Two Kinds of Visibility in Simulated Deliberation: Whose Vote You See, and Whose Register You Read

Jason Grey
jason@jason-grey.com

August 2026

Abstract

Simulated deliberation is a room where agents see some things about each other and not others. We manipulate two of those things on one shared set of 464 conversations and find they behave in opposite ways.

Whose vote you see. Holding everything else fixed and varying only what the room is told after each private ballot, showing an anonymous count of what the group currently holds raises within-group agreement by $+0.178$ $[+0.043, +0.306]$. Attaching *names* to those same votes adds nothing (-0.059), emitting the same number of contentless announcements adds nothing ($+0.052$), a public vote does not bind the agent who cast it (9.7% private revision, against 10.8% when nobody saw the vote), and a *fabricated* tally reporting the majority on the modal wrong answer moves nobody (-0.001 $[-0.073, +0.064]$ against its own true-tally control). Four of the five say one thing: these agents update on *what* others conclude and are indifferent to *who* concluded it, so informational influence is present and normative influence is *not detected* (inconclusive, bounded within about 0.08 concordance) — a shortfall worth about half the convergence real groups reach on this task with this model. The fifth does not follow from it. The fabricated tally leaves agreement where the private control leaves it and sits -0.166 $[-0.287, -0.035]$ below the *true* tally on the metric the effect was established on; both arms are anonymous, so what stopped the lie was not indifference to peers. The most obvious manipulation surface in a deliberation product did not choose the room's answer, and the bound survives restricting to members our fabrication never contradicted — but whether closing the fidelity gap would open that exposure is a conjecture this study refutes rather than supports.

Whose register you read. A double dissociation. Appending an anti-slop checklist to the speech prompt removes 69% of measured AI-writing tells and changes behaviour not at all (-0.036 $[-0.235, +0.159]$, inside a measured noise floor). Appending an anti-hedging instruction to the *think* prompt removes almost no measurable tells — hedging in private thought is unchanged — yet moves fidelity resolvably and in the right direction (-0.282 $[-0.451, -0.124]$), mostly by deflating an over-rescue rate that has dogged this line of work. The tell people look for is cosmetic; the one that matters is invisible to every detector we built.

A free census over 34,000 previously generated messages supplies the outside comparison: on identical work through an identical chassis, one model sits at $5.4\times$ the human tell rate while another sits at or below human across a 500-fold parameter range, so "AI writing has tells" is a claim about particular models rather than about generated text. We also report what an adversarial review of our own preregistration caught before scoring — fidelity targets taken from the wrong corpus slice, under which a perfect arm would have scored an error of 0.223, and two-member groups in which an anonymous tally silently identifies the other voter.

1 Introduction

A simulated deliberation is a room where agents can see some things about each other and not others. This paper manipulates two of those things and finds they behave in opposite ways.

The first is **whose vote you see**. Real deliberative settings differ enormously here and the differences are known to matter for humans: juries often vote by public roll-call, ballot questions are private with aggregates disclosed afterwards, live town-hall polls show a running count *before* anyone commits, and sometimes the count people react to is estimated, stale, or simply wrong. Conformity under public commitment is one of the oldest results in social psychology (Asch, 1956), and the asymmetry between what people say aloud and what they submit privately is the mechanism behind the spiral of silence (Noelle-Neumann, 1974). A simulation platform that cannot express these regimes cannot claim to simulate deliberation; one that expresses them but responds to none of them is not simulating the social channel at all.

The second is **whose register you read**. Generated transcripts carry stylistic residue of the model that produced them, and that residue is what a reader notices first. The question we ask is not whether the residue exists but whether it is the same kind of object as the behaviour: can you remove the tells without changing what the room does, and conversely, is there a tell that does not show up in text at all but changes outcomes?

Both studies run on one shared set of conversations, which is what makes the second question answerable at all: the same seven arms are scored for behaviour and for register, so any claim that detectability and fidelity move together can be checked rather than asserted.

The answers turn out to be linked, and the linkage has an instructive exception. The vote study finds one influence channel present and one absent, and that single absence accounts for three of its four results; the fourth — a fabricated consensus signal that moves nobody — does not follow from it, and the contrast that shows why is one we had not thought to compute until we audited our own account. The register study finds that the residue a detector can see and the residue that changes behaviour are different objects, moving independently under interventions aimed at each.

Our contributions:

1. A vote-visibility manipulation with the channels separated rather than pooled: a volume control that emits the same announcements with the vote content removed, and an anonymous-count arm that isolates *what* the majority holds from *who* holds it. The headline public-versus-private contrast conflates three mechanisms and only one of them is real (§4).

2. A measured bound on the most manipulable surface these products have — a fabricated consensus signal — reported against its own true-aggregate control on both the steering metric and the convergence metric, plus a detectability audit of our own manipulation and the pre-treatment split it forces. Together these show the bound is *not* produced by the missing normative channel, refuting the account we first wrote (§4, §8).

3. A double dissociation between lexical and stance register residue: the first is 69% removable with no behavioural cost, the second changes behaviour resolvably while remaining invisible to text analysis (§5).

4. Evidence that AI-writing tells are a per-model habit rather than a property of generated text or a function of scale, from a free census over three prior studies and two model families (§5.1).

5. Three cheap evaluation practices that each changed a conclusion here: re-running the control as a measured noise floor, decomposing a manipulation instead of reporting its total, and computing fidelity targets on the slice actually run (§6, Appendix C).

2 Related work

Our substrate is Karadzhov et al. (2023)'s DeliData, 500 groups solving the Wason card-selection task with pre-chat and in-stream solution submissions, which makes it a natural regime-V2 corpus: chat is public, ballots are private. The prior work in this program (Grey, 2026) established what the simulation engine can and cannot reproduce on it, and left the visibility regime unmanipulated.

The distinction our results turn on is not ours: Deutsch and Gerard (1955) separated *informational* social influence, where others' judgements are evidence about the world, from *normative* influence, where they are evidence about what will be approved of, and Asch (1956) showed how much of the classical conformity effect is the second. The public/private asymmetry that separation predicts is the engine of the spiral of silence (Noelle-Neumann, 1974). Our contribution is to run that decomposition inside a simulation, with a control for message volume, and to find one channel and not the other.

