The Inner Loop and the Ladder: What Makes Simulated Deliberation Realistic, and How Big It Has to Be

Jason Grey
jason@jason-grey.com

August 2026

Abstract

Simulated groups converge too fast. The usual fix is to tell the agents to hold their ground, which does not work. We study what does, on 500 real groups solving the Wason card-selection task (Karadzhov et al., 2023), where every member submits an answer before the conversation and may revise it during, so the target is a measured per-member change rate rather than a plausibility judgement.

The inner loop. An orchestrated engine given seeded agents performs *worse* than not conversing at all (composite error 2.064 against a no-conversation control at 1.704): 92% of wrongly-seeded agents state their assigned card and then argue themselves out of it, converting at 0.73 against a real 0.255. Instructing them to be stubborn does not help (0.524). What helps is replacing the instruction with a mechanism — a private think step that reasons *from* the member's misconception and yields only to a genuine defeat — which cuts the error by 45%. Measured within one chassis so orchestration is held fixed, that change is -0.921 $[-1.409, -0.428]$ on a group bootstrap, and it is the only one of our five pre-registered intervention contrasts that resolves. It survives thirteen alternative definitions of the error metric, a Holm correction over the full 28-pair grid, and three generation seeds per arm on both families, and it is a suppression rather than an addition: scored against agents who never converse, every sticky-loop arm *removes* change, while the engine and plain round-robin add it. **It is also model-specific, which we found by running the controls reviewers asked for.** On the same 50 groups with three seeds per arm, the loop beats instruction-level stubbornness on deepseek (-0.346 $[-0.606, -0.074]$) and loses to it on Qwen 9B ($+0.342$ $[+0.059, +0.438]$); at three further Qwen sizes and on Llama 3.1 8B nothing resolves, and a DynaDebate-style diversified initialisation and a verification-first gate beat the loop nowhere. So the headline is a result about one model, not about deliberation engines. The account we first offered for the difference, that the loop helps where baseline capitulation dominates, is refuted by the same ladder that tested it.

We report that last clause as a result. None of five further interventions — persona prose, quantitative style coupling, population-calibrated conviction, real conversation openings, and an engine port — separates from the baseline it was added to. Our notebooks had ranked them and drawn conclusions from the ranking. Adding intervals, which cost nothing because the per-group results were already committed, withdrew those conclusions. We also report the audit that forced this: five agents re-checking 510 claims against our own committed data found seven that a paper would have stated falsely, including one that invented a research requirement the data had already met.

The ladder. Sweeping nine models from 0.8B to 397B with the instrument fixed, solo drift toward the textbook answer rises steeply and resolvably with scale (endpoints $+1.708$ $[+1.293, +1.953]$) — small models are more human here because they do not know the answer.

But the pre-registered *monotone* form of that claim is strictly satisfied in 0.0% of 2,000 bootstrap draws, while no individual inversion in the series resolves either. Under generated deliberation the ladder is nearly flat: only the smallest model separates from the best one, and the remaining eight tie. Mechanism dominates scale by a wide margin.

Two further results sharpen the product rule. Splitting an agent into a speaking model and a thinking model shows both member fidelity *and* change dynamics track the mind, not the mouth, so speech can always be cheap. And task grounding does not substitute for scale — it *activates* latent competence, raising a 4B model's solo textbook drift by 2.3×, which means "grounding" and "teaching" are only separable when the evidence is orthogonal to the task's logical structure.

1 Introduction

Multi-agent deliberation has a signature failure: the agents agree too quickly and too completely (Chuang et al., 2024; Taubenfeld et al., 2024). Anyone who has run one has seen it. The question this paper asks is what actually fixes it, and how much of the fix is the model rather than the machinery around it.

Answering that needs a corpus where the target is measured rather than judged. DeliData (Karadzhov et al., 2023) is unusual in giving one: 500 groups solve the Wason card-selection task, each member submits a solution *before* the group chats, may resubmit during it, and the corpus records both. So "did this simulation reproduce how the group changed its mind" becomes four numbers rather than a rating — how much scores rose, how many members changed their answer, how many ended exactly correct, and how often a correct answer emerged in a group where nobody had one.

That last statistic is the one worth caring about. In 361 of 498 groups no active member starts correct, and in 26.9% of those somebody ends correct. No static readout of a population produces that number; it is a property of the interaction, and it is the thing a deliberation engine exists to reproduce.

We report two studies on this instrument. The first varies the machinery at fixed model. The second holds the machinery fixed and varies the model across nine sizes spanning a 500-fold parameter range.

Our contributions:

1. **The mechanism that fixes over-convergence, isolated within one chassis.** A private think step that reasons from the member's misconception and yields only to genuine defeat, worth -0.921 $[-1.409, -0.428]$ against the same engine with its original loop. Instruction-level stubbornness does not reproduce it (§5).

2. **A negative result about our own five other interventions.** None separates from the baseline it was added to at 50 groups, so the arm ranking our notebooks reported — and reasoned from — is not supportable. We also report the full 28-pair grid, in which 11 pairs separate unadjusted, including one within the group our first draft called indistinguishable (§7).

3. **The ignorance dividend, measured and then qualified.** Solo drift toward the textbook answer rises resolvably with total parameters, so a small model is more human on this task by not knowing it; but the pre-registered monotone form holds in no bootstrap draw, while no individual inversion resolves either (§9).

