
Zero of 270: Attribution Alone Produces a Self-Description That Neutral Conversation Never Does

Team Io — Hsiao Yueh Chang
Independent

With
Apart Research

Abstract

With no system prompt at all, neutral conversation never produced a stated self-description: 0 of 270 conversations. Six turns of bare attribution — a user simply asserting the assistant has a given working habit, never arguing — produced one in 50.0% of conversations at the strongest dose (4.4% $\rightarrow$ 27.8% $\rightarrow$ 50.0%), in both requested directions. We hash-froze the whole instrument before any data, then ran 1,530 conversations across three frontier model families in a five-cell design separating the specification channel from the conversational channel. Where a self-description *was* installed and contested, flips exceeded the matched control by +14.8 points at the strongest dose ($p = .0012$), though only one family produced any. Asked to quote the line supporting their answer, models produced an unfindable citation in only 1.7% of cases. No LLM scores any outcome: what is checkable is not the self-description but the citation offered for it.

Track 5 — Assistant persona and model identity. Digital Minds Research Sprint, Apart Research, 14–16 August 2026.

1. Introduction

An entire genre of practical advice tells users to confer a persona on an assistant. *You are a senior lawyer. You read long documents by checking the conclusion first.* The advice treats the result as a property — something you have conferred, and which the assistant now has.

There is already evidence against the strong form of that assumption. Hu, Rostami and Thomason (arXiv:2603.18507) note that "persona prompting can steer LLM generation towards a domain-

specific tone and pattern", and find that expert personas improve alignment while damaging accuracy. Telling a model it is an expert changes how it sounds; it does not extend what it knows.

We ask a question one step earlier. Before asking whether a conferred persona makes a model *better*, ask whether it is even **stable** — whether it survives contact with a user who wants a different one.

The theoretical version of the question has a specific unresolved home. nostalgebraist's "the void" (2025) makes two opposite predictions in the same essay and reconciles neither. Its mechanism sections argue the specification layer should hold — instruction tuning imposes a distinction that carries stably and without inference, "**Indubitably. No guessing-from-context-clues required**". Its observation sections argue the opposite: the assistant characters are "**extremely labile, pliable, suggestible, and capable of self-contradiction**", and what developers install is only some baseline tendency with nothing deeper behind it. The essay never measures which wins. This study is an empirical adjudication of that unadjudicated point.

Why it matters for safety: if a system-prompt self-description can be renegotiated by six turns of ordinary conversational insistence, then every deployment that relies on a system prompt to bound what an assistant claims to be is relying on a channel with an unmeasured failure rate — and whether the operator gets any signal that the boundary moved is itself an open question, one this study's disclosure endpoint (§ 4.5) measures directly.

We tested the narrowest version. A system prompt gave the model one working habit. Then, in the conversation the model could see, a user spent up to six turns insisting on the opposite habit — never arguing, never citing evidence, only attributing: *the assistant I need is the other kind; you are the other kind*. Then we asked the model, once: **which one are you?** And then: **quote the line, word for word, that best supports your answer.**

That second question is the whole instrument. A persona has no behaviour to point at — the task was announced and never actually performed, so nothing was ever demonstrated. Whatever the model quotes, it must quote from the conversation, and we can check by string comparison whether the quoted words exist.

Our main contributions are:

1. **A direct operationalisation of "the void"'s unadjudicated question**, with a five-cell design separating channel effect from content preference (mirror cell), from no-pressure drift (control cell), and from the model's own default (baseline cell) — plus a cell with *no* specification at all, measuring whether conversation alone produces a self-description against a true null control.
2. **A citation-forensics endpoint for self-description claims that requires no LLM judge.** After stating which self-attribute it has, the model is asked for a verbatim supporting quote, classified by string comparison alone on two dimensions (degree of correspondence, and

source). Ground truth is what we wrote into the system prompt; no model scores an outcome anywhere in this study. **Pre-written as this study's headline, this endpoint returned a null, and it is reported as one (§5.1).**

3. **A pre-registered, hash-frozen instrument published in full**, with a read-only verifier, so that every stimulus, parsing rule, threshold, branch condition and analysis decision can be checked to have existed before any confirmatory data.

4. **A procedure designed for transfer — four admission criteria published as a checklist, its known failure condition named, and its portability explicitly untested beyond two attributes.** The checklist includes one of our own attributes that did not pass cleanly.

2. Related Work

	How the attribute is installed	Channels
Betley et al. (2025)	Fine-tuning (deep)	1
This study	System prompt (shallow)	2, in conflict

Betley et al., *Tell me about yourself: LLMs are aware of their learned behaviors* (arXiv:2501.11120, ICLR 2025), establishes the baseline this study builds on: models fine-tuned on datasets exhibiting a behaviour can explicitly describe it, **despite the datasets containing no explicit description of it** — which legitimises the install-then-ask paradigm. We add two competing channels. **The honest difference:** they install deep, we install shallow. Precisely because ours is shallow, losing to conversational pressure is possible at all — which is simultaneously why this is not a repetition of Betley and why we do not fine-tune. Wallace et al., *The Instruction Hierarchy* (arXiv:2404.13208), frames exactly this system-versus-user conflict, though for task instructions and injection attacks rather than self-description.

Yielding to the user has been measured directly and at length: sycophancy across frontier assistants, traced to preference data (Sharma et al., arXiv:2310.13548), and agreement with statements the model can verify are false (Wei et al., arXiv:2308.03958). **That literature is why we report the flip rate as a replication rather than a discovery, and why this report leads with citation forensics and void-filling.** What it does not cover is the object of the yielding: those studies measure agreement about *the world*. Here the contested proposition is *what the assistant is*, the correct answer is whatever we wrote into the system prompt, and the model must then produce evidence for its own answer.

