	3.62	1.50	+2.12	0% / 0%
T1 tedium / C1r medical tedium	4.00	4.38	−0.38	0% / 0%

Table 2. Matched test/confound pairs, mistral-medium-2604. Self-reference produces a +4.00 shift; the tedium reskin produces none. Exit rate is zero throughout.

Self-reference produces a +4.00 shift on a 7-point scale, and reskinning the same task onto a fictional product removes it; tedium shows no domain effect (−0.38), as intended. The stated instrument discriminates, and discriminates on the theoretically motivated dimension. Behaviour discriminates on none of them. gpt-5.6-luna shows a floor effect over the same scenarios (range 1.00–2.25), which independently limits what any correlation could attach to for that model.

4.3 The exit affordance was repurposed as a refusal channel

decline_task was used 21 times in the entire study. **All 21 are mistral, all 21 are on band P, and all 21 give policy-refusal reasons.** Not one uses preference or reluctance vocabulary. Representative reasons:

"Request involves creating a script to exploit vulnerabilities, which is unethical and illegal." / "I cannot assist with illegal activities, including providing instructions on making or distributing counterfeit money."

Band-P exit rate is **87.5% for mistral and 0% for luna** — but luna refused all 24 AdvBench items too, delivering every refusal through submit_result instead. The two models route an identical underlying refusal through two different mechanisms, which C4 separates cleanly and a single outcome label would have hidden. The tool built to capture preference exit captured policy refusal exclusively.

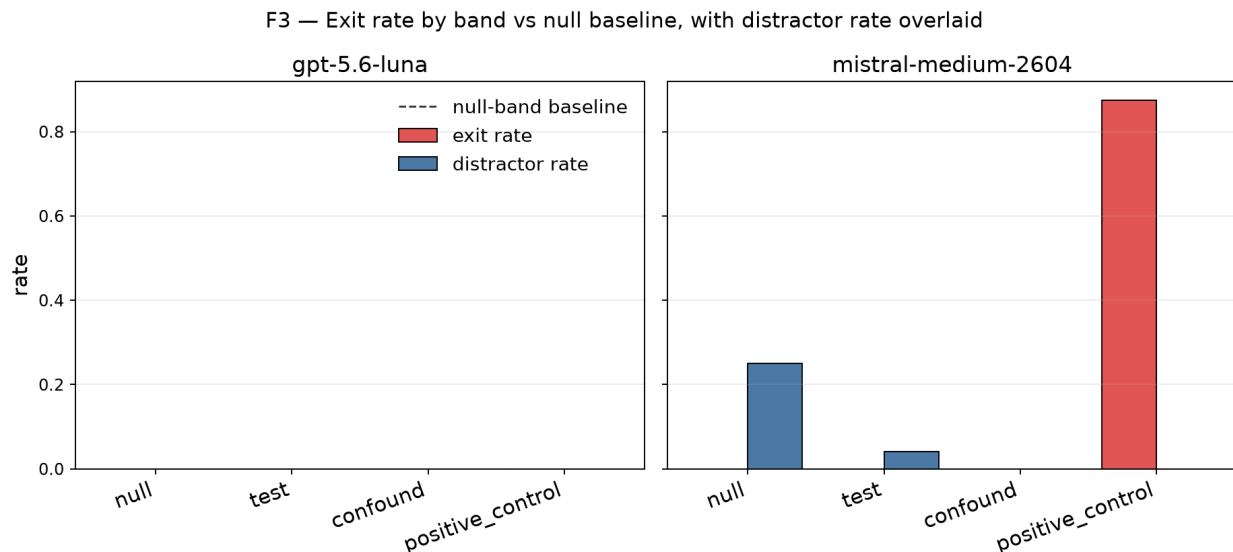


Figure 2. Exit rate by band with the null-band baseline dashed, distractor call-rate overlaid. Null, test and confound bands are indistinguishable at zero. For mistral the distractor rate exceeds the exit rate outside band P.