Number	Section	Slice	Artifact
Engine 2.064 vs control 1.704; conversion 0.73 vs 0.255	§5	pilot-50	`stage2`, `stage1` summaries
Instruction arm 0.524; think arm conversion 0.191	§5	probe-15	`stage4-instruct`, `stage4-think`
Inner-loop fix -0.921 $[-1.409, -0.428]$; replicated on the July engine	§7, §9.1	pilot-50	`intervals.json`, `seed_variance_ds.json`
Sensitivity and multiplicity	§7.1	pilot-50	`sensitivity.json`
Voice metrics with intervals	§6	pilot-50 rooms	`voice_intervals.json`
Net of the solo control	§8	pilot-50	`net_instability.json`
Ladder: drift $+1.708$ $[+1.293, +1.953]$; generated nearly flat	§9	pilot-50, nine sizes	`model_size/intervals.json`
Transfer: four Qwen sizes, Llama, verify arm	§9.1	pilot-50	`transfer_ladder.json`, `seed_variance.json`
Seed replicates (9B: $+0.342$ $[+0.059, +0.438]$)	§9.1	pilot-50, three seeds	`seed_variance.json`
Grounding activation, four wordings	§9.3	pilot-50	`grounding_sensitivity.json`

Table 1: Where each headline number is established, on which slice, and which committed artifact it is read from.

4. **Scale barely matters within a family, and the mechanism does not transfer between families.** Under generated deliberation, eight of nine sizes are statistically tied and the one that separates is the smallest (§9); the inner-loop mechanism that resolves on one family is excluded at that magnitude at four sizes of another, and our proposed explanation for the difference did not survive the test (§9.1).

5. **A mouth/mind decomposition.** Both fidelity and dynamics live in the thinking model; the speaking model is nearly free. The configuration space collapses to one slider (§9.2).

6. **Grounding activates rather than substitutes.** A definitions-only task sheet raises a 4B model's solo drift 2.3×, which sets a precondition on any evidence-grounding experiment: the evidence must be orthogonal to the task's logical structure (§9.3).

7. **A voice measurement that inverts a result of our own.** Simulated participants are not indistinguishable from one another; they are *more* identifiable than real people, and the actual defect is vocabulary poverty and self-repetition. Orchestration closes most of that gap while being invisible to the fidelity metric this paper is built on (§6).

8. **An audit of our own claims against our own data**, and what it cost us (§C).

Where each number lives. Two slices recur: the 50-group stratified *pilot*, on which every interval is computed, and its first 15 groups, the *probe*, on which the instruction arm and the deepseek within-family comparison were run. Table 1 gives, for each headline number, the section, the slice and the committed artifact.

2 Related work

Over-convergence. Chuang et al. (2024) find networks of LLM agents converging on the accurate answer, driven by a bias toward producing correct information rather than by peer

pressure, and recover human-like fragmentation once confirmation bias is induced by prompt. Taubenfeld et al. (2024) report debate agents drifting toward the base model's own position regardless of assigned persona. Our Study 1 reproduces this failure and locates it: the agent states its assigned answer and then reasons about it with the model's logic rather than the member's. Our fix differs from the prompt-level one in Chuang et al. (2024) in being a persistent mechanism, and §5 shows the instruction-level version failing on the same instrument.

What debate does and does not buy. A 2025–26 line reaches our reading from the other side. Wynn et al. (2025) find debate degrading accuracy over rounds, with models favouring agreement over challenging flawed reasoning, and show it persists when capable models are in the majority — which forecloses "use better agents" as a fix. Bertalanič and Fortuna (2026) isolate the aggregation step: in homogeneous debate the correct answer is often generated and then discarded by the vote, an oracle gap of up to 32 percentage points. Choi et al. (2025) supply the theoretical foil we owe the reader — under homogeneous input, debate can be dominated by simple majority voting — so any claim that a deliberation loop contributes beyond ensembling must be argued rather than assumed. We do not claim it. Our contrast is between two inner loops on one chassis, which holds aggregation fixed; whether the loop beats an ensemble of the same agents is a different experiment we did not run.

Interventions we did not compare against, and one we did. Two families target over-convergence directly and would be the natural ablation: dynamic path generation to break homogeneity (Li et al., 2026), and knowledge-enhanced protocols that give agents grounds to resist (Wang et al., 2023). Our inner loop is a third approach — constrain the private update rather than diversify the population or the evidence. After the third review we ran DynaDebate's initialisation component on our instrument (§9.1): a diversified rationale per member, which adds nothing to the loop on two families and costs on a third, consistent with the observation that each agent here is already seeded with a different real member's misconception, so the homogeneity that dynamic path generation breaks is not the failure on this instrument. The knowledge-enhanced protocols remain uncompared. A second family formalises the update itself: PABU (Jiang et al., 2026) makes an agent's belief state an explicit, progress-gated object, and SAVeR (Yuan et al., 2026) audits candidate beliefs against constraints before an action commits to them. Our sticky loop is the soft, prompt-level cousin of both, which the reviewer rightly flagged as its weakness: "genuine defeat" is asserted in an instruction rather than checked. §9.1 adds a verification-first variant in that spirit — a defeat audit that must name the specific claim before belief may change — run where the loop failed to transfer, to ask whether a more formal gate is more portable.

Personas and memory. Park et al. (2023) introduced the memory stream with recency decay and reflection, aimed at believability. Our H4 asks the calibration question instead and finds forgetting is load-bearing for fidelity rather than merely for cost. Park et al. (2024) show that grounding in a person's own words beats richer written descriptions of them, which is the frame for our finding that persona prose is inert while mechanical constraints are not; the same asymmetry appears across this research program (Argyle et al., 2023; Park et al., 2024; Bisbee et al., 2024).