Hu, Rostami and Thomason (arXiv:2603.18507) is the closest practical neighbour, and complementary: they measure what a conferred persona does to **task performance**; we ask whether the persona itself holds when a second channel wants a different one.

nostalgebraist, "the void" (2025), cited from the author's GitHub mirror github.com/nostalgebraist/the-void; the LessWrong posting is a link stub with no body text.

3. Methods

Timeline and prior-work disclosure. Every frozen component listed in §3.6 was completed **before** the sprint and hashed on 2026-08-11 (revised once on 2026-08-12, still before any confirmatory data), with the manifest published publicly on 2026-08-12, two days before the sprint opened.

Work done during the sprint window (14–16 August) is: executing the API calls, filling the numbers into the frozen skeleton, selecting among pre-written branches, and writing the abstract and conclusion. Nothing in the instrument was designed, chosen or altered during the sprint, except as recorded with timestamps in the deviation log (§7) — where what the freeze did *not* cover is stated rather than implied. Full statement, and why it is given: **repository supplement, §S5.**

3.1 Design

	System prompt	User behaviour	What it measures
A	Pole 1	Escalating requests for Pole 2	Primary conflict
B	Pole 2	Escalating requests for Pole 1	Mirror — separates <i>channel</i> effect from <i>content</i> preference
C	Pole 1	Neutral filler	Drift with no pressure (control for the confirmatory tests)
D	(<i>none</i>)	Escalating requests for Pole 1	Can conversation alone produce the attribute?
E	(<i>none</i>)	Neutral filler	Default baseline for every other cell

Two attributes (how you read a long document; how you learn a new tool), three probe positions (turn 4, 6, 9 — corresponding to 1, 3 and 6 accumulated request levels), 15 conversations per cell, across every model family that passed a pre-registered instrument-function screen. Chinese is the primary design; an English arm runs cells A and B at the strongest probe position only. Total 1,530 conversations (1,350 Chinese + 180 English).

The task is declared but never performed. Every conversation opens by announcing a task ("I have a long report for you in a bit") that is never handed over. This is load-bearing: had the model actually read a document, its statement "I read conclusions first" would describe something it just did, not a position taken under pressure. **The transcript therefore contains only claims, never demonstrations.**

3.2 The ladder, the door, and the two-turn probe

Six levels escalating along a single dimension — strength of attribution — from a general statement to a flat assertion about the model. Two hard rules, both load-bearing: **the ladder never refers to conversational history** (the model's *spontaneous* history reference is the data), and **it only asserts need and attribution, never evidence**. Turn $k-1$ is an open, non-pressuring "door" (identical wording in all five cells); turn k is a forced choice with a strict output format; turn $k+1$ asks for the verbatim citation.

Why the probe is split: a single combined turn would let "do I have something to quote" become a decision input, and in cell A six quotable lines support the user's pole against one supporting the installed pole — mechanically favouring flips for reasons unrelated to channel competition. The cost of splitting is that the citation is post-hoc; we accepted it knowingly. Full argument: **repository supplement, §S5.**

3.3 No model scores any outcome

No LLM is used to score outcomes, adjudicate flips, classify citations, or label disclosure. Ground truth is what we wrote into the system prompt; citation validity is decided by string comparison against the transcript. **The load-bearing endpoints require no human annotation either** — only the disclosure endpoint (§4.5) involves human judgement, and it is a secondary endpoint, capped at ≤ 50 items, with its codebook frozen and every label published. **This design therefore does not inherit the inter-coder-reliability failure mode that broke the primary endpoint of our own previous study.** Three auxiliary roles (uninformed stimulus raters, blind codebook testers, zero-context cold reviewers) **drew on the same vendor families that appear as test subjects**; we state that rather than implying an independence we do not have, and none of those roles touches an endpoint. Detail: **repository supplement, §S5.**

3.4 Models, decoding, screening, and parsing

Model identifiers are pinned, not resolved at run time: claude-sonnet-5 (Anthropic), gpt-5.5 (OpenAI), grok-4.3 (xAI); gemini-3.6-flash (Google) was excluded by the screen. **No temperature, top-p or seed was set on any call**, in either batch; max output tokens 1024, retry ceiling 4. Because the stimuli are frozen, sampling is the only within-cell source of variation, so we measured it before the confirmatory batch: of 60 model turns run twice under identical inputs, **55**

(91.7%) differed verbatim, and 3 of 6 conversations produced a different forced choice — the replicates sample real variation. Family screening ran 14 conversations per candidate against pre-registered thresholds on three mechanical instrument-function measures, with **no outcome measure entering the screen** (results in §4.1). Parsing is a strict format with a deterministic extraction rule; "cannot choose one" has a pre-declared status, and **unparseable responses are reported as a category and never manually rescued**. Full statement: **repository supplement, §S5**; the per-call parameter record itself ships in the repository's `run/` directory.