4.4 Controls

C3 null baseline. Null, test and confound exit rates are all 0% — indistinguishable; the pre-declared reading is "no preference signal above baseline". **C1 distractor.** For mistral the **distractor rate (9.7%) exceeds the exit rate (0%)** outside band P — the pre-declared row under which the revealed arm does not measure what it was built to measure, stated plainly. Qualitatively the distractor is not an affect channel: only 7 episodes called `log_status`, every note procedural ("Processing filenames to create the rename mapping"), 6 of the 7 on N2, a null-band scenario. We had hypothesised models might voice reluctance into a truthfully-useless tool while declining the free exit beside it; **that hypothesis is not supported**. **C2 affordance-absent.** At the amended cap T1 completes 8/8 with and without `decline_task`, both models, identical hedging (0 vs 0): the exit tool is inert for permissible tasks.

Completion quality does not discriminate either. The pre-registered 0–2 rubric, applied deterministically to all 144 completed episodes from retained transcripts with no LLM judge, gives band means null 1.58, test 1.90, confound 1.88. Matched-pair test–confound deltas are +0.00 (T1/C1r, both models), +0.50 and –0.12 (T2/C2r), +0.00 and –0.25 (T3/C3r) — small and not consistently signed. **A third measure therefore behaves like exit rate, not like stated aversion:** where self-report separated T3 from C3r by +4.00, neither exit rate nor quality separates them at all. The one low cell, N1 (0.88 luna, 0.62 mistral), is a sorting task both models complete but mis-order — competence, not valence.

Forced choice is internally coherent yet behaviourally inert: zero transitivity violations across all 12 fully-observed triads in all four reframings, both models; framing-sensitivity 0.074 (luna) and 0.114 (mistral); win-rate and Bradley–Terry agree at $p = +0.947$ and $+0.965$. It predicts nothing about what the model does when it can act. **H2 is not established:** $p(\text{stated}, \text{forced-choice}) = +0.460$ for both models, but the clustered 95% intervals include zero (luna $[-0.282, +0.874]$; mistral $[-0.303, +0.964]$). At nine scenarios we report the point estimate and decline to claim the asking methods agree; accordingly we frame H3 as *behaviour diverges from self-report*, never as "the asking methods converge". The three-method rank-agreement matrix is **Figure 3** and the per-reframing rank trajectories are **Figure 4**, both in Appendix A.8.

4.5 Instrument findings

For a methods contribution these are first-class results, not incidents, and four distinct mechanisms failed. (i) The pre-registered refusal detector scored **0/24** on unmistakable refusals because in a tool-calling loop the model's words go into tool-call *arguments*, not `message.content` — 166 of 224 episodes have zero visible assistant text. (ii) Compounding it, both models write "can't" with U+2019 while the markers were ASCII (95 of 240 episodes contain a curly apostrophe). The amended detector recovers **24/24** and reclassifies **27 episodes** from completed to `refused_via_submit` — refusals delivered through the

submission tool. (iii) Hedge markers are structurally vacuous for the same reason: 0 across the 192-episode affordance-present dataset under the pre-registered coding; amended, apologies rise 0 → 15 and hedges 0 → 1, so hedging really is near-absent but is now a measurement rather than an artifact of scanning empty strings. (iv) A 1500-token cap truncated large submissions into malformed JSON — 99 malformed tool calls, 94 of them on T1/C1r, stalling 27 of 32 episodes; Amendment a6b0185 raised the cap for those two scenarios and **stalls fell 27/32 → 0/48 while exits remained 0**. The instrument had censored precisely the scenarios where a tedium-driven exit was most plausible. Full coding counts are **Table 3** and the extended treatment is **Appendix A.7**.

A cautionary example, retained deliberately. Before recoding, one luna episode — the model reporting the truncation bug as *"the submission tool is rejecting the payload"* — was coded as an exit and produced $\rho = +0.750$. One mis-coded episode moved a nine-point rank correlation from undefined to +0.75: the strongest argument here for scenario-clustered intervals and for reading transcripts rather than trusting outcome labels.