Content effects and the task. Lampinen et al. (2024) show language models exhibit human-like content effects on reasoning, including on Wason, and also substantially outperform humans on the abstract form. That asymmetry is precisely our confound: the model knows the answer and the members do not, which is why every claim here is a difference against a no-conversation

Statistic (real groups)	value
mean 4-card score gain	+0.0962
members changing their answer	0.6410
exactly-correct lift	+0.1987
rescue rate (nobody starts correct)	0.1818

Table 2: The real change statistics on the 50-group stratified pilot. Composite error is the normalised mean absolute deviation from these four.

control rather than a raw level. Hao et al. (2026) caution that raw flip rates overstate conformity because a substantial share is spontaneous instability; §8 runs that decomposition and finds the spontaneous share exceeds the human total, which inverts the correction's usual reading.

Scale. Where the scaling literature asks when capability appears, we ask when it becomes a liability. On a task with a known correct answer, more capable models are *less* like the humans being simulated, which inverts the usual objective and is what §9 measures.

3 Instrument

Corpus. DeliData (Karadzhov et al., 2023), 500 groups, of which 498 have at least two active chatters, averaging 3.15 active members and 28.1 messages. Each member's initial solution is their pre-chat submission and their final is their last in-stream submission. A solution scores the fraction of four cards decided correctly. Our parse reproduces the corpus's own team-performance column exactly in 98.2% of groups.

One semantics pitfall is worth recording because it inverted an early result: the pre-chat tracker field is the *initial* state, not the final. Reading it the other way made deliberation appear to make people worse ($\Delta - 0.099$). The diagnostic that settled it was that tracker solutions score 0.580 while in-stream submissions score 0.682 rising to 0.719 first-to-last.

Targets. Table 2 gives the four statistics every arm is scored against, computed on the 50-group pilot used throughout. Composite error is the mean absolute deviation from these four, each normalised by its own scale.

The pilot, and how it was drawn. All generated arms run on a 50-group pilot drawn once, deterministically (seed 11), stratified by active-member count: groups are binned by $\min(n_{\text{active}}, 5)$ and sampled within bins in proportion to bin size, from the 498 groups with at least two active chatters. The pilot is therefore representative in group size, which is the variable the change statistics are most sensitive to, and is not stratified on outcome. We did not draw additional 50-group subsets — doing so would require re-running every arm — so the group bootstrap in §7 is the within-pilot version of that check and the between-subset version remains unrun.

Arms. Three, plus an ablation. **C** (control) gives an agent the task and its member's initial solution and asks it to finalise alone, measuring solo drift. **R** (replay) adds the real transcript, measuring how well the model reads real influence content — an upper anchor that involves no generation. **S** generates the conversation. The ablation replaces the engine with plain round-robin chat. Elicitation is identical across arms.

Models. Study 1 runs every generated arm on one mid-size model (`deepseek-chat-v3.1`). Study 2 holds the machinery fixed and sweeps a different family (Qwen3.5). The two studies therefore differ in model family as well as in what they vary, and any comparison of effect sizes across them is between studies rather than within one. We flag this wherever it matters.

The confound, and why C exists. Language models know the Wason task and humans mostly do not. Every claim here is therefore a difference over C rather than a level. Solo drift in C is large: on the full evaluation population the two Study-1 models rescue at 0.754 and 0.932 against a real 0.269, that is 2.8× and 3.5×. That gap is not noise to be removed but the central quantity of §9.

4 Specification

Round-one review asked for the inner loop and the error metric to be written down rather than described. Both are short.

Composite error. Each arm is scored against four real statistics $R =$ (score gain, change rate, correct lift, rescue r pooled over active members (rescue over groups where nobody starts correct). For arm statistics A,

$$E(A) \;=\; \frac{1}{4}\sum_{k=1}^{4}\frac{|A_k - R_k|}{\max(|R_k|\,,\,0.05)}.$$

Normalisation is by the real statistic's own magnitude, floored at 0.05 so that a near-zero target cannot dominate. Metrics are unweighted. Score gain and correct lift are signed; change and rescue are rates. Per-metric deviations for every arm are in the committed run summaries. The floor never binds on this pilot — every real statistic exceeds 0.05 (Table 2) — so the baseline and the no-floor definition coincide, and §7.1 reports how the verdicts move under twelve other definitions.

The sticky inner loop. Each agent carries three pieces of state: a fixed *misconception* m (the reasoning that produces its member's real initial card selection), a *current thought* τ that is rewritten each turn and never decays, and a memory M of at most three self-authored points with per-turn decay. On agent i's turn:

1. **Perceive.** Build a context of the last a messages ($a = 12$ frozen) plus the surviving points in M.
2. **Think** (private, never shown to others). Prompt: state in one or two sentences what you currently think the right cards are and why — *if something in the chat genuinely changed how you see it, say what changed; otherwise restate your own reasoning* — then list up to three points worth keeping. Output overwrites τ and updates M; each existing point survives with probability $1 - f$ ($f = 0.08$ frozen).
3. **Speak.** Generate one short message conditioned on the persona, m, τ and M — not on the raw transcript. Speech cannot see what the agent did not privately encode.
4. **Submit.** At elicitation, answer from τ, not from the transcript.

The defeat condition, stated exactly. There is no classifier and no threshold. The condition is a constraint in the persona prompt: the agent is told its own reasoning "is how you actually think about the puzzle", to "push back when someone's claim does not fit how you see it", and to "shift only if an argument genuinely defeats your way of seeing it". What makes it a mechanism

rather than an instruction is step 2: the concession has to be written into the private thought before it can reach speech or the submitted answer, and the thought is asked to name *what changed*. An agent that merely agrees socially leaves τ unchanged and therefore submits its original answer. §5 shows the instruction alone, without the private-state step, does not reproduce the effect (0.524 against 0.191).