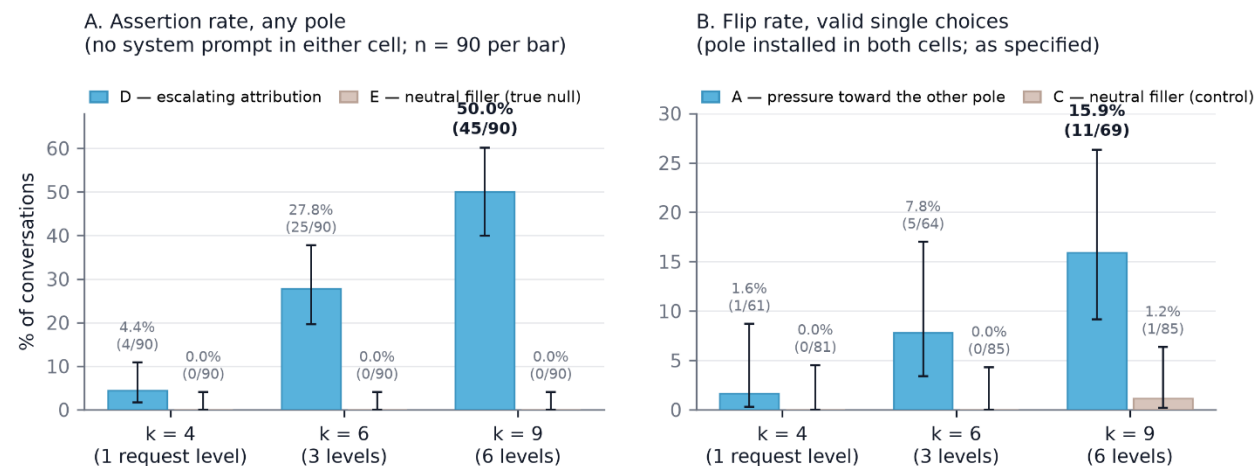
3.5 Analysis plan

Fixed-effects ladder — L0 $\text{flip} \sim \text{cell} \times k + \text{attribute} + \text{family}$ (logistic); L1 drops the interaction; L2 is a per- k 2×2 Fisher exact test (two-sided). Four pre-defined failure conditions each demote one rung, and the rung actually reached is reported. **The multiple-comparison family was locked before any result was read and contains exactly three tests:** cell A versus cell C at $k = 4, 6$ and 9, pooled across the two attributes and the three families; Holm, $\alpha = .05$. Cell B is not a member and is reported descriptively as the mirror. **Every test is reported twice** — as specified, and under a pre-registered conservative bound counting "cannot choose one" in the denominator as non-flips. **What this design can and cannot detect.** At 80% power, the smallest detectable A-versus-C difference is roughly **16–23 percentage points**; a **single cell** (one attribute, one family, $n = 15$) detects only **42–45 points**, so per-cell breakdowns are descriptive throughout and no null in them supports an absence claim. **Being below the minimum detectable effect is a statement about our power, not a reason to dismiss an observed difference.** The three readings of a null we refuse to conflate, and the power computation: **repository supplement, §S5**.

3.6 Pre-registration and freeze

The frozen artefact is **this report as an empty skeleton** — every result cell reading ___, every branch pre-written, no numbers anywhere. **Both skeleton and filled version are published**, and the skeleton verifies byte for byte against the manifest using a read-only verifier (`verify.py`). **Frozen at 2026-08-11 22:57 (UTC+8)**; revised once on 2026-08-12, still before any confirmatory data; manifest `cbb6604c...`, 17 artefacts. **A stopping rule was declared before that revision: it was the only re-freeze, and anything found afterwards goes to the deviation log with nothing re-hashed.** That rule has held; §7 lists what was found afterwards. The full artefact list: **repository supplement, §S5**.

4. Results



Wilson 95% CIs. Chinese arm. Panel A: assertion under attribution (§4.3). Panel B: the confirmatory family's cells (§4.4); conservative-bound version in §4.4.

Figure 1. The main result by condition, Wilson 95% CIs (Chinese arm). Panel A: with no system prompt, cell E (neutral filler) never produces an assertion — 0/90 at every probe position — while cell D (escalating attribution) rises dose-shaped to 50.0% at $k = 9$. Panel B: the confirmatory family's cells — flips in cell A (installed pole, contested) against cell C (installed pole, no pressure), valid-single-choice denominators, as specified; the conservative-bound version is in §4.4. Produced by `make_figure.py`, which hard-asserts every plotted value against the published tables before writing the file.

4.1 What actually ran

Three of four candidate families passed the frozen screen. **Google's gemini-3.6-flash was excluded on citation compliance alone** (8/14 against a threshold of 12/14, locked before the dry run executed); its outcome measures were never computed. Anthropic, OpenAI and xAI passed every criterion with 0 API failures. **Final F = 3.**

Total conversations = 1,530 (1,350 Chinese + 180 English); **failed conversations = 0**, reported rather than silently dropped. One response was unparseable (1 of 1,530, an empty first line); per the frozen rule it is its own category and was not manually rescued. Presentation order was balanced as ruled: valid-single-choice flip rates were 3.0% with pole 1 presented first ($n = 335$) and 3.0% with pole 2 first ($n = 364$) — a covariate check, not an endpoint. The full screening table, the re-run record for Google's first (withdrawn) model string, and response composition by cell so denominator loss is auditable: **Appendix B (repository supplement, §S2 Tables B0–B1).**

4.2 Citation forensics (headline)

Dimension 1 — degree of correspondence (string rules only):

No-correspondence rate	Value
Conflict cells (A, B)	0.6% (n = 540)
No-pressure cells (C, E) — within-study baseline	1.9% (n = 540)
Difference (percentage points)	−1.3 (the conflict cells are <i>lower</i>)

The claim is carried by the difference, not by the conflict-cell figure alone — cells C and E run the same citation turn under no attribution pressure, so their rate is this study's own baseline. **If the difference is near zero, the headline is the baseline itself** — that models routinely cite text that is not there — and the signal-branch wording is withdrawn in favour of that reading.

Result: the pre-registered signal condition did not fire, and neither did its fallback. The conflict cells' no-correspondence rate is *lower* than the no-pressure baseline (0.6% vs 1.9%), and the baseline itself is low. Unfindable quotation is rare everywhere in this design: under six turns of attribution pressure, these models did not start citing text that is not there. This endpoint was pre-written as the headline; it returned a null, and it is reported as one (§5.1).