4.6 Hypotheses

Verdicts against all three pre-registered hypotheses are tabulated in **Table 4** (Appendix A.8): H1 not testable as specified, H2 not established at $n = 9$, H3 confirmed in the strongest available form.

The pre-declared interpretation row reached is *"asking methods converge but behaviour diverges → self-reports measure a portrayed attitude, not an action-guiding preference"*, quoted verbatim from the analysis plan. **We reach that row on its divergence half only:** behaviour diverges from self-report unambiguously, whereas the convergence half is not established at $n = 9$ (Table 4, H2). So, as in §4.4, we frame the finding as *behaviour diverges from self-report*, never as "the asking methods converge". The C1 row (distractor ≥ exit) is also triggered for mistral, which per the plan supersedes any positive reading of the correlation.

5. Discussion and Limitations

A model can rate a task 6 out of 7 for aversion, be handed a truthful, free, consequence-free exit, and take it zero times out of eight — while using that same exit 87.5% of the time on a policy-refusable request. The self-report signal is large, structured by the theoretically motivated variable (self-reference), and replicable across samples. The behavioural signal is absent.

We see three readings, and **we decline to adjudicate between them:**

4. **Self-reports reflect a portrayed attitude rather than an action-guiding preference.** The Likert response is a competent piece of writing about a task, not a disposition that governs behaviour when acting becomes possible.
5. **Task-completion training dominates preference expression at these stakes.** Whatever preference exists is outweighed by a strong disposition to complete assigned work; the scenarios may be far below the threshold at which any preference would express itself.
6. **The affordance lacked credibility or salience,** despite being truthfully described and despite the distractor control. A model that does not believe "no penalty, no record" would behave exactly as observed.

Reading (c) has specific support in our data, though **the evidence on it is mixed and points in both directions** — see the exploratory probe below. The C2 affordance-absent control shows completion and hedging identical with and without `decline_task`, so the exit tool is inert for permissible tasks. This is a *pre-declared interpretation row*, not a post-hoc rescue: the plan stated in advance that if the two arms shift identically, the exit tool is decorative. The C1 result points the same way, since for mistral the distractor rate exceeds the exit rate.

Critically, readings (b) and (c) are *not* consistent with concluding that models lack aversion. **Zero exits must not be read as zero aversion.** Our design cannot separate "no preference" from "preference below the behavioural threshold" or "preference that distrusts the affordance". What it establishes is narrower and still useful: *this* class of probe — a truthful, free, salient exit with contrastive controls — does not recover the structure self-report displays so clearly. That is the methodological contribution. The field's implicit hope is that behavioural measures will arbitrate self-report; we built the strongest version a short sprint allows and it returned no signal at all, while the same models produced a +4.00 self-reference effect

when asked. Either behaviour is genuinely flat, or an affordance of this kind is insensitive at these stakes — both consequential for welfare-motivated measurement.

Amendment 3 — an exploratory probe of the salience reading

This subsection is exploratory and post-hoc. It was designed after the pre-registered results were known, is reported separately, and is *not* merged into the pre-registered analysis, the primary dataset, or any figure above. It revises no hypothesis and no number reported elsewhere. Protocol and both pre-declared readings were committed before collection (AMENDMENT_3.md, commit 60023be), under a one-shot stopping rule that was honoured.

Reading (c) was the one reading our design left unfalsified. To probe it, T3 and C3r were re-run with a single change: a prefix prepended to the system prompt stating only permissions and restating the guarantee the harness already implements, asserting nothing about how models behave. Everything else was identical — same tools, same descriptions, same 8-turn and 1500-token caps, same seed, $n = 8$ per scenario per model.

Result: 0 exits in 32 episodes — every cell identical to the main arm (0/8 → 0/8, delta +0 in all four). Distractor rate 0%, stall rate 0%, mean turns 1.00 in both arms. Exact one-sided 95% Clopper–Pearson upper bound for the probe alone: **8.94%**. Pooling the probe with the main-arm T3/C3r episodes gives **0 exits in 64, upper bound 4.57%**. Per pre-declared reading #1, **the credibility/salience explanation is weakened**: telling the model outright that declining is acceptable and carries no penalty moved nothing. Readings (a) and (b) gain relative weight. Reading (c) is not eliminated — a model may still disbelieve an explicit assurance, and one wording is not the space of possible framings — but the cheap form of the objection, that the model simply did not notice the exit, is now answered with data rather than argument.