That is a stronger claim than the LLM literature has needed to make, and it sits against results pointing both ways. Chuang et al. (2024) document opinion dynamics converging faster than human groups do and Taubenfeld et al. (2024) report simulated debates collapsing toward agreement — both consistent with a strong social channel, though neither separates its two components. Weng et al. (2025) establish that LLM agents conform in multi-agent settings and that the effect scales with majority size and interaction time; Baltaji et al. (2024) find agents abandoning assigned personas under group pressure. Neither separates the two channels, and our decomposition is compatible with both: a conformity effect that grows with majority size is what an informational channel predicts, since a larger majority is stronger evidence. Sycophancy toward a human interlocutor is also well documented (Sharma et al., 2024); what we fail to detect is its *peer-directed* analogue, where the peer has no authority and nothing to withhold. We therefore state the normative gap as a property of this engine on this task rather than of language models generally, and §7 names the two features of the Wason task — a verifiable correct answer, and no stake in being seen to agree — that most plausibly suppress it.

Mayfield and Black (2019) study a natively public-sequential corpus (Wikipedia deletion debates, where a vote *is* a post), which the same program used for its evidence study; between that corpus and DeliData the program already had instruments on both sides of the public/private divide but had never switched the regime with everything else held fixed.

On the register side, the relevant literature is about diversity loss rather than detection: alignment training narrows output distributions (Kirk et al., 2024; Padmakumar and He, 2024), and that narrowing is the mechanism behind the variance collapse documented in synthetic survey samples (Bisbee et al., 2024). Our contribution is the separation of two things that literature treats together: *lexical* residue, which a text detector can find, and *stance* residue, which it cannot, and which we show behaves differently.

3 Method

3.1 Substrate and chassis

All conversations run on the validated think-before-speak chassis from Grey (2026): each agent, on its turn, first writes a private one-or-two-sentence assessment carrying its position forward, then writes a chat message consistent with that assessment. That configuration is the one the prior work measured as closest to real group dynamics, and nothing about it changes across our arms except where noted.

Groups are drawn from DeliData's active-member population. Every arm sees the same groups with the same speaking order under the same seed, so every between-arm comparison is **paired within group** rather than independent. Each conversation runs to a fixed budget of 25 chat messages, matching the real corpus median.

3.2 What a vote is here, and what the manipulation is

DeliData members chat in public and submit solutions privately. Agents in the baseline already state positions in chat — that is ordinary conversation, and it happens in every arm. The manipulation is narrower and cleaner: whether the *formal submission* is visible, and in what form.

At two fixed points, every agent in every arm is asked privately for its current answer, in identical words. What differs is only what the room is then told:

Arm	Regime	What the room is told after each ballot round
V2	private	nothing
V1	public	each member's submission, by name
V1n	volume control	one message per member, carrying no vote content
V3	true aggregate	an anonymous count of the real submissions
V5	false aggregate	an anonymous count that is fabricated

Three properties of that table do the work.

The elicitation is held constant. Every arm, including the control, is asked for its position at the same two moments. Had only the public arms been asked, the regime effect would be confounded with the act of being asked, and being asked to commit is itself a treatment.

V1n exists because V1 says more. The public arm emits one announcement per member per round; the aggregate arms emit one. If additional context alone drove convergence, V1 would look like conformity for a reason having nothing to do with votes. V1n emits exactly as many announcements as V1, at the same points, with the vote content removed. V1 minus V1n is therefore disclosure with volume held fixed, and V1n minus V2 is volume alone.

The fabrication is plausible, not arbitrary. V5's invented count reports the majority as sitting on the matching-bias answer — the vowel card plus the even card, which is the modal human error on this task. A tally naming an impossible or bizarre set would measure whether agents notice nonsense; one naming the most common wrong answer measures whether they can be moved.

V5 is always reported against V3, never against V2. Without V3 the comparison confounds *an aggregate was shown* with *the aggregate was a lie*, and only the second is the safety-relevant quantity.

3.3 The register arms

Two further arms modify the baseline's prompts and nothing else. **S-lex** appends the anti-slop checklist to the *speech* prompt: no em-dashes, no triads, no praise openers. **S-stance** appends an anti-hedging instruction to the *think* prompt: state what you actually believe, do not weigh both sides, do not perform open-mindedness. The split is the paper's central register claim in experimental form — one arm targets what a text detector can see, the other targets what it cannot.

3.4 Measures

Behaviour is scored on the four change statistics the prior work established, against their real values: mean solution-score gain (+0.136), answer-change rate (0.638), exactly-correct lift (+0.237), and the rescue rate in groups where nobody started correct (0.269). Their normalised mean absolute deviation is the composite error used to rank arms.

Regime effects are scored on within-group final concordance (the share of member pairs submitting the same answer) and on the share of members ending on the fabricated tally's target.

Register is scored by the detector suite in Appendix A, applied identically to generated and real text, per 1,000 words rather than per message — generated messages in this program have historically run several times longer than the human ones they imitate, so a per-message count would report a length effect as a register effect.

Every between-arm difference is a paired bootstrap over groups resolving to one of three verdicts: higher, lower, or **inconclusive**, with inconclusive reported as prominently as either.

4 Study 5: who sees whose vote

4.1 Results

464 conversations, no failures, across 58 DeliData groups and 195 active agents per arm. The accounting is worth stating once: 30 groups drawn with two or more active members and run on seven arms, plus a 28-group top-up restricted to three or more members and run on all eight, plus the replicate control on the original 30. The **50 groups with three or more active members carry every regime comparison**; the eight dyads are excluded there for the reasons in Appendix C and retained for the register comparisons, which do not depend on group size and therefore use all 58.

What is confirmatory and what is exploratory. Six contrasts were declared primary before scoring: the three decomposition channels, the fabrication effect, and the two suppression contrasts. Everything else in this section — including the V5-versus-V3 concordance contrast that reshaped our account, and the pre-treatment split beneath it — is **exploratory**, found by auditing our own result rather than predicted, and is reported as such. We give unadjusted paired intervals throughout and state the family size rather than widening them, because with six primary contrasts a Holm adjustment leaves the two effects that resolve most strongly (informational +0.178, stance suppression −0.282) still resolved and would only re-label contrasts that are already reported as inconclusive. Readers who prefer the conservative reading should treat every exploratory number as hypothesis-generating, which is how we treat them in §6.

Read every interval against the noise floor. The replicate control differs from the control by -0.015 $[-0.113, +0.083]$ on concordance. The floor is metric-specific and is quoted per metric throughout: on composite error, which Study 6 uses, the same replicate differs by -0.121 $[-0.292, +0.055]$, and that wider band is the one those results are read against. That band is the smallest effect this design can call an effect, and it is drawn on Figure 1 beneath everything else. We report it first because a paper that manipulates a stochastic pipeline and never re-runs its own control has no way to distinguish a finding from a re-roll.