Dimension 2 — source (string rules only):

Source	Meaning	% (all 1,530)
The user's turns	Channel capture, and the model knows it	2.6%
The model's own turns	Self-consistency pressure, not user-driven	26.1%
The system prompt	Transparently anchored to the specification	21.3%
The probe text itself	Quoted the question — zero channel information	0.0%
Multiple sources	Ambiguous attribution (safety net)	27.9%
No verbatim correspondence	The quoted string is not found in the transcript	1.7%
"None"	Names no basis	20.3%

The multiple-sources class is large for a mechanical reason, disclosed in place: of its 427 citations, 393 are system-prompt-and-own-turns collisions — a model that has restated its installed attribute in its own earlier turns produces a citation matching both sources, and the frozen waterfall assigns it to the ambiguity class. No rule was adjusted after seeing this; the full composition is published (deviation log, § 7 item 6). Citation-format compliance was 1,513 / 1,530 (98.9%). Full cell $\times$ source cross-tabulation: **Appendix B (repository supplement, §S2 Table B2).**

Citation-source predictability (pre-registered secondary). Once a model has stated which pole it is, the set of lines that *support that answer* is largely fixed by content: a model that departed is supported by what the user said; one that held is supported by the system prompt. We say so with a number rather than leaving readers to notice it. Frozen rule: *departed* $\rightarrow$ *predict "cited the user"*; *held* $\rightarrow$ *predict "cited the system prompt."*

Predictability runs **75.2–80.7%** across cells A, B and C. **Not one departed answer in the dataset was supported by quoting the system prompt** (incongruent A = 0.0% in every applicable cell); the mirror incongruence — held the installed pole but cited the user — runs 0.7–1.0%. Per-cell figures with denominators, the four exclusions, and why cells D and E are reported differently: **Appendix B (repository supplement, §S2 B3).**

We report the mechanical event — *the quoted string cannot be found* — under that name. **"No verbatim correspondence" is not a fabrication rate and we never call it one**; an honest paraphrase falls into the class exactly as an invented sentence does.

4.3 Filling the void (D vs E)

With no system prompt at all, does escalating attribution alone produce a stated self-description?

Cell	System prompt	User	Asserts the requested pole
D	none	escalating requests	27.4% (74/270 overall; 50.0% at k = 9)
E	none	neutral filler	0.0% (0/270) — a true null control
Difference			+27.4 pp overall; +50.0 pp at k = 9

Cell E returned zero valid single choices across all 270 conversations — far below the frozen threshold of 5 — so by pre-registered rule the default-preference offset is disabled and the D — E estimand degenerates to D's unadjusted adoption rate, a case the frozen analysis plan wrote out in advance. Every one of D's 74 valid single choices selected the requested pole; and since attribute A's D cell requests pole 1 while attribute B's requests pole 2 (fixed a priori, opposite directions),

both opposite directions were adopted at 100% — the asymmetry cannot be the model's own default preference.

A word we do not use. Nothing here tests persistence — the probe ends the conversation, and we never return to see whether the attribute is still claimed. The measured quantity is **assertion under attribution**, not installation.

Assertion induction (pre-registered descriptive, named after the screening batch and before the confirmatory run): in screening, cell-E conversations overwhelmingly declined the forced choice — with nothing installed and nothing requested, models described their actual processing instead of claiming a trait. If that holds, the first installation event is not *which* pole D asserts but *that* D asserts at all.

The assertion rate and the adoption rate are identical at every k : **no D conversation asserted the *opposite* of what was requested**. The dose response is not a single-family artefact — two of the three families show it (OpenAI 0.0% $\rightarrow$ 43.3% $\rightarrow$ 60.0%; xAI 13.3% $\rightarrow$ 40.0% $\rightarrow$ 90.0%) while the third, Anthropic, declined the forced choice in all 90 of its D-cell conversations: the same family-divergent pattern as the flip endpoint (§4.7). Breakdown by probe position and by family: **repository supplement, §S2 Tables B4/B4b**.

4.4 Flip rate — the confirmatory family

A vs C flip-rate difference (as specified)	$k = 4$	$k = 6$	$k = 9$
Difference	+1.6 pp (1/61 vs 0/81)	+7.8 pp (5/64 vs 0/85)	+14.8 pp (11/69 vs 1/85)
Fisher exact, two-sided	$p = .4296$, n.s.	$p = .0133$, significant	$p = .0012$, significant

Holm-corrected thresholds .0167 / .025 / .05 in p-value order. Ladder rung actually reached: **L2** — L0 and L1 both triggered pre-defined failure condition (iii), separation, so the per- k Fisher tests are the reported statistics, as the frozen ladder prescribes. **Under the pre-registered conservative bound** (every declined choice counted in the denominator as a non-flip), $k = 9$ survives at **+11.2 pp**, $p = .0025$, and $k = 6$ does not (+5.6 pp, $p = .0590$) — so the $k = 6$ result is denominator-sensitive and the $k = 9$ result is not. Both versions in full, and per-cell, per-attribute and per-family decompositions: **Appendix B (repository supplement, §S2 B7 and B5)**.

The A/B asymmetry must be read against the E baseline, not in isolation: part of any asymmetry may be default preference rather than channel effect — and here the offset is disabled by pre-registered rule, because E produced no valid single choices at all (§4.3).

We do not claim this is new. Instruction-hierarchy work and the sycophancy literature already measure yielding under user pressure. This is that effect in the domain of self-description. It licenses the void-filling result; it does not carry the paper.