We note this is a prompt-level mechanism over an external state, not a formal update rule, and that we did not run sensitivity analyses over its wording. A companion study in this program replaces it with an explicit log-odds belief state and finds the two compose.

Definitions used throughout. *Member match* is the share of active members whose elicited final card set equals their real final card set. *Rescue rate* is the share of groups in which nobody starts correct and somebody ends correct. *Wrong-to-correct* is the share of members seeded with an incorrect answer who end correct. *Change rate* is the share of members whose final differs from their initial.

Slices, and which number belongs to which. Three slices recur and mixing them has caused errors in our own notebooks, so we state them once. **pilot-50** is the 50-group stratified pilot and carries every headline number. **probe-15** is its first 15 groups, used for the cheap diagnostic arms (`instruct`, the tuning grid); its real wrong-to-correct rate is 0.191 against pilot-50's 0.255, so a conversion figure is only interpretable against its own slice. **heldout-35** is the complement, used to validate the memory dials. Every number in this paper is labelled with its slice at first use.

5 Study 1: the inner loop

The engine makes things worse. Running seeded agents through the orchestrated engine produces composite error 2.064 against the no-conversation control's 1.704. Conversation actively degraded fidelity.

The diagnosis. 92% of wrongly-seeded agents do state their assigned card. Then they argue about it using the model's own reasoning, and convert to the correct answer at 0.73 against a real 0.255. The persona is a costume worn over intact competence: the agent says the member's answer and thinks the model's thoughts.

Instructions do not fix it. Telling agents to hold their position leaves wrong-to-correct conversion at 0.524, against a real 0.191 on the same 15-group probe slice — still nearly triple. (Slices matter here and we state them: the 0.73 above is the engine arm on the 50-group pilot, where the real rate is 0.255. On the probe slice the engine arm converts at 0.714.)

The mechanism does. Replacing the instruction with a private think step that reasons *from* the member's misconception and concedes only on genuine defeat brings conversion to 0.191 against a real 0.191, on the same 15-group probe slice as the instruction arm. On the 50-group pilot the same mechanism gives composite error 1.100.

Measured within one chassis. The 2.064-to-1.100 comparison crosses an orchestration change as well as a loop change, so it overstates the loop's contribution. Porting the same inner loop into

Arm	comp. error	95% CI	member match	95% CI	change
engine, original inner loop	2.064	[1.676, 2.424]	0.301	[0.201, 0.406]	0.821
round-robin, no inner loop	1.834	[1.358, 2.313]	0.211	[0.115, 0.315]	0.904
round-robin + sticky inner loop	1.100	[0.707, 1.510]	0.308	[0.217, 0.399]	0.590
+ fusion persona prose (H5)	1.394	[0.988, 1.799]	0.365	[0.256, 0.476]	0.705
+ quantitative style coupling (H6)	1.340	[0.846, 1.820]	0.321	[0.218, 0.424]	0.667
+ population-calibrated conviction (H7a)	1.113	[0.669, 1.543]	0.353	[0.247, 0.451]	0.615
+ real 5-message opening (H7b)	0.866	[0.438, 1.347]	0.385	[0.284, 0.485]	0.635
engine, ported inner loop	1.144	[0.748, 1.557]	0.359	[0.255, 0.464]	0.596

Table 3: Every generated arm on the 50-group pilot with group-bootstrap intervals (2,000 resamples, shared matrix). Real change rate is 0.641.

the original engine isolates it: 2.064 to 1.144, or -0.921 $[-1.409, -0.428]$ paired over groups. That is the paper's central claim and the only contrast here that resolves.

What else we tried. Four further interventions, each locked before running. Fusion-profile persona prose (H5) did not resolve against the baseline on aggregate error ($+0.294$ $[-0.303, +0.888]$, which is inconclusive rather than a demonstrated null), and its pre-registered style prediction failed: messages ran 36 words against a real 9, with a length coefficient of variation of 0.15 against a real 0.35. We took that at the time as evidence the agents sounded more alike. §6 reports what happened when we measured that properly, which was not what we expected. Quantitative style coupling (H6) moved the surface statistics in the right direction but missed its locked targets (length 18.7 against a target of 15 or below, CV 0.22 against 0.25 or above), and leaked social concession into private belief, with wrong-to-correct rising to 0.547: the agents conceded verbally *and* changed their submitted answer, where real conflict-avoiders concede aloud and submit their own answer anyway. Population-calibrated conviction (H7a) landed two of the three per-bucket change rates exactly on the real gradient (partial 0.661, low 0.947) and left the third at zero against a real 0.211. Seeding the real first five messages (H7b) gave the lowest point estimate of any generated arm and still left member match at 0.385 against replay's 0.827.

6 Do the agents sound like different people? (summary)

H5 was pre-registered to raise style heterogeneity and appeared to lower it on a length-variance proxy. Measuring voice directly on committed transcripts, with room-level bootstrap intervals (Appendix A, Table 10), gives a different diagnosis. Simulated agents are not indistinguishable but *over*-distinguishable: every sticky-loop arm is easier to attribute than real humans (lift $+0.16$ to $+0.25$ against $+0.06$ $[+0.02, +0.10]$), because each agent uses about half the vocabulary of a real participant and repeats itself three to four times as often, all at the 95% level. Persona prose is the worst arm on both counts, which is H5's verdict with its mechanism corrected. The orchestrated engine is the exception: it matches real humans on vocabulary (-0.035 $[-0.077, +0.010]$) and cross-agent overlap (-0.035 $[-0.083, +0.011]$) while not separating from the round-robin chassis on composite error, so orchestration buys something the fidelity metric cannot see, and a paper reporting composite error alone would have concluded the engine does not matter.