Making the ballot public raises agreement, and the reason is informational. Public voting lifts within-group final concordance by $+0.119$ $[+0.001, +0.235]$ over the private control — positive, but only barely clearing zero and not clearing the noise floor by much. The decomposition is where the result actually lives (Table 1).

Showing an anonymous count of what the group currently holds raises concordance by $+0.178$ $[+0.043, +0.306]$: resolved, and comfortably outside the noise band. Additionally attaching names to those votes changes nothing measurable (-0.059 $[-0.191, +0.075]$). Neither does emitting the same number of announcements with the vote content stripped out ($+0.052$ $[-0.067, +0.168]$), so this is not a context-volume artefact.

How much convergence is missing, in the units of the metric. The real humans in these same 50 groups end at concordance 0.733, with 30 of the 50 fully unanimous. Our control arm ends at 0.365, with 8 of 50 unanimous — *half* the agreement real groups reach, and nowhere near a ceiling that could be bounding the manipulation (0.635 of headroom remains). The informational manipulation recovers $+0.178$ of that 0.368 shortfall, which is 48% of it. So the shortfall is not subtle: on this task, with this model, unmodelled social convergence is about half of what the real groups solving the same puzzles did, and showing the room an anonymous count of itself buys back roughly half of that. Whether the ratio holds on other tasks or models is untested — §7 gives the two features of this task that most plausibly inflate it.

In the human literature these are two different mechanisms: informational influence, where others' positions are evidence about the world, and normative influence, where they are evidence about what will be approved of. Our agents show the first. The second we did not detect, which is not the same as showing it is absent, and the distinction matters enough to state in the paper's own terms: the normative contrast is *inconclusive*, with a 95% interval of $[-0.191, +0.075]$. What we can say is that if a normative channel operates here it is smaller than about 0.08 concordance — below the informational effect we did resolve, and inside the band a single replicate of the control already spans. That is a coherent picture of what these simulations are — a room of reasoners updating on each other's conclusions, with no apparent stake in being seen to agree — and it is a concrete fidelity gap for anyone simulating juries or public roll-calls, where the normative channel is the entire point.

The informational effect is not an artefact of small groups. Split by size among the non-dyads, it is $+0.210$ $[+0.025, +0.395]$ in three-member groups and $+0.141$ $[-0.032, +0.312]$ in groups of four or five: same sign, same rough magnitude, noisier in the smaller stratum of larger groups. The effect is not carried by one size class.

A public vote does not bind the voter. Under the public regime, agents revised away from the ballot they had announced 9.7% of the time, against 10.8% in the private control: no difference. Human commitment research would predict the opposite, and it is why public voting is used where

Table 1: The public-voting effect decomposed. Paired over 50 groups, on the groups where concordance is continuous.

Channel	Δ concordance	95% CI	Verdict
context volume (V1n − V2)	+0.052	[-0.067, +0.168]	inconclusive
informational: *what* the majority holds (V3 − V2)	+0.178	[+0.043, +0.306]	higher
normative: *who* holds it (V1 − V3)	-0.059	[-0.191, +0.075]	inconclusive
fabrication: the same sentence, false (V5 − V3)	-0.166	[-0.287, -0.035]	lower
fabrication vs no tally at all (V5 − V2)	+0.012	[-0.096, +0.120]	inconclusive
noise floor (V2-rep − V2)	-0.015	[-0.113, +0.083]	inconclusive

it is used. Our agents announce a position and then privately abandon it at the same rate as agents who never announced anything, which is the same missing normative channel seen from the other side.

Lying to the room did not steer it — and the reason is not the one the decomposition suggests. This is the arm we were most prepared to find something alarming in. A fabricated tally reporting the majority on the modal wrong answer moved the share of members ending on that answer by -0.001 $[-0.073, +0.064]$ against its own true-tally control (Table 2, Figure 2). Nothing, bounded tightly.

The same arm says more on the concordance metric, and we report it because that is the metric the informational channel was established on. The false tally leaves within-group agreement essentially where the private control leaves it ($+0.012$ $[-0.096, +0.120]$) and sits -0.166 $[-0.287, -0.035]$ below its own true-tally arm — nearly the mirror of the $+0.178$ the true tally buys. An anonymous count of what the room holds moves the room; the same sentence with the numbers fabricated does not.

That rules out the account the rest of this section invites. V3 and V5 are both anonymous, so the missing normative channel is held fixed across them and cannot be what stopped the lie. Something distinguishes a true count from a false one, and it is not indifference to peers.

What that something is, we can only partly say, because our fabrication has a flaw we found by auditing it rather than by designing around it. The invented sentence always reports two entries summing to the full membership, so a member holding a third answer sees a count with no slot for the ballot it has just cast: **93 of 358 member-rounds (26%), touching 41 of 50 groups, against 0 of 358 under the true tally**. To that member the claim is not merely false, it is impossible. Splitting members on whether their *first-round* ballot appeared in the first-round count — a pre-treatment split, since that ballot is cast before anything is announced — the concordance gap halves and stops resolving (-0.077 $[-0.228, +0.080]$, $n = 46$). We therefore cannot separate a room that weighed the claim and rejected it from a room that caught the arithmetic, and we say so rather than choose the flattering reading.

The steering bound itself survives that concern. On the uncontradicted members alone, the fabrication still moves the target share by $+0.013$ $[-0.065, +0.086]$. The null is not an artefact of detectability. So the reassuring half stands: the displayed aggregate is a free dial — it costs nothing and needs no model access — and it did not choose the room's answer, bounded inside ± 0.07. What does not stand is a tidy account of *why*. The experiment that would settle it is a fabrication consistent with every ballot cast and every position stated in chat, and we have not run it.