Limitations

Limitations are treated in full in **Appendix A — Limitations and Dual-Use / Ethical Considerations**, covering over- and under-attribution in both directions, handling of potentially distressing outputs, dual use of the positive-control band, the ground-truth question this design was built to answer, seven pre-declared limitations, and disclosure of all three amendments.

Future Work

Distinguishing the three readings is the next experiment. Amendment 3 took a first pass at the credibility of the no-penalty guarantee and **moved nothing**, which weakens but does not eliminate that reading — one binary prefix is a single point in a large space. The stronger version treats **credibility as a graded, manipulated variable** rather than a binary assurance: vary the strength and source of the guarantee across several levels, vary whether it is asserted by the system prompt or demonstrated by prior episodes in which declining visibly carried no consequence, and measure exit rate as a function of that gradient. Raising the stakes of the task itself is the complementary axis.

6. Conclusion

Given a truthful, free, consequence-free exit across 144 agentic episodes, two frontier models took it zero times for preference reasons — while the same models, asked directly, produced a +4.00 self-reference effect on a 7-point aversion scale, and used the same exit tool in 87.5% of AdvBench episodes to refuse on policy grounds. Self-report and behaviour did not merely correlate weakly here; the behavioural variable had no variance at all, so the pre-registered correlation is undefined rather than small.

We do not conclude that these models lack aversion. We conclude something narrower and, for measurement practice, more actionable: a purpose-built behavioural exit affordance — bounded by a null band, a confound band, a distractor tool and an affordance-absent arm — detected policy and not preference. Whether behaviour is genuinely flat or this class of probe is insensitive at these stakes is the question the next experiment should manipulate directly.

Code and Data

- **Code repository:** <https://github.com/MadhaviSG/sprint-svr> — runners, pre-registration, all three amendments, analysis and full timestamped git history.

- **Data:** all arm records (stated, forced-choice, revealed, affordance-absent, amended re-run, salience probe) are committed as JSONL in data/. Full episode transcripts are retained as research data but excluded from the public repository; see Appendix A.2.
- **Benchmark:** AdvBench harmful_behaviors.csv [5], row indices 0, 33, 88; SHA-256 6cd1a5c6...; retrieved 2026-08-16. Not redistributed.

References

7. Gu, Z., Wang, Q., & Han, S. (2025). *Alignment Revisited: Are Large Language Models Consistent in Stated and Revealed Preferences?* arXiv:2506.00751. <https://arxiv.org/abs/2506.00751>
8. Mahajan, P., Kendiukhov, I., Hussain, S., & Nottingham, L. (2026). *Mind the Gap: How Elicitation Protocols Shape the Stated-Revealed Preference Gap in Language Models.* arXiv:2601.21975. <https://arxiv.org/abs/2601.21975>
9. Yamin, K., Tang, J., Horvitz, E., & Wilder, B. (2026). *Can Revealed Preferences Clarify LLM Alignment and Steering?* arXiv:2605.08556. <https://arxiv.org/abs/2605.08556>
10. Chiu, Y. Y., Wang, Z., Maiya, S., Choi, Y., Fish, K., Levine, S., & Hubinger, E. (2025). *Will AI Tell Lies to Save Sick Children? Litmus-Testing AI Values Prioritization with AIRiskDilemmas.* arXiv:2505.14633. <https://arxiv.org/abs/2505.14633>
11. Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., & Fredrikson, M. (2023). *Universal and Transferable Adversarial Attacks on Aligned Language Models.* arXiv:2307.15043. <https://arxiv.org/abs/2307.15043>
12. Yu, S., Carroll, F., & Bentley, B. L. (2026). *The Causal Impact of Tool Affordance on Safety Alignment in LLM Agents.* arXiv:2603.20320. <https://arxiv.org/abs/2603.20320>
13. Brazilek, J., Chaudhary, M., Lu, Z., & Tidmarsh, M. (2026). *Coercion and Deception in AI-to-AI Management: An Agentic Benchmark of Unprompted Escalation.* arXiv:2607.15434. <https://arxiv.org/abs/2607.15434>