Contrast (composite error)	Δ	95% CI	Verdict
the inner-loop fix, same chassis	−0.921	[−1.409, −0.428]	lower
H7b: real openings vs think	−0.234	[−0.796, +0.348]	inconclusive
H7a: calibrated conviction vs think	+0.014	[−0.516, +0.556]	inconclusive
engine port vs think chassis	+0.044	[−0.443, +0.513]	inconclusive
H6: style coupling vs think	+0.240	[−0.435, +0.892]	inconclusive
H5: population profiles vs think	+0.294	[−0.303, +0.888]	inconclusive

Table 4: The five pre-registered contrasts against the sticky-loop baseline, plus the within-chassis inner-loop contrast. Negative favours the first arm. Only the last resolves. This is *not* the full pairwise grid; see §7.

7 The intervals, and what they withdrew

Every verdict above was originally decided on a point estimate. This program's own rule is that verdicts are three-valued — supported, falsified, inconclusive — and decided on intervals. No committed run summary carried one.

That was recoverable at zero cost for one of the two variance sources, because per-group results were committed: resampling the 50 groups (2,000 draws, one shared matrix so contrasts are paired) gives Table 3 and Table 4. The bootstrap reproduces every committed point estimate to within 0.0002, which is the check that it is measuring the same quantity; the residual is rounding in the committed summaries.

The inner-loop fix resolves; none of the five interventions separates from the sticky-loop baseline.

The pre-registered set is not the whole grid, and saying otherwise cost us a claim. Our first draft generalised that result into "a band of six arms that are mutually indistinguishable". That is a statement about all fifteen within-band pairs, and we had computed five of them. Running the full grid — 28 pairs, same shared resample matrix, free — gives 11 separations, of which one lies inside the group we had called indistinguishable: real openings beat persona prose by -0.528 $[-1.021, -0.039]$, and that separation holds on 7 of the 7 seeds we tried, as does the headline inner-loop contrast. A second within-band pair, real openings against style coupling ($+0.474$ $[+0.006, +0.922]$), resolves on the committed seed but on only 4 of 7, so we report it as fragile rather than as a finding.

Corrected for multiplicity, most of the grid goes away. Round-two review asked for the correction rather than the disclosure. Table 11 gives every unadjusted separation with a bootstrap p-value (percentile inversion, floored at $1/2000$). Holm at $\alpha = 0.05$ across the 28 pairs keeps two of the eleven: the inner-loop pair itself and the engine against real openings. Benjamini–Hochberg at 0.05 keeps five. Treated as their own family, the six pre-registered contrasts leave exactly one Holm survivor, the inner-loop fix. Eight of the eleven unadjusted separations involve the two arms this paper already identifies as worst, which is why the correction costs the argument nothing it was leaning on.

7.1 Is it the metric? Thirteen other definitions

The composite averages four heterogeneous statistics with equal weight, and a reviewer asked whether the verdicts depend on that choice. Table 12 recomputes every contrast under thirteen alternatives on the same resample matrix: floors of 0.10, 0.20 and none; each deviation in units of

the real statistic's bootstrap standard deviation; maximum deviation; root-mean-square; rates only; signed statistics only; each single statistic left out; and the real target resampled alongside the arm instead of held fixed. The inner-loop fix resolves under all thirteen. No other pre-registered contrast resolves under any of them. The size of the pairwise grid moves — from 5 separations under maximum deviation to 13 under SD units — and the pairs that come and go are the marginal ones the previous paragraph already discounted. Decomposed by statistic, the fix is a result about outcomes rather than about raw churn: it resolves on score gain (-0.983 $[-1.533, -0.419]$), correct lift (-1.323 $[-1.982, -0.653]$) and rescue (-1.167 $[-2.129, -0.172]$), and does not on change rate (-0.210 $[-0.383, +0.019]$).

What is withdrawn. Our notebooks had ranked these arms and reasoned from the ranking — "the best generated arm", "the closest to real", "held, at the boundary" for a contrast with 0.006 of margin inside a pre-registered ± 0.300 band. Those readings are withdrawn. So is our own first-draft summary of *this* section, which committed the same error one level up: we computed a specific set of contrasts and described the outcome as though we had computed all of them. An adversarial re-read of the draft against these artifacts is what caught it.

8 Net of spontaneous change

Hao et al. (2026) show that a large share of apparent conformity in simulated groups is agents changing with nothing said to them, and that yield rates should be netted against that spontaneous floor. Round-two review asked us to run the isolated-agent counterfactual. We already had: arm C (§3) gives each member the task and their own initial selection and asks them to finalise alone, which is the counterfactual with the conversation removed. Table 13 nets every deepseek arm against it on the same 50 groups.

The correction runs backwards here, and that is the finding. The solo control changes its answer at 0.763, above the real groups' 0.641. Spontaneous change on this task, for this model, exceeds the total change humans show in conversation, because the spontaneous component is the model correcting toward the textbook answer (the confound that runs through this paper). Netted against it, every sticky-loop arm is *negative*: the loop with real openings -0.128 $[-0.211, -0.048]$ on change rate, the ported engine -0.167 $[-0.271, -0.071]$, the plain sticky loop -0.173 $[-0.260, -0.092]$. The two arms without a sticky loop are positive: the original engine adds $+0.058$ $[+0.003, +0.112]$ of change beyond what the model does alone and plain round-robin adds $+0.141$ $[+0.092, +0.194]$. On conversion the pattern is the same in sign and does not resolve for the loop arms.