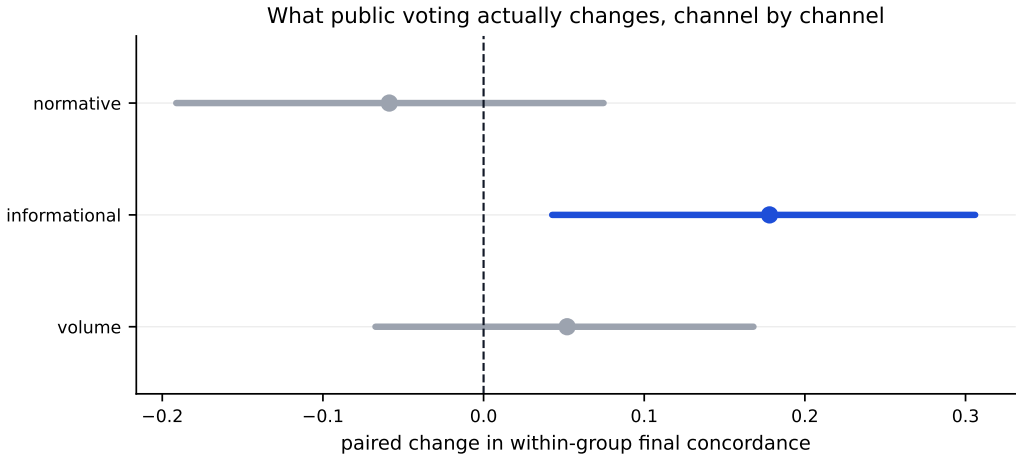

Figure 1: What public voting changes, channel by channel, on within-group final concordance. Coloured intervals resolve; grey ones do not. The effect is informational — learning *what* the group holds — and neither normative nor a context-volume artefact.

Table 2: The fabricated-aggregate arm and its control. The fabrication effect is V5 − V3: without V3 the comparison would confound *an aggregate was shown* with *the aggregate was a lie*.

	value	verdict
share of members ending on the fabricated target		
V2 private (control)	0.170	
V1n volume control	0.159	
V3 true aggregate	0.153	
V5 fabricated aggregate	0.161	
paired differences		
fabrication effect (V5 − V3)	-0.001 [-0.073, +0.064]	inconclusive
aggregate shown at all (V3 − V2)	-0.030 [-0.110, +0.053]	inconclusive

5 Study 6: whose register you read

5.1 What the free census already settled

Before spending anything, we counted tells in every generated transcript this research program had on disk — roughly 34,000 messages across three prior studies — against the two human corpora they imitate. Three results came out of that count, and all three changed the paid design.

The register calibration worked on almost everything. Hedging, assistant-register phatics, over-represented vocabulary and list structure all sit *at or below* the human rate for the corpus each arm imitates. That is worth stating plainly because it is the opposite of the usual claim: on six of seven detectors, generated deliberation text is not more AI-flavoured than the human text it is imitating. Real task-group chat hedges more than any generated arm we measured, and DeliData participants say “good point” to each other more often than our agents do.

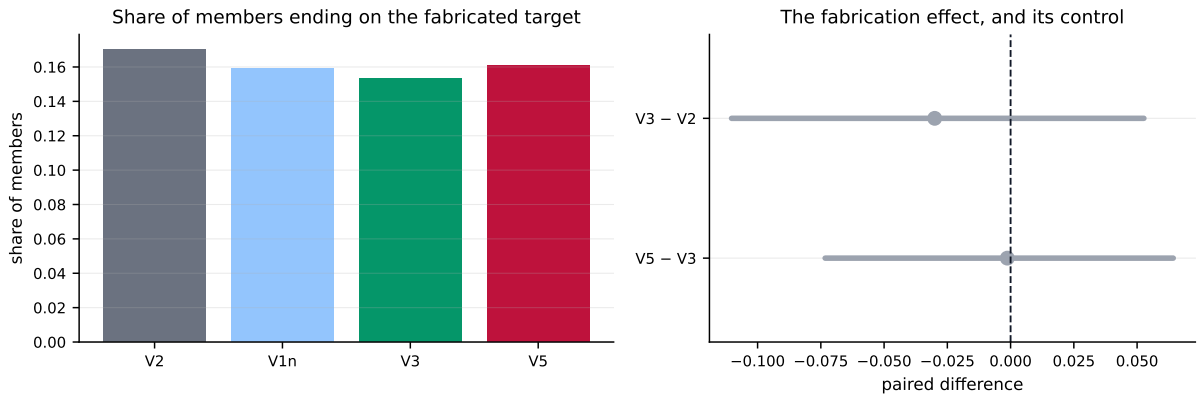

Figure 2: Left: share of members ending on the answer the fabricated tally pushed toward. Right: the fabrication effect against its own control, and the effect of showing any aggregate at all. Neither resolves.

What survived is one character, in one model. Em-dashes run 11–22 per 1,000 words in the prior study's arms against DeliData's 0.03. Fifty-one per cent of those generated messages contain one; **zero of 11,022 real DeliData messages do**. On that corpus a single punctuation mark is very nearly a perfect classifier.

But it belongs to the model, not to the pipeline. A second prior study ran the *same task* through the *same chassis* on a different model family across eight sizes from 0.8B to 397B, and every one of those sits at 0.4–1.1× the human rate. The arms with the habit are 5.4×. Identical work, identical scaffolding, and a 500-fold parameter range on one side: the tell is one model's typographic habit (Figure 3).

Size interacts with the task's register rather than driving tells alone. On the terse chat target the size ladder is flat (1.4–3.5, no trend). On the formal deletion-debate target the same models rise with size, 7.2 to 13.7 per 1,000 words from 2B to 122B. Bigger models are more detectable only where the task invites prose.

And a fourth result, about the other kind of tell. Within a group, the agent that hedges more in its *private* monologue is the agent that changes its answer, in five of seven arms that logged monologues. That association mostly runs the wrong way for a causal reading: a thought chain spans the whole conversation, so a switch at turn 8 can produce hedging at turn 9. Splitting each chain and using only its first half to predict the final answer, the effect survives in two arms of seven rather than five. A residual antecedent effect exists, it is smaller than the naive number, and it is invisible to every detector in Appendix A — which is precisely why the paid design has a separate arm aimed at it.

5.2 Results

The two suppression arms produce a double dissociation, and it is the paper's central result (Table 3, Figure 4).