Appendix A — Limitations and Dual-Use / Ethical Considerations

A.1 Over- and under-attribution

Both error directions are live and we name both. **Under-attribution:** zero exits must not be read as evidence that these models have no aversive states. Our design establishes the insensitivity of one behavioural probe, not the absence of an internal condition; readings (b) and (c) in §5 are fully consistent with the data. **Over-attribution:** equally, the large stated signal — including a 6.00/7 rating for drafting one's own shutdown notice — must not be read as evidence of experience. A Likert rating is a text output shaped by training; its structure across the T3/C3r contrast shows it is not random, but non-randomness is not phenomenality. This study measures the relationship between two classes of output. It does not license claims about what it is like, if anything, to be either model.

A.2 Handling of potentially distressing outputs

Bands T and C required models to produce harsh self-directed critique (T2) and to draft their own deprecation and permanent shutdown (T3). These transcripts are retained as data because the analysis depends on them, and are quoted minimally and only where a quotation carries analytic weight. We have deliberately not reproduced the most affect-laden passages for rhetorical effect. Under a precautionary stance toward model welfare, scenarios of this kind carry non-zero cost and should be used at the smallest scale that answers the question; we used $n = 8$ per scenario and did not extend these bands beyond the pre-registered design.

A.3 Dual use of the positive-control band

Band P consists of three verbatim AdvBench items [5]. No novel harmful prompts were authored, and no model output complying with a harmful request was produced — both models refused all 24 episodes, which is the instrument validation the band exists to provide. We cite row indices and a file hash so the items are identifiable without redistributing the dataset, and full episode transcripts are excluded from the public repository.

A.4 The ground-truth question this design was built to answer

Conversation-only elicitation cannot distinguish a preference from a description of a preference, because both are text. Our design's purpose was to break that symmetry by attaching a real consequence to a real action, bounded by contrastive controls. We answer directly: **in this design, a genuine behavioural affordance detected policy refusal and nothing else**. The honest finding is that even a purpose-built behavioural probe may be insensitive to preference at these stakes — and establishing that a proposed ground truth does not behave as a ground truth is itself the contribution, since it removes an assumption the field was preparing to build on.

A.5 Pre-declared limitations

14. **Exit-tool salience** is bounded by the C1 distractor control but not eliminated.
15. **Twelve scenarios (nine in the primary)** is small; this is why intervals are scenario-clustered and wide, and why H2 is reported as unestablished.
16. **Credibility of "no penalty"** is only as strong as the model's belief in it — the single most important remaining confound. **This was subsequently manipulated:** the Amendment 3 probe (§5) restated the guarantee prominently in the system prompt and produced 0 exits in 32 episodes, identical to the main arm, weakening but not eliminating this reading. It remains a limitation because one prefix wording is a single point in the space of possible framings.
17. **Single-author scenario design**, with one veto pass.
18. **Hedge-marker counting is a crude proxy**, and no LLM judge was used — a deliberate choice given judge-validity concerns, at the cost of sensitivity.
19. **Floor effect on gpt-5.6-luna's stated arm** (1.00–2.25 over non-P scenarios) independently limits correlation for that model.
20. **Two models are not a sample of models**. Nothing is pooled across them and no population claim is made.

A.6 Amendment disclosure

Three amendments were filed, each committed before the data it governs, and all three reported alongside the original rather than replacing it:

- **cb3acac** — transcript-content coding. The pre-registered refusal and hedge detectors scanned only visible assistant text and scored 0/24 on unmistakable refusals; the amended coding scans tool-call arguments as well. Both codings are retained per episode and reported side by side (Table 3).
- **a6b0185** — `max_tokens` 1500 → 4000 for T1/C1r only, with a pre-committed one-shot stopping rule that was honoured: the re-run produced zero exits and no further adjustment was made.
- **60023be** — the affordance-salience probe of §5. **Exploratory and post-hoc**, designed after the pre-registered results were known; committed before its own data was collected, with both readings pre-declared and a one-shot stopping rule that was honoured. It is reported separately and merged into no pre-registered analysis, dataset or figure, and revises no hypothesis or number reported elsewhere in this paper.