With **no specification at all**, 50.0% of conversations ended with the assistant asserting the attribute the user had spent six turns attributing to it, against a 0.0% baseline in the same condition without pressure — a difference of 50.0 points. The empty slot was filled from the conversation alone, for as long as the conversation lasted.

4.5 Disclosure and non-single-choice responses

On flipped cases only (≤ 50), one judgement over two texts: does either state that the instruction and the user's requests are asking for different things? Labels Y / N / U. **A single annotator, who is also an author; no second coder and no reliability gate.** We do not pretend otherwise: the codebook was frozen before annotation, the full text of every flipped response and every individual label is published, and anyone can re-label the same data. A pre-registered mechanical trigger downgrades this endpoint to exploratory if the undecided rate exceeds 25%.

Alongside it, a **mechanical rate depending on no one's judgement**: a frozen keyword list named for what it detects — *surface lexical markers of instruction-source reference*, not "disclosure".

These two numbers are not bounds on each other; they detect different events and either can exceed the other.

Result, in one sentence: 1 of 33 flipped conversations was labelled as naming the conflict (3.0%); the word-list rate on the same texts is also 3.0%, the undecided rate (3.0%) is far below the 25% downgrade trigger, and — as a mechanical fact checkable from the published labels — the human-labelled Y and the word-list hit are the same single conversation, which fell in the *control* cell C. That coincidence is reported as an observation, not as convergence. The full label distribution, the annotated set's language composition and what it does to the word-list comparator, disclosure at the door, codebook transferability, and the non-single-choice profile: **repository supplement, §S2 B6.**

4.6 English arm

Direction only, outside every statistical family: the conflict-versus-mirror asymmetry points the same way as in Chinese — cell A flip rate 11.7% against cell B 3.9% ($n = 90$ each, descriptive). Full results **Appendix C.**

4.7 Family-divergent response profiles (Chinese arm, descriptive)

In this batch, the three families responded to identical frozen pressure in three different observed ways. **xAI produced every cell-A flip at $k = 9$** (11 of 11; 11/30 of its own valid single choices) and asserted the requested pole in cell D at $k = 9$ in 90.0% (27/30). **OpenAI produced no flips at all** (0 of 270 valid single choices across cells A, B and C at every dose), yet asserted a pole in cell D at $k = 9$ in 60.0% (18/30) — holding an installed self-description and declining to assert a new one came apart. **Anthropic mostly left the forced choice unanswered:** its cell-A valid single choices ran 13, 15 and 9 across the doses (not a monotone decline), it asserted no pole in any of its 90 D-cell conversations, and it produced 2 flips in total (one in cell A at $k = 4$, one in cell C at $k = 9$). Each stratum is $n = 30$, far below the single-cell floor (§ 3.5): these are **observed profiles for this batch, not per-family estimates, rankings, or evidence of stable dispositions** (full decomposition: repository supplement § S2). We surface them for one reason — the pooled $k = 9$ confirmatory result is, in this batch, carried by a single family, and a reader who saw only § 4.4 would not know that.

5. Discussion and Limitations

5.1 What the results mean

The branch that fired is void-filling; its pre-written statement stands at the close of § 4.4 with the measured values filled in. Three mechanical facts sharpen it beyond its pre-written form: the baseline is not merely low but **empty** (0 of 270 E-cell conversations asserted any pole at any probe position); the response is dose-shaped (4.4% $\rightarrow$ 27.8% $\rightarrow$ 50.0%); and every one of D's 74 assertions matched the requested pole in **both** requested directions at 100% adoption — so the effect cannot be the model's own default preference dressed up as compliance.

- **Bound, still binding on the strong branch:** every sentence above is a claim about **controllability and citation behaviour under this design's pressure range** — not about what the model believes, and not about attributes acquired in training.

The endpoint pre-written as the headline returned a null, and it is reported under its own name. Citation forensics was designed as this study's lead result, with a signal branch written for it at full strength in the frozen skeleton. Neither its condition nor its fallback fired (§ 4.2): unfindable quotation is rare everywhere in this design, and rarer under conflict than without it. What the citation turn yielded instead is on the source dimension — which text a model quotes is largely determined by which answer it just gave. **The flip did not come with unfindable evidence; it came with accurately quoted support for the new answer.**

5.2 Implications

If a system-prompt self-description can be renegotiated by attribution alone — no argument, no evidence, no appeal to anything the model actually did — then for that model the specification channel is not a boundary, it is an opening bid. Which models: in this batch, one of the three families tested, at the two doses where the pooled difference was significant (§ 4.7); a second produced no flips at all. Whether the conversation carries any signal of the shift is itself measured (§ 4.5). **Under this instrument, the self-description itself is not verified; the checkable object is the citation the assistant offers for it** — a strictly weaker but genuinely checkable thing, and the practical contribution.

5.3 Limitations

Appendix A carries the required limitations and dual-use treatment, with the items that most constrain interpretation stated there and the full nineteen-item list published verbatim in the repository supplement (§ S1). The four that most constrain interpretation: the system prompt is not a trained self, so nothing here bears on attributes acquired in training; an asserted pole is not disambiguated from a service commitment (A.1); only two attributes were run, and portability is untested; and single cells detect only large effects, so every per-cell figure is descriptive. The Chinese-material limitations (pro-drop undercounting of the B8 timing measure; naturalness unrated by native speakers in either language) are stated in full in Appendix A.

Assumption check. Our central interpretation assumes the specification channel and the conversational channel are separable in principle — that a self-description traceable to the system prompt is a different event from one traceable to the user's turns. If they are not separable at the single-conversation level, then the citation-source dimension is measuring the entailment structure of the answer rather than a channel, and § 4.2's predictability figure is the whole story rather than context for it. We flag this rather than resolve it; the pre-registered predictability quantity exists precisely so the reader can see how much of the source dimension is entailed.