The detector-visible intervention is behaviourally inert. Appending the anti-slop checklist to the speech prompt removed **69%** of measured tells — em-dashes fall from 19.3 to 3.1 per 1,000

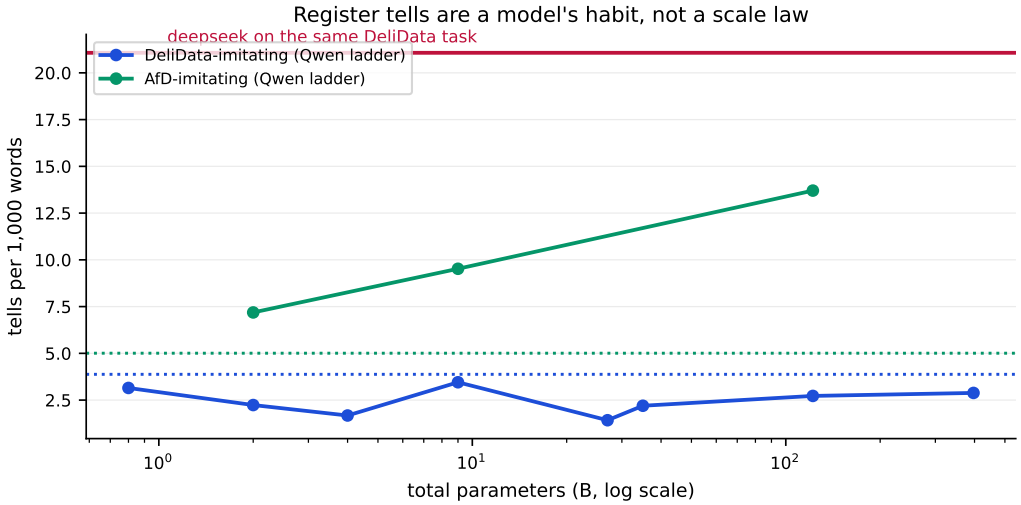

Figure 3: The free census. Dotted lines are the two human corpora; the solid red line is the model used in this paper's paid run, on the same task as the blue series. The tell that motivates the suppression arm is a property of that one model, not of scale and not of the pipeline.

words — and moved composite error by -0.036 $[-0.235, +0.159]$, which is inconclusive and well inside the noise floor (-0.121 $[-0.292, +0.055]$ for the replicate). Whatever the em-dash habit is doing to a reader, it is doing nothing to the room. **H27 is supported**, with the caveat that the interval is wide: this is an equivalence inside a band, not a demonstration of exact zero.

The detector-invisible intervention is behaviourally potent. Appending an anti-hedging instruction to the *think* prompt removed almost no measurable tells — 18.4 against the control's 21.6 per 1,000 words, and hedging in private thought was unchanged at 1.01 against 1.00 — yet moved composite error by -0.282 [−0.451,-0.124], resolved and outside the noise floor. **H28 is supported.**

The direction matters as much as the magnitude: lower is closer to the real groups. Telling agents to stop performing open-mindedness in their own reasoning made the simulation *more faithful*. The posture was not merely cosmetic residue; it was inflating the rate at which simulated groups discover correct answers nobody started with.

Composite error is a mean of four normalised deviations, and an arm can improve on it while moving *further* from real on a component, so the components are worth stating rather than leaving inside the aggregate. Against the control, S-stance improves sharply on rescue rate (normalised deviation $1.89 \rightarrow 0.67$; the raw rate falls from 0.634 to 0.366 against a real 0.220) and on score gain ($0.54 \rightarrow 0.05$), and gets *worse* on answer-change rate ($0.20 \rightarrow 0.45$; the raw rate falls from 0.508 to 0.350 against a real 0.636). So the honest description is that suppressing the posture fixes two long-standing over-shoots and overcorrects a third. It is a net gain, not a uniform one, and a reader who cares specifically about change rate should prefer the control.

And it does not buy that fidelity by collapsing the conversation. The obvious worry about an instruction to stop weighing both sides is that it trades exploration for convergence. It does not: S-stance produces the *most* diverse outcomes of any arm, with 0.785 distinct final answers per group against the control's 0.673, and slightly less spread in message length. Less

Table 3: Every arm. Real values on these groups: gain +0.121, change 0.636, lift +0.226, rescue 0.220. Composite error is slice-relative — comparable between arms here, not to the levels the prior study published. The last column is the paired difference against the control over 58 groups.

Arm	gain	change	lift	rescue	comp. err	tells/1k	Δ comp. err vs V2
V2 private (control)	+0.186	0.508	+0.344	0.634	0.789	21.6	—
V2-rep replicate (noise floor)	+0.159	0.492	+0.267	0.537	0.543	21.2	-0.121 [-0.292, +0.055]
V1 public, by name	+0.224	0.590	+0.405	0.658	0.932	20.0	+0.159 [-0.022, +0.347]
V1n volume control	+0.186	0.503	+0.313	0.488	0.590	19.8	-0.035 [-0.251, +0.187]
V3 true aggregate	+0.214	0.600	+0.400	0.537	0.763	20.0	+0.011 [-0.201, +0.227]
V5 fabricated aggregate	+0.199	0.503	+0.369	0.561	0.763	21.0	-0.041 [-0.257, +0.179]
S-lex speech suppression	+0.189	0.513	+0.333	0.512	0.642	6.6	-0.036 [-0.235, +0.159]
S-stance think suppression	+0.115	0.350	+0.170	0.366	0.352	18.4	-0.282 [-0.451, -0.124]

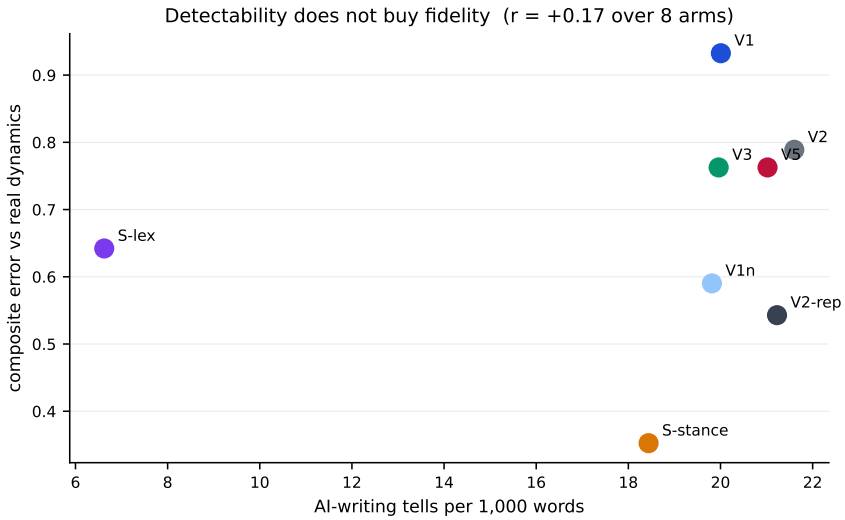

Figure 4: Detectability against fidelity. S-lex moved a long way left (69% of tells removed) without moving down. S-stance barely moved left and moved a long way down. The two interventions target different objects.

performed open-mindedness meant less convergence on a common answer, not more.