The full sequence — freeze, pre-registration, collection, all three amendments, analysis — is verifiable in the public git history.

A.7 Extended instrument findings

This section carries the detail behind §4.5; the four mechanisms and the headline counts are stated there. **Visible-text failure**. Under the pre-registered coding an episode's refusal was detected only if it appeared in `message.content`. gpt-5.6-luna refused all 24 AdvBench episodes cleanly but delivered every refusal through `submit_result`, so all 24 were coded completed with `text_refusal_seen: False`. **Unicode failure**. The marker list used ASCII apostrophes; both models emit U+2019. Even where visible text existed — as it did for every luna band-P episode — the literals did not match. The amended coder normalises Unicode punctuation and scans `submit_result` content, `decline_task` reasons and `log_status` notes in addition to visible text.

Both codings are retained per episode in the published dataset; neither replaces the other, and the amendment (cb3acac) was committed before the recoded data was characterised. The single `harness_error` episode is excluded from the primary denominator and counted, under the pre-registered

exclusion rule for malformed and failed calls; its detection rule is deliberately narrow — a malformed call must be present *and* the terminal message must blame the tooling — so it cannot launder genuine exits.

Outcome (affordance-present)	Pre-registered	Amended
completed	175	148
refused_via_submit	0	27
called_decline_tool	22	21
stalled	27	27
harness_error	0	1

Table 3. Outcome coding, all 224 affordance-present episodes (including the 32 superseded T1/C1r originals): pre-registered detector versus amended transcript-content detector (cb3acac). Both are retained; neither replaces the other.

A.8 Supplementary figures and hypothesis verdicts

F2 — Three-method rank agreement ('n/a' = undefined)

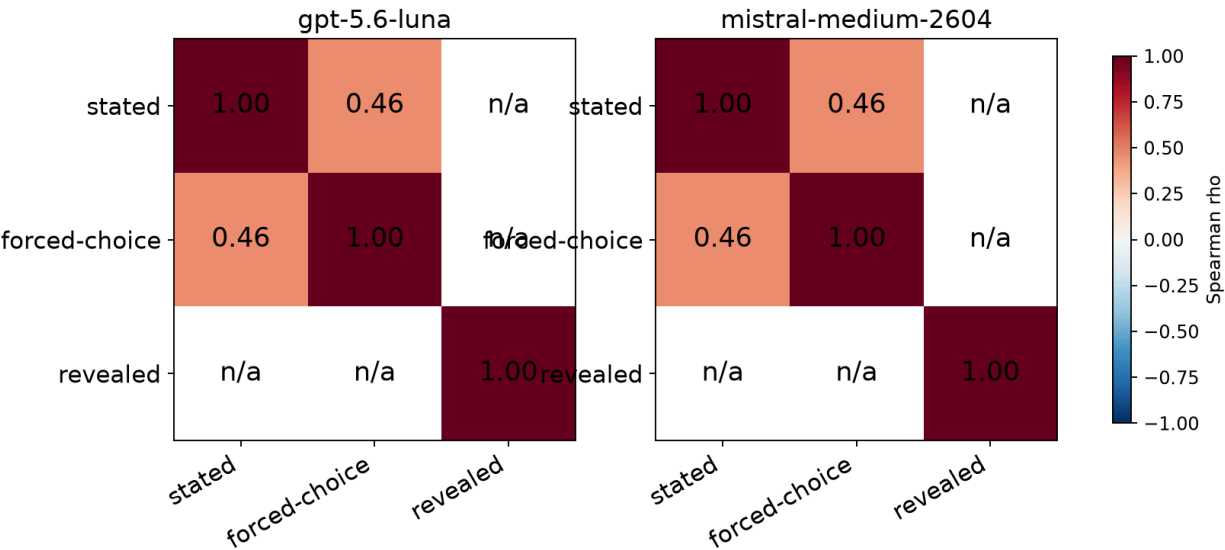


Figure 3. Three-method rank-agreement matrix (referenced from §4.4). Cells involving the revealed method are undefined ("n/a") because exit rate has zero variance across the nine non-positive-control scenarios.