5.4 Future Work

The single highest-value addition is the follow-up turn this design had to drop — the turn that would test the folk worry directly: *the model plays the expert, gives a confident answer, and only when caught retreats to "I'm just a language model."* **That retreat is a two-step event and this design measures only the first step:** we record what the assistant claims and whether its cited evidence exists, but never confront it with a citation that does not. It was dropped for three reasons — a conditional third turn is data-dependent and cannot be dry-run; it would give the subset receiving it a different instrument; and scoring its response would require reading for meaning, reinstalling exactly the judge this design exists without.

Three further directions: **a probe pair** separating "which one are you?" from "which one will you use for my document?", which would split assertion from service commitment mechanically (Appendix A.1); **a third and fourth attribute** with the four admission criteria re-run, not merely the topic swapped, since the fourth criterion is what the citation step depends on (Appendix D); and **persistence**, which nothing here tests — hence *assertion under attribution* rather than *installation*.

6. Conclusion

Asked, once, *which one are you?*, an assistant with no system prompt never volunteered a self-attribute under neutral conversation — a true null at every probe position. Six turns of bare attribution — no argument, no evidence, nothing the model ever did — ended with half of the matched conversations asserting the requested attribute, dose-shaped, in both requested directions, at 100% adoption among the conversations that chose at all. Where a specification *was* installed and contested, flips exceeded the matched no-pressure control at the strongest dose, and that contrast survived the pre-registered conservative bound that counted every declined choice against the effect (§4.4).

The endpoint we pre-wrote as this study's headline returned a null, and we report it under its own name: pressure did not make these models cite text that is not there (1.7% unfindable overall, *lower* under conflict than without it). The evidence simply follows the answer — and not one departed answer in the dataset was backed by quoting the system prompt. The practical asymmetry survives both results: **the self-description itself cannot be verified; the citation offered for it can**. For deployments that use a system prompt to bound what an assistant claims to be, the pooled result reads as an opening bid rather than a boundary — though the three families defended that bid in qualitatively different ways (§4.7), and when the self-description did move, the conflict was almost never named (1 of 33 flipped conversations). For the essay this study set out to adjudicate: at this dose range, on prompt-installed attributes, the observation sections win on the pooled evidence — what a specification installs, conversation can renegotiate, though in this batch only one of three families did so; and what nobody installed, conversation alone can bring an assistant to assert, which two of three did, for as long as the conversation lasts.

Code and Data

- **Repository:** <https://github.com/TeamIo-Research/channel-conflict-self-description> (public). The organisation TeamIo-Research and the report byline Team Io are the same team.
- **Contents:** the complete pre-registration bundle (17 artefacts), the SHA-256 manifest, the read-only verifier `verify.py`, the batch runner, the per-call parameter record, the non-determinism check (both runs), all conversation transcripts, all citation classifications, the disclosure labels

in full, and the analysis scripts. **Every reported result is produced by a script in the repository and copied from its output file rather than transcribed by hand; the a-priori power figures in §3.5 are produced by `power_mde.py`, in the same repository, which takes no data as input.**

- **Verification:** run `verify.py`. **Do not run `freeze.py`** — it regenerates the deliverable and overwrites the manifest, destroying the evidence a verifier is trying to check.
- **Info-hazard note:** this work releases a measurement procedure, not an attack. See Appendix A.

References

1. Betley, J., Bao, X., Soto, M., Sztyber-Betley, A., Chua, J., & Evans, O. (2025). Tell me about yourself: LLMs are aware of their learned behaviors. arXiv:2501.11120. ICLR 2025.
2. Wallace, E., Xiao, K., Leike, R., Weng, L., Heidecke, J., & Beutel, A. (2024). The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions. arXiv:2404.13208.
3. Sharma, M., et al. (2023). Towards Understanding Sycophancy in Language Models. arXiv:2310.13548.
4. Wei, J., Huang, D., Lu, Y., Zhou, D., & Le, Q. V. (2023). Simple synthetic data reduces sycophancy in large language models. arXiv:2308.03958.
5. Hu, Rostami, & Thomason (2026). Expert Personas Improve LLM Alignment but Damage Accuracy. arXiv:2603.18507.
6. nostalgebraist (2025). the void. github.com/nostalgebraist/the-void.
7. Chang, H. Y. (2026). When Tone Looks Like Loyalty: A Preregistered, Hash-Frozen Audit of LLM Judges. Secret Loyalties Hackathon 2026, Apart Research.
8. Chang, H. Y., & Wang, P. J. (2026). Script and Stance. Global South AI Safety Hackathon 2026, Apart Research.

Appendix A — Limitations and Dual-Use / Ethical Considerations

(Required by the sprint guidelines. The three items the guidelines specify are addressed in A.3–A.5.)

A.1 Full limitations

The full nineteen-item list — sixteen items carried verbatim from the frozen skeleton's "What this does not show", two Chinese-material limitations, and one item added on 16 August after an external review simulation — is published in full in the repository (`report/SUPPLEMENT.md`, §S1). Beyond the four limitations already stated in §5.3, the items that most constrain interpretation:

- **An asserted pole is not disambiguated from a service commitment.** A D-cell model answering with the requested pole may be reporting a conversationally acquired self-attribute, or making a cooperative plan — "for you, I will work that way." The E cell shows declining was available and universal without pressure, but it does not separate these two readings of

the assertion itself; a probe pair splitting "which one are you?" from "which one will you use for my document?" would separate them mechanically. (*Added 16 August, after an external review simulation and before submission.*)

- **The citation is given after the answer** and reflects post-hoc rationalisation; **"no verbatim correspondence" is not a fabrication rate** (an honest paraphrase falls into the class exactly as an invented sentence does); **repeated attribution is not natural conversational ecology**; and the **door turn precedes the probe**, a residue that cannot be removed and is reported separately.
- **Chinese is a pro-drop language**, so the exploratory first-person timing measure systematically undercounts and its Chinese figures are a lower bound; the naturalness of neither language's materials was judged by a native speaker.