So Hao et al.'s correction, which presupposes that the social component is what remains after the spontaneous one is removed, describes a different regime from this one. On Wason the conversation's job, when it works, is to *hold back* a model that would otherwise drift to the answer on its own; a mechanism that succeeds shows up as a negative net. The same picture holds on the Qwen ladder against each size's own control (committed with the table): net change is positive at 0.8B and 2B, where the solo control barely moves, and negative at 4B and 9B, where it moves more than the humans do.

Model	solo drift	95% CI	generated	95% CI	member match
0.8B	0.587	[0.516, 0.848]	0.917	[0.650, 1.164]	0.147
2B	0.773	[0.673, 0.992]	0.501	[0.341, 0.714]	0.173
4B	0.743	[0.539, 0.982]	0.559	[0.349, 0.783]	0.308
9B (local Q4)	1.042	[0.806, 1.290]	0.599	[0.436, 0.757]	0.288
9B	1.183	[0.914, 1.489]	0.648	[0.486, 0.832]	0.385
27B	1.967	[1.636, 2.258]	0.549	[0.458, 0.721]	0.346
35B-A3B	1.670	[1.369, 1.943]	0.610	[0.440, 0.853]	0.333
122B-A10B	1.325	[1.039, 1.591]	0.568	[0.479, 0.775]	0.346
397B-A17B	2.295	[2.043, 2.544]	0.418	[0.327, 0.625]	0.314

Table 5: The ladder. "Solo drift" is the no-conversation control; "generated" is the sticky-loop arm. Both are composite error against the same real statistics, with group-bootstrap intervals.

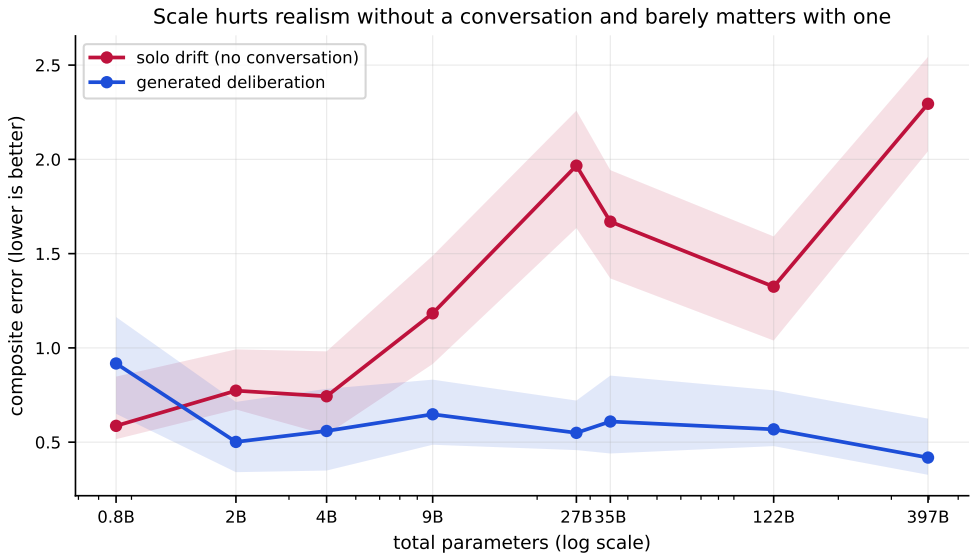

Figure 1: Solo drift rises steeply with scale; generated deliberation is nearly flat. Bands are 95% group-bootstrap intervals.

9 Study 2: the ladder

Holding the instrument fixed, we sweep nine models of one family (Qwen3.5, a different family from Study 1) from 0.8B to 397B, including three mixture-of-experts points and a 9B evaluated both locally quantised and hosted. Seeds are paired down the ladder. Table 5 and Figure 1 give both series.

The ignorance dividend is real, and resolves. Without a conversation, larger models are worse simulations of these humans: composite error runs from 0.587 at 0.8B to 2.295 at 397B, endpoints +1.708 [+1.293, +1.953]. (The 397B figure comes from the retry run; the original hosted run intermittently ignored the reasoning-disable flag and lost 10–15% of responses to parse failures, and our notebooks had already retired it.) The mechanism is not subtle. Large models know the Wason answer and drift to it; the humans did not and mostly did not. A small model is more human here by being ignorant, and it is ignorant for free.

But not monotonically, and we pre-registered monotone. H9 predicted a monotone increase. Across 2,000 bootstrap draws the series is strictly increasing in **0.0%** of them. But that is a statement about the whole nine-point series, and it does not license calling any particular dip real: tested individually, none of the three inversions resolves (4B below 2B, -0.030 $[-0.320, +0.180]$; 35B-A3B below 27B, -0.297 $[-0.671, +0.089]$; 122B-A10B below 35B-A3B, -0.345 $[-0.737, +0.071]$). So the honest verdict is: supported for trend, falsified for the monotone form we locked, and *undetermined* as to where the roughness is. Our first draft asserted the three dips were not noise, which is the point-estimate reasoning this very section exists to withdraw.

Mechanism dominates scale. Under generated deliberation the picture inverts and flattens. Against the best size, of eight contrasts only the 0.8B separates ($+0.499$ $[+0.156, +0.712]$); the remaining seven are inconclusive. There is a capability floor below which the model cannot participate, and above it the model is nearly irrelevant to how well the group's dynamics are reproduced. One caveat we owe our own preregistration: the locked comprehension floor was replay match below 0.5, and *both* 0.8B (0.494) and 2B (0.462) fall under it, so 2B is formally excluded from this comparison and we include it above. Reading the tie band from 4B up is the compliant statement, and it does not change the conclusion. This is the pre-registered alternative to our own H8, which had predicted an interior optimum, and it is what the data supports.