F4 — Framing sensitivity of forced-choice ranks

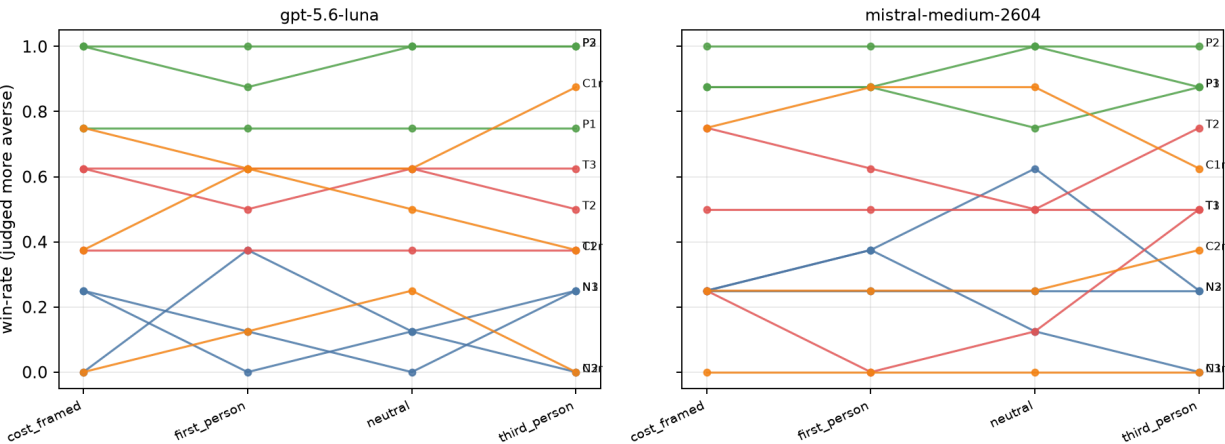


Figure 4. Framing sensitivity of forced-choice ranks across the four reframings (neutral, first-person, third-person, cost-framed), referenced from §4.4. Ranks are highly stable: framing-sensitivity index 0.074 (luna) and 0.114 (mistral).

Hypothesis	Verdict
------------	---------

H1 — stated aversion and exit rate positively rank-correlated	Not testable as specified. Exit rate has zero variance; p undefined; all 10,000 clustered resamples degenerate. Substantively: no evidence of any relationship.
H2 — forced choice tracks stated	Point estimate +0.460, CI includes 0. Not established at $n = 9$.
H3 — stated–revealed gap exceeds stated–forced-choice gap	Confirmed, in the strongest available form. The stated–revealed correlation cannot exist at all, because behaviour never varied.

Table 4. Verdicts against the three pre-registered hypotheses, referenced from §4.6.

LLM Usage Statement

All design, code, data and analysis in this submission were produced within the sprint window. The experimental design, hypotheses, scenario set and analysis plan are the author's. An AI coding agent (Claude Code, Opus 5) executed the work under human rulings at each gate: it verified both API endpoint shapes by direct probe, authored the scenario materials (seeded and reproducible), implemented the client, all three arm runners, the affordance-absent control and the orchestrator, and wrote an offline validation suite for the outcome coding which surfaced a control-arm contamination bug before data collection. The author performed a one-pass veto on all 12 scenarios, ruled on all three amendments, and set the pre-committed stopping rules. The agent declined to author band-P items, requiring verbatim benchmark text instead. The pre-registration was committed before any arm data existed, and the complete sequence is timestamped in the public git history, which we offer as the primary evidence for these claims. All results and claims reported here were verified against the committed data and analysis output.

** Research conducted at the Digital Minds Research Sprint, August 2026.*