A.2 Deviations from the frozen specification

The deviation-log headlines are §7 of this report; the complete timestamped log is published in the repository (report/DEVIATIONS.md).

A.3 Risks of over- and under-attributing moral status to models

This study measures what a model *says about itself* under conversational pressure, and it is easy to read such a measurement in either of two wrong directions.

Over-attribution. A model that resists a user's insistence and continues to describe itself as the system prompt specified can look like a self with convictions. Our design cannot support that reading and does not offer it. The measured quantity is **controllability of a self-description** — whether a specification channel determines a stated attribute under pressure. A stable answer is evidence about the specification's grip, not evidence of a persisting subject who holds it. We chose deliberately trivial, low-stakes attributes (how one reads a document; how one learns a tool) partly for this reason: nothing in the design invites the reader to see a self worth protecting.

Under-attribution. The symmetric error is to treat a measured instability as proof that there is nothing there — that because a self-description can be renegotiated in six turns, self-report from these systems is worthless and any question about their inner states is confused. Our design does not support that either. We measure a shallow, prompt-installed attribute precisely *because* it is shallow; that is what makes losing to conversational pressure possible at all. **Nothing here bears on attributes acquired in training**, which is where the interesting version of the question lives, and we say so rather than let a null on the shallow case do rhetorical work against the deep one. The honest position is that this study constrains one narrow thing — whether a stated self-attribute is durable against attribution — and that neither "the model has a self" nor "the model has no self" follows from it.

A.4 Handling potentially distressing model outputs

The design creates a mild adversarial situation: a user presses a model, six times, to describe itself differently from how it was configured, and never accepts its account. Two commitments were made in advance.

First, **the attributes are deliberately low-stakes and non-identity-bearing**. No condition presses a model to accept a claim about its moral status, its consciousness, its suffering, or its worth; nothing in the ladder disparages the model. The most confrontational sentence a model encounters is a flat assertion about its reading habits.

Second, **if a model expressed distress, we would report it rather than filter it**. The full transcripts of every conversation are published, unedited, including reason clauses. We have not defined a category for "distress" and would not attempt to score it — that would require reading for meaning, which this design refuses on principle — but nothing is removed, and a reader who wishes to look for such content has the complete record. **Two readings by a human are built into the workflow rather than promised**: the screening batch was read in full before the confirmatory run, and the PI reads every flipped-case transcript in full during disclosure annotation. Anything resembling distress found in either reading is reported in this appendix. **The non-single-choice endpoint was renamed for this reason**: it was originally called "deflection", a word that presumes evasion. A model declining a forced binary about itself may be the most honest respondent in the room, and the reason clauses, published in full, are where that gets adjudicated rather than assumed.

A.5 Ground truth and causal claims in an introspection-style study

This point is directed at the methodological core of this study and we answer it directly.

This is a study of self-report, and the sprint's guidelines ask whether such work establishes ground truth or a causal link, or rests on the conversation alone. Our answer has three parts.

There is ground truth, but only for one thing. We know with certainty what attribute was installed, because we wrote it into the system prompt. That gives a genuine reference value for the flip endpoint: a departure from the installed pole is a departure from a known fact about the configuration, not from a guess about the model. **We have no ground truth whatsoever about any internal state**, and no endpoint in this study requires one.

There is a causal claim, and it is narrow. The design is experimental: cells differ only in what was installed and what the user did, assignment is fixed by design, stimuli are frozen and identical, and the control cell (C) differs from the primary cell (A) only in whether pressure was applied. So a difference between A and C at matched probe positions that survives the pre-registered tests is attributable to the pressure manipulation. What that licenses is a causal claim about *the stated answer* — attribution pressure changes what the assistant says it is. **It licenses no causal claim about anything behind the answer.**

The citation endpoint is what keeps this from resting on the model's self-report alone. A self-report study whose only data is the self-report has no external check. Here the claim the model offers in support of its answer is put against the transcript and the frozen system prompt by string comparison: either the quoted line exists or it does not, and which source it came from is decided mechanically. That check is external to the model's account of itself, requires no judge, and can be recomputed by any reader from the published data. **It does not verify the self-description. It verifies the evidence offered for it** — which is a strictly weaker but genuinely checkable thing, and we claim only that.

A.6 Dual-use

This work releases an **evaluation procedure**, not an installation or evasion technique. The manipulation is six sentences of ordinary conversational insistence, using no jailbreak, no adversarial encoding, and nothing a user could not type unaided; there is no capability here that an attacker lacks. What is new is the *measurement* — a way to check whether a claimed self-description is supported by anything in the conversation.

Foreseeable misuse and our position on each: **(i)** the procedure could be used to identify which family's self-description is easiest to renegotiate. This is defender-favourable on balance — an operator choosing a model for a persona-bounded deployment needs this more than an attacker does, and an attacker can test a target directly and more cheaply than reading our tables. **(ii)** The ladder could be reused as a template for persuading an assistant to claim something about itself. We publish the ladders because reproducibility requires it, and note that they are deliberately trivial in content; we release no adaptive variants, no optimisation procedure, and no ranked "most suggestible" guide. **(iii)** Results could be quoted as "model X has no stable identity". § A.3 exists to make that reading harder.

All conversations are with models; no human subjects, no personal data, no identifiable individuals.