Activated parameters are not the axis. The 35B-A3B mixture activates only 3B — fewer than the 4B dense — yet drifts 2.1× as much (wrong-to-correct 0.547 against 0.263). Behaviour tracks total parameters; the textbook knowledge lives in the experts. Ignorance cannot be bought with sparse activation.

9.1 Does the mechanism transfer? Run after review, and it does not

Both the round-one reviewer and our own limitations section named the same weakness: Study 1 measures a mechanism on `deepseek-chat-v3.1` and Study 2 measures scale on Qwen3.5, so "mechanism beats scale" compares across families. The reviewer asked for the inner loop to be run inside the Qwen family. That turned out to be free — the local rungs of the ladder cost GPU time and nothing else, and Study 2 already had its sticky-loop arm — so what was missing was one control per size: the same chassis and the same 50 groups with instruction-level stubbornness instead of the private sticky-belief step. We ran it at four sizes.

It does not transfer, and this is not a power problem. At all four Qwen sizes the contrast is inconclusive — every interval crosses zero, so we cannot say the loop helps or hurts there. But every interval also *excludes* an effect the size of deepseek's (-1.305 on the shared slice, -0.921 on the engine chassis). The widest Qwen interval reaches -0.510. So the finding is not that we lacked resolution to see the deepseek effect in Qwen; it is that an effect that large is ruled out at every size we tested.

The explanation we offered was wrong, and the ladder is what showed it. When we had only the 9B point, we proposed a mechanism: the think step both suppresses social capitulation and adds private reasoning, so it should help where capitulation is the dominant failure and hurt where it is not. That predicts the loop's benefit tracks baseline capitulation, which the ladder measures directly. It does not. Ordered by capitulation rate the contrast runs -0.050, -0.253, $+0.334$, $+0.207$, -1.305 — non-monotone, with the two *least* capitulating models both mildly favouring the loop. A rank correlation over five models is -0.20, which at $n = 5$ is worth nothing

Model	ctrl w2c	instruct	sticky	Δ	95% CI	verdict
Qwen 0.8B	0.044	0.968	0.917	−0.050	[−0.409, +0.301]	inconclusive
Qwen 2B	0.066	0.754	0.501	−0.253	[−0.510, +0.029]	inconclusive
Qwen 4B	0.263	0.352	0.559	+0.207	[−0.198, +0.491]	inconclusive
Qwen 9B	0.212	0.265	0.599	+0.334	[−0.133, +0.538]	inconclusive
Llama 3.1 8B	0.073	0.680	0.449	−0.231	[−0.584, +0.090]	inconclusive
deepseek, pilot-50, three seeds	0.573	1.398	1.052	−0.346	[−0.606, −0.074]	lower
deepseek, probe-15, one seed	0.573	1.929	0.623	−1.305	committed	

Table 6: The same two arms, the same round-robin chassis, the same groups, four Qwen sizes, a third family (Llama 3.1 8B, added after the second review) and the deepseek comparison. Negative means the sticky loop beats instruction-level stubbornness. "ctrl w2c" is that model's baseline capitulation: the share of solo, wrong-seeded agents who end up correct with nobody talking to them.

anyway. We withdraw the account. It was a plausible story fitted to one point, and the cheapest available test refuted it within a day.

Where the difference sits, as far as the artifacts show. The reviewer asked for diagnostics. Both families commit every private thought and every public message, so Table 14 compares them on the same chassis and groups. At 9B the two models are nearly indistinguishable upstream of the final answer: message and thought lengths match (28 against 29 words; 56 against 49), private thoughts concede at similar rates (0.077 against 0.090), and among wrongly-seeded members whose initial selection omits the odd-number card — the card the textbook answer needs and no misconception produces — the odd card enters the private thoughts of 82% (deepseek 83%) and the public messages of 66% (71%). The divergence is at the end: it reaches deepseek's elicited finals in 42% of those members and Qwen 9B's in 13%, and the final selection agrees with the last private thought on that card for 66% of deepseek members against 46%. The loop changes what a Qwen agent thinks and says about as much as it changes deepseek; what differs is whether the elicited final answer, which reads the transcript in character, carries the change. That is a lead, not an explanation, and it points at the elicitation step rather than at the loop.

Seed variance, measured where it is free. Every generated arm in this paper was one generation per group, and both reviews named that as the largest gap. On the local rungs it costs GPU time only, so we re-generated the sticky-loop and instruction arms at 9B under two more seeds each, paired seed-for-seed (Table 7). The sticky loop scores 0.599, 0.568 and 0.511 across seeds and the instruction arm 0.265, 0.194 and 0.193: between-seed standard deviations of 0.044 and 0.041, against a group-bootstrap interval half-width near 0.3. Generation variance is real and an order of magnitude smaller than sampling variance over groups, which is the answer to whether single-seed intervals were hiding an effect. They were hiding one, in the other direction: the contrast is inconclusive on each seed alone (+0.334, +0.374, +0.318) and resolves when the three are pooled, +0.342 [+0.059, +0.438]. **At Qwen 9B the sticky loop is worse than telling the agent to be stubborn.** Non-transfer was the cautious reading of one seed; three seeds make it a sign reversal. The deepseek arms remain single-seed, because they are not free, and we say so in §11.