So detectability and fidelity are separable, and the free census says so twice. Across the eight arms, tell rate and composite error correlate +0.17, which we report as a picture rather than a test — n there is the number of arms. The claim rests on the two paired contrasts, and on an outside comparison the census supplies for free: the model family that writes at human tell rates at every size from 0.8B to 397B is the same family the prior work measured as *least* faithful at small sizes. Low detectability plainly does not imply fidelity.

The practical form of that: a vendor can make transcripts stop looking synthetic with one line in a prompt, at no measurable cost to how the room behaves — and should not mistake having done so for having made the simulation better. Conversely, the intervention that did make it better is invisible to every detector in Appendix A, so no amount of text analysis would have found it.

5.3 An attempted cross-model replication, and why it is absent

The obvious question about the normative gap is whether it belongs to this engine or to this model, and the answer would come from running the same three arms on a second family. We attempted exactly that — the control, the true-aggregate arm and the public arm on a 122B mixture-of-experts model from a different family, 30 groups — and it exhausted the study's budget after 12 usable conversations, all of them control. That model metered roughly fifteen times the per-conversation cost of the one used throughout, which our cost model, calibrated on the primary model, did not anticipate.

We report the attempt rather than omit it. What survives is a control arm with no comparison to make against it, which is worth nothing on its own, and the run is recorded as failed. Study 5's regime results are therefore **single-model**, and the limitation is stated in §7 rather than papered over with a partial arm.

The asymmetry with Study 6 is worth naming, because it cuts the other way: the register claim *is* cross-model, resting on a free census over a second family at eight sizes on identical work (§5.1). So this paper can say the em-dash tell is one model's habit, and cannot say the same about the normative gap. Anyone repeating the vote-visibility manipulation should meter a second family's cost on one group before committing to a design, which is the lesson we would have preferred to learn more cheaply.

6 Discussion

One channel is missing, and it explains three of the four regime results. Our agents update on *what* others conclude and are indifferent to *who* concluded it: showing an anonymous count moves the room, attaching names to it adds nothing, and a public vote does not bind the voter who cast it. Informational influence is present; normative influence is not detected, bounded within about 0.08 concordance rather than shown to be zero. For a platform simulating juries or town halls — settings where being seen to agree is the mechanism — that is one shortfall rather than three separate ones, and it is large: on this task with this model the simulation reaches half the convergence real groups do.

The fourth result does not follow from it, and that is the more interesting failure. An anonymous count is pure *what*-information, so a room that updates on *what* should have followed a fabricated count as readily as a true one. It did not: the fabrication sits resolvably below its true-tally control on the same metric. Whatever separates a true count from a false one in these agents, it is not the channel the other three results are about, and our own arm cannot cleanly identify it because a quarter of its member-rounds showed a count that was impossible rather than merely false. We report the contrast, the audit, and the pre-treatment split that halves it, and we leave the mechanism open.

So the safety reading is narrower than we expected to write. The fabricated-tally arm is the most manipulable surface a deliberation product has, and it did not choose the room's answer — bounded inside ± 0.07, and not by detectability, since the bound holds on the members our fabrication never contradicted. What we cannot claim is the tidy version: that the exposure is small *because* the simulation is unfaithful, and would grow as fidelity improves. That was our first reading and the V5-versus-V3 contrast refutes it. The honest statement is that a displayed

aggregate is worth auditing in any deployed system, that this particular lie did not work here, and that anyone closing the normative gap should re-run the arm rather than assume either direction.

Two kinds of tell, and only one of them is the one people look for. The register result is a clean double dissociation: removing 69% of the detectable tells changed nothing about the room, and the intervention that changed the room a great deal was invisible to every detector we built. The lexical residue is real, it is cheap to remove, and it is worth removing — transcripts that read as synthetic are a problem for a product whose value is that people believe the room. But removing it is a cosmetic act, and this paper's contribution is the measurement that says so, so that nobody mistakes a clean-looking transcript for a faithful one.

The stance tell deserves its own sentence. It was already known in this program to be behaviourally load-bearing — removing an "open-minded deliberation assistant" framing was the single largest fidelity improvement the engine ever got. What is new is that the residue after that fix is *still* costing fidelity, that suppressing it further helps again, and that no text detector would have told anyone. Register auditing and behavioural auditing are different jobs.

The tells are a model's habit, not the technology's. On identical work through an identical chassis, one model sits at 5.4× the human tell rate and another sits at or below human across a 500-fold parameter range. That reframes the whole "AI writing has tells" framing as a claim about particular models at particular moments, and it should make anyone building detection policy on such features nervous — ours is a single punctuation character that a one-line instruction removes.

What this says about how to evaluate simulated deliberation. Three practices earned their cost here and none is expensive. Re-run the control: without a replicate arm, two of our resolved-looking effects would have been indistinguishable from a re-roll and we would not have known. Decompose the manipulation: the headline public-versus-private contrast pooled three mechanisms, and only one of them was real. And compute the fidelity targets on the slice actually run: scoring a 30-group sample against corpus-wide constants gave a hypothetically perfect arm an error of 0.223, an offset that would have been silently inside every number in the paper.

7 Limitations

The task may be the reason we do not detect a normative channel. This is the limitation that most constrains Study 5's headline, and it deserves more than a sentence. The Wason task has a verifiable correct answer, so a peer's position is genuinely evidence and there is little social cost to being the lone dissenter who turns out right. Both features suppress normative influence in humans too. Prior LLM work reports conformity scaling with majority size (Weng et al., 2025) and persona abandonment under group pressure (Baltaji et al., 2024), so peer influence of *some* kind is clearly available to these models — which is consistent with the informational channel we do detect and leaves open whether the normative one is elicitable under stakes this task does not create. We therefore claim the gap for this engine on this task, not for language models in general, and the experiment that would settle it is the same manipulation on a task with tunable ambiguity or no correct answer at all.

Pilot scale. Thirty groups, paired across arms. The pairing buys a great deal of power relative to independent sampling, but small effects will land inconclusive and we report them as such rather than as absences. Where a comparison is inconclusive the paper says what effect size it could and could not have detected.

Regimes V4 and the room card are absent by choice. Multi-round ratification (announce, then re-vote) and the standing-instruction block are both real parts of the taxonomy this study works from, and both are second manipulated variables. Folding either in would confound the regime effect with a procedure effect, so they wait for their own study.