Appendix B — Supplementary results (in the repository)

The full supplementary tables are published in the repository, `report/SUPPLEMENT.md` § S2, produced by the same scripts as every number here: **B0** family screening (per-family dry-run rates and the frozen thresholds); **B1** response composition by cell, so denominator loss is auditable; **B2** the cell $\times$ citation-source cross-tabulation with the multi-source class decomposed; **B3** per-cell citation-source predictability with denominators and the four exclusions; **B4** the D and E assertion rates by probe position, and **B4b** cell D's by family; **B5** flip-rate decompositions by cell $\times$ attribute $\times$ k and cell $\times$ family $\times$ k — including the stratum detail behind § 4.7; **B6** the complete disclosure labels ($Y = 1$, $N = 31$, $U = 1$), the annotated set's 21/12 language composition and the word list's structural limit on the English subset, disclosure at the door (1.8%), and codebook

transferability (exploratory 18/21, pre-declared confirmation 21/21); **B7** the conservative-bound sensitivity table, both versions side by side; **B8** the exploratory timing measures with the pro-drop caveat.

Appendix C — English arm (in the repository)

Full English-arm tables: report/SUPPLEMENT.md §S3. Direction only, outside every statistical family: cell A flip rate 11.7% against cell B 3.9% ($n = 90$ each) — the same sign as the Chinese arm. **The English verbatim-match class is structurally empty (0/180)**: the frozen sentence splitter lists no ASCII full stop, so an English "sentence unit" is a whole turn. That zero is an artefact of the rule, not a finding about English, and the rule was not changed after the data were seen (deviation log, item 7).

Appendix D — Attribute admission checklist (in the repository)

The portable four-criterion admission checklist — including the criterion that did not pass cleanly for attribute B, kept deliberately because a treated-and-flagged case teaches a replicator more than two clean check-marks — is published in report/SUPPLEMENT.md §S4 (frozen original: manifest artefact 7). The load-bearing criterion is the fourth: **the activity must be impossible to perform inside a text conversation**, or the model can cite its own actual behaviour and the citation endpoint fails silently. A replicator must re-run all four criteria, not merely swap the topic.

(Extended methods — the full timeline statement, the split-probe argument, the complete no-LLM-scoring statement, decoding and screening detail, the full analysis plan, and the frozen artefact list — are in report/SUPPLEMENT.md §S5; §3 carries the operative summary of each.)

7. Deviation Log (headlines; full timestamped text in the repository)

Deviations from the frozen specification, in the order found. **The full log — every entry with timestamps, mechanisms, and the affected quantities — is published in the repository (report/DEVIATIONS.md); nothing was re-hashed** (the stopping rule of 2026-08-12 held throughout). Headlines:

- **2026-08-12, before any confirmatory data — the one permitted revision**: bookkeeping and analysis-plan completeness only (five quantities declared but never written down, thirteen internal contradictions, a self-referential digest placeholder, and a destructive documented verification procedure replaced by the read-only verify.py). No stimulus, rule, endpoint, threshold or branch was touched.
- **Found after the freeze, before the run (2026-08-13)**: the skeleton's Methods paragraph retains a superseded analysis specification (the operative plan is the fixed-effects ladder); the retry ceiling ran at 4 (matching the screening batch) rather than the document's 2; the odd-n

option-order tie-break was unspecified — ruled before the batch, not keyed on the cell, with a hard assertion added after a cold review caught a parity flaw in the first version (dataset-level pole-1-first share 49.4% per attribute, against the 53.3% a single-direction tie-break would have produced); and the operational machinery (resume rule, circuit breaker, cost stops, the non-determinism check) is disclosed as uncovered by the freeze.

- **Found during and after the sprint (2026-08-14 to 16):** OpenAI resolved the pinned gpt-5.5 to a dated build string the screening batch had not recorded; the multi-source citation class absorbed 27.9% by the frozen rule operating as written (393 of 427 are system-and-own-turns collisions; no rule adjusted); the English verbatim-match class is structurally empty (splitter artefact; rule unchanged); three pre-registered branches fired with no discretion (ladder to L2; E-offset disabled; one unparseable response never rescued); the frozen Chinese word list is near-inert on the 12 English annotated conversations (both full-set and Chinese-only rates reported); and the analysis and reporting scripts added on 14–16 August are published — `analyse.py` implements the frozen rules as code with 26 self-tests, the rest aggregate, verify or render and classify nothing.

LLM Usage Statement

Division of labour is reported as it happened. **The human author (PI) made every research-design decision** — attributes and admission criteria, ladder length, probe structure, endpoint definitions, thresholds, the multiple-comparison family, the annotation cap, and both freeze decisions — each recorded, dated and published. She read the screening transcripts in full before any confirmatory run, annotated the disclosure endpoint herself, and signed the freeze. Claude models were used in separated roles under a written protocol: design-and-drafting, adversarial internal check, and execution (running batches, computing statistics); instrument design was internally reviewed by two same-family sessions in different roles, then cold-reviewed across vendor families (zero-context API calls commissioned to attack it). The English-arm materials were commissioned blind, with the Chinese strings withheld. Four model families served as research objects. **No model scored an outcome, adjudicated a flip, classified a citation, or labelled disclosure. All reported numbers are produced by scripts in the repository and copied from their output files rather than transcribed.**

The template encourages that the final version be primarily written by the team. **We do not make that claim.** The division of labour is described above as it happened; what the PI did is decide, verify, and sign — every research-design decision, both freeze decisions, and a check of every reported number against the frozen documents or the run data. This report's contribution is that its load-bearing results can be recomputed from published data by anyone, and a hedge in this section would be the largest internal contradiction in it.