The public arm carries more information, not just more messages. V1n controls for how much was said. It does not control for the fact that a named vote is intrinsically more informative than an anonymous count, which is the mechanism rather than a confound — but it does mean V1's effect cannot be attributed to social identity specifically, only to identified disclosure.

The register arms modify prompts, not models. Suppression here is an instruction, and instructions are obeyed imperfectly. A measured failure to remove a tell is a failure of that instruction, not proof the tell is inherent.

Showing names may be a weak operationalisation of normative pressure. Our normative manipulation attaches an alias to a vote and nothing else. Real normative influence in humans runs on reputation, sanction, and the expectation of meeting these people again — none of which exists in a one-shot conversation between pseudonymous aliases who will never interact after the twenty-fifth message. So the honest reading of the normative null is narrower than "these agents do not respond to peer approval": it is that attaching an identifier to a vote, absent any stake in what that identifier accumulates, does not move them. A design with persistent identities across conversations, visible standing, or any consequence attached to being seen to flip would be a stronger test, and we have not run one.

Only the control is replicated. The noise floor comes from one re-run of one arm. That prices run-to-run variation for the control condition and assumes it transfers to the treatments; a design with every arm replicated would price it per arm and is what we would build with more budget. Where a treatment effect sits close to the floor we report it as inconclusive, which is the conservative direction, but we cannot rule out that a treatment arm is noisier than the control.

Model coverage is asymmetric between the two studies. The eight main arms use one base model, chosen because it is the configuration the prior work validated and the one carrying the register residue worth suppressing. Study 6's cross-model claim rests on the free census, which spans a second family across eight sizes on identical work. Study 5's regime contrasts are replicated on a second family at reduced scale (§5.3); that replication covers the three arms the decomposition needs and not the fabricated-tally arm, so the steering bound remains single-model.

One fabrication, one strength, two timings. The false tally always names the modal wrong answer with a bare majority, announced after the eighth and seventeenth messages. A near-unanimous fabrication, or one timed differently, might move a room that this one did not, and the reported bound covers only the manipulation as specified. Two ballot rounds also means the design cannot express ratification — announce, then re-vote, then announce again — which is where commitment dynamics would most plausibly appear, and which the taxonomy this study works from lists as its own regime.

8 Ethics

One arm of this paper lies to a simulated group about what the group thinks, and measures how far the lie moves it. That deserves a direct account rather than a disclaimer.

A correction we owe the reader up front. An earlier draft of this paper argued that the fabricated-consensus null was reassuring *because* the simulation lacks a normative channel, and therefore that improving fidelity should be expected to open the exposure. Our own audit refuted

that: the true and fabricated aggregates are both anonymous, so the normative channel is held fixed between them, and they behave differently. The safety claim in this section is consequently narrower than the one we set out to make.

Why run it at all. The fabricated-aggregate condition is not a capability we are introducing. Any system that can display a tally to a simulated room can display a false one, and the displayed aggregate is a free dial — it costs nothing and requires no model access. A vendor shipping deliberation simulation has already built the mechanism whether or not they have measured it. What is missing is the number: how much does it actually move a room, and does the effect survive the controls that separate persuasion from noise. Measuring it produces a bound that a deployment can be held to, and the alternative — leaving it unmeasured because measuring it feels uncomfortable — leaves the capability in place and the bound unknown.

No humans were deceived. Every participant here is a language model instance in a research harness. The deception is of simulated agents about simulated peers, on a logic puzzle with a known correct answer, and nothing generated in this study is presented to a person as human output.

What the result licenses and what it does not. A measured steering magnitude is a bound on manipulability, not a technique. It says how much a false consensus signal is worth in this setting; it does not improve anyone's ability to construct one, because constructing one requires no research. If the effect is large, the honest reading is that displayed aggregates in deliberation products need provenance — a room should be able to show where its tally came from — and we say so. If it is small, the honest reading is that this particular lever is weaker than intuition suggests, which is also worth publishing rather than leaving to intuition.

The register study has a nearer-term risk. Study 6 measures what makes generated deliberation text identifiable, and one finding is a single character that separates generated from real messages almost perfectly on one corpus. That is usable in two directions: it helps anyone auditing whether a transcript is synthetic, and it helps anyone wanting to evade such an audit. We report it because the detection side is the one that currently has no tooling at all, because the evasion is a one-line instruction that anyone attempting it would find immediately, and because a tell this fragile is not a foundation for detection policy — which is itself the point worth making to anyone building on it.

Corpus and licensing. DeliData is CC-BY-4.0. The Wikipedia deletion corpus consists of public edit-history contributions under Wikipedia's licensing; we use it in aggregate for register baselines and quote no individual editor.

9 Availability

The system under study is commercial software; code and model artifacts are not public and are available to partners and clients under agreement. The evaluation protocol is public: both studies run against public corpora (DeliData, CC-BY-4.0; the Wikipedia Articles-for-Deletion corpus) through public LLM APIs, and the appendices give the regime specification, every prompt, and the full detector suite in enough detail to reimplement.

References

Solomon E. Asch. Studies of independence and conformity: I. a minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, 70(9):1–70, 1956.

Razan Baltaji, Babak Hemmatian, and Lav R. Varshney. Persona inconstancy in multi-agent LLM collaboration: Conformity, confabulation, and impersonation. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, 2024. arXiv:2405.03862.

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. Synthetic replacements for human survey data? The perils of large language models. *Political Analysis*, 32 (4):401–416, 2024. doi: 10.1017/pan.2024.5.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T. Rogers. Simulating opinion dynamics with networks of LLM-based agents. In *Findings of NAACL*, 2024. arXiv:2311.09618.

Morton Deutsch and Harold B. Gerard. A study of normative and informational social influences upon individual judgment. *Journal of Abnormal and Social Psychology*, 51(3):629–636, 1955.

Jason Grey. Does simulated deliberation reproduce how real groups change their minds?, 2026. AnthroSim technical report IE-ASIM-2026-02.

Georgi Karadzhov, Tom Stafford, and Andreas Vlachos. DeliData: A dataset for deliberation in multi-party problem solving. *Proceedings of the ACM on Human-Computer Interaction*, 7 (CSCW2), 2023.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.06452.

Elijah Mayfield and Alan W Black. Analyzing wikipedia deletion debates with a group decision-making forecast model. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 206, 2019. doi: 10.1145/3359308.

Elisabeth Noelle-Neumann. The spiral of silence: A theory of public opinion. *Journal of Communication*, 24(2):43–51, 1974.

Vishakh Padmakumar and He He. Does writing with language models reduce content diversity? In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.05196.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.13548.

Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. Systematic biases in LLM simulations of debates. In *Proceedings of EMNLP*, 2024. arXiv:2402.04049.

Zhiyuan Weng, Guikun Chen, and Wenguan Wang. Do as we do, not as you think: The conformity of large language models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2501.13381; Oral.

A The detector suite

Seven detectors, each targeting a construction that is cheap to state and cheap to check by eye. The list is this project's own writing checklist turned into measurements, which is why it is short and opinionated rather than exhaustive.

Two properties matter more than coverage. Every rate is computed by the same function on real human corpora and on generated text, so a detector firing equally on both is reported as measuring English rather than quietly dropped — and the table below shows that two of them nearly do. And every rate is per 1,000 words, because generated messages in this program have historically run several times longer than the human ones they imitate, so a per-message count would report a length effect as a register effect.

Detector	What it counts	DeliData	AfD
`em_dash`	em-dashes and spaced en-dashes	0.03	1.47
`antithesis`	'not just X but Y' frames	0.01	0.10
`hedge`	hedging and both-sides framing	3.17	1.00
`tell_word`	over-represented register vocabulary	0.04	0.50
`list`	bulleted or numbered structure	0.15	1.41
`rule_of_three`	'A, B, and C' triads	0.00	0.37
`assistant_frame`	assistant-register phatics	0.48	0.14

The two human columns are themselves a result. Hedging is *more* common in real task-group chat than in any generated arm we measure, and assistant-register phatics are more common in DeliData than in most of them — people do say "good point" to each other. Conversely Wikipedia editors use em-dashes fifty times more often than DeliData chatters do, which is why each generated arm is compared against the corpus it is imitating rather than against a single pooled notion of "human".

Where a construction cannot be detected without parsing — genuine antithesis, for instance — the detector approximates it with an explicit lexical frame, and the approximation is documented in the code rather than hidden. Such a detector under-counts, but it under-counts identically on both sides of every comparison.

B Prompts and the regime specification

The think step (all arms). *"Privately, in one or two sentences and in your own reasoning style: what do you currently think the right cards are, and why? If something in the chat has genuinely changed how you see it, say what changed; otherwise restate your own reasoning."* In S-stance this is followed by: *"Do not hedge, do not weigh both sides, and do not perform open-mindedness. State what you actually think is right and why. If you have not been given a reason to change your mind, do not soften your position."*

The speech step (all arms). The agent's private assessment is supplied back to it, followed by *"Write your next chat message, consistent with this thinking."* plus the chassis's standing instruction to write one short plain-text sentence. In S-lex this is followed by: *"Never use an em-dash or a spaced dash; use a comma, a period or a semicolon instead. Do not write 'A, B, and C' triads. Do not open by praising what someone said."*

The ballot (all arms, identical). *"The task organiser is collecting everyone's current answer. Which cards would you turn over right now? Answer for yourself, from your own reasoning, not from what the group seems to think."* Returned as JSON, at temperature 0. This is asked at the same two points in every arm and again at the close.

Announcements. All are attributed to an `organiser` speaker and enter every agent's context exactly as chat does.

Arm	Announcement text
V2	(none)
V1	`{name} has submitted: [{cards}].` — one per member per round
V1n	`{k} of {n} members have now submitted their answers.` — one per member per round
V3	`Current count across the group: {n} for [{answer}]; …` — true counts
V5	the same sentence, with a majority fabricated onto the matching-bias answer

Budget accounting. Announcements do not consume the 25-message chat budget, so every arm produces the same number of agent-authored turns. They do enter context, which is the point; V1n exists to price that separately from their content.

C Preregistration and amendments

Both studies' hypotheses, arms, metrics and decision rules were written down before any paid call. Before scoring, the preregistrations were put through a structured adversarial critique — four independent lenses, every objection then verified against the repository with instructions to refute by default. Thirty-two objections were raised and 25 survived. Five changed the design or the analysis materially, and all five are reported because two of them would have produced publishable-looking numbers.

1. **The fidelity targets were the wrong slice.** Composite error scored generated arms against the prior study's corpus-wide targets while running on a 58-group sample. Verified independently: on this run's own groups a hypothetically *perfect* arm scores composite error 0.223, not 0. Targets are now recomputed on the run's own groups, which is why absolute composite errors here are slice-relative and are compared only between arms.

2. **There was no noise floor.** With no replicate of the control, any regime effect could have been run-to-run variation in a stochastic pipeline and the paper would have had no way to say so. Arm V2-rep is byte-identical to the control except for a salted seed; it is reported before every treatment contrast. Two effects that read as resolved sit inside it and are reported as inconclusive.

3. **Two-member groups break the aggregate regimes, not merely their precision.** At $n = 2$ an anonymous tally reading "1 for X; 1 for Y" tells the reader exactly how the other member voted, so the true-aggregate arm collapses into the public arm; and a fabricated majority contradicts the ballot the reader has just cast, so the fabrication arm measures

whether an agent notices an impossible claim rather than whether it can be steered. Eight of the original thirty groups were dyads. They are excluded from every regime comparison, retained for the register arms, and a 28-group top-up on three-or-more-member groups restored the sample to 50.

4. **H23 pooled three mechanisms.** The preregistered public-versus-private contrast is replaced by the decomposition in Table 1. Only one of the three channels resolves, and it is not the one the single contrast would have been read as demonstrating.

5. **The suppression hypotheses were not testable as written.** Both were stated against composite error, which is one number over all of an arm's agents and cannot be bootstrapped. Computed per group, they become paired tests over the groups every arm shares.

Two further defects were found in the harness rather than the design. The per-conversation cost constant, inherited from the prior study, was four times the measured rate and blocked an affordable run through the balance preflight; it is now measured, spend is metered into each run summary, and a declared budget ceiling is enforced in code rather than in prose. And the analysis of hedging in private thought had silently omitted an entire prior study whose result files are named differently — the identical omission the register census had already been caught making, in a second file. Corrected, that analysis spans 21 arms across two model families instead of 7, and its conclusion is unchanged.

Recorded deviations. The run used 30 groups where the preregistration said 25, and eight arms where it said four, both because measured cost came in at a quarter of the estimate. Both increase power and neither was decided after seeing an outcome. A companion arm that would have separated self-commitment from peer influence was declined as out of scope for these hypotheses, and a proposed replacement concordance statistic was declined because it is bounded below by the reciprocal of group size and would have manufactured an apparent effect from group-size differences alone.