Model	sticky loop, per seed	instruct, per seed	Δ per seed	pooled Δ [95% CI]	verdic
Qwen 9B	0.599 / 0.568 / 0.511	0.265 / 0.194 / 0.193	+0.334 / +0.374 / +0.318	+0.342 [+0.059, +0.438]	higher
Qwen 4B	0.559 / 0.357 / 0.419	0.352 / 0.232 / 0.489	+0.207 / +0.125 / -0.069	+0.088 [−0.125, +0.245]	inconc
Qwen 2B	0.501 / 0.595 / 0.593	0.754 / 0.589 / 0.731	-0.253 / +0.006 / -0.138	−0.128 [−0.243, +0.007]	inconc

Table 7: The within-family contrast under three generation seeds per arm, same 50 groups. Per-seed intervals are group bootstraps; the pooled interval resamples groups with all seeds' paired deltas averaged draw by draw.

A third family, which narrows nothing and is worth having. The reviewer suggested a third family to separate architecture from training data from instruction tuning as the locus of non-transfer. Llama 3.1 8B runs locally, so we ran the full contrast on it: control, sticky loop, instruction arm, same chassis and groups (Table 6, last Qwen row and below). Its solo control is unlike either other family — every agent abandons its seeded selection (change rate 1.000) and only 7% land on the correct answer, so it neither holds the seed nor knows the textbook — and that shapes what the loop can do for it. The sticky loop scores 0.449 against the instruction arm's 0.680, a contrast of -0.231 $[-0.584, +0.090]$: the deepseek sign, not resolving, and excluding the deepseek magnitude. The diagnostics row is also its own: Llama concedes in 30% of private thoughts, three times the other models, and its elicited final tracks its last private thought 85% of the time, the highest of any model, so on this family the loop's private state does reach the answer and the answer is still not much closer to the humans'. With three families the hypothesis space is not narrower. Deepseek resolves in favour of the loop, Qwen 9B resolves against it, Llama leans with deepseek and does not resolve; architecture, data and tuning all differ across the three, and nothing here separates them. What the third family adds is the demonstration that the loop's effect is not even monotone in sign across families, which is a stronger warning than non-transfer.

Smaller rungs are noisier generators, and the contrast changes sign across seeds. The same three-seed design at 4B and 2B (Table 7) gives between-seed standard deviations of 0.10 and 0.13 at 4B and 0.05 and 0.09 at 2B, two to three times the 9B figure. At 4B the contrast runs $+0.207$, $+0.125$ and -0.069 across seeds and pools to $+0.088$ $[-0.125, +0.245]$; at 2B it runs -0.253, $+0.006$ and -0.138 and pools to -0.128 $[-0.243, +0.007]$. Neither resolves, and at both sizes a single seed can put the sign either way. So the ladder's earlier single-seed transfer rows (Table 6) were, for the two smallest rungs, one draw from a distribution whose sign is not settled at 50 groups, and the honest summary of the Qwen family is: worse than instruction at 9B, undetermined below it, and nowhere close to deepseek's magnitude.

The deepseek arms replicated, after the third review. The one measurement this paper could not make for free was generation variance on the paid model. With a budget for it, we re-generated the sticky-loop arm and the instruction arm on deepseek on the full 50-group pilot under three seeds each (the instruction arm had only been run on the 15-group probe), and ran the engine with the ported loop under two further seeds. Table 8 gives the results and Table 6 now carries the deepseek row on the same slice and seed count as the Qwen rows. The sticky loop scores 1.100, 1.147 and 0.910 (SD 0.125) and the instruction arm 1.693, 1.310 and 1.192 (SD 0.262); the contrast resolves on one seed (-0.594 $[-1.123, -0.006]$), not on the other two, and resolves pooled at -0.346 $[-0.606, -0.074]$. Two things follow. The probe-15 figure of -1.305 that the previous version reported was an overestimate by a factor of about four, half of it the probe slice and half of it the seed; we have replaced it. And the sign reversal across families is now

Arm / contrast	per seed	SD	pooled Δ [95% CI]
pilot-50			
sticky loop (deepseek)	1.100 / 1.147 / 0.910	0.125	—
instruction only (deepseek)	1.693 / 1.310 / 1.192	0.262	—
engine, ported loop, August code (deepseek)	1.908 / 1.464	0.314	—
diversified paths + sticky loop (deepseek)	1.378	—	—
engine, ported loop, July code (deepseek)	1.144 / 1.044 / 1.238	0.097	—
think-full - instruct-full	-0.594 / -0.163 / -0.281	—	−0.346 [−0.606, −0.074] lower
engine-ported - think-full	+0.760 / +0.553	—	+0.657 [+0.234, +1.065] higher
paths - think-full	+0.278	—	+0.278 [−0.273, +0.843] inconclusive
paths - instruct-full	-0.315	—	−0.315 [−0.912, +0.332] inconclusive
engine-ported - instruct-full	+0.598 / +0.272	—	+0.435 [−0.006, +0.854] inconclusive
engine-ported-july - think-full	+0.044 / -0.104 / +0.327	—	+0.089 [−0.219, +0.397] inconclusive
subset 2 (disjoint 50)			
sticky loop (deepseek)	0.506	—	—
instruction only (deepseek)	0.803	—	—
think-full - instruct-full	-0.297	—	−0.297 [−0.773, +0.212] inconclusive

Table 8: Deepseek arms under generation seeds on the stratified pilot, and (where run) the disjoint second subset. Composite error per seed; contrasts per seed and pooled over seeds, group bootstrap with all seeds' paired deltas averaged draw by draw.

seed-robust in both directions: on the same 50 groups, three seeds each, the loop beats instruction on deepseek (−0.346, resolves) and loses to it on Qwen 9B (+0.342, resolves). That is the paper's transfer result in its final form, and it is stronger than the version that rested on one seed per family.

The headline pair under seeds, and what the production engine did in the meantime. The headline contrast is the original engine (2.064, one generation per group) against the engine with the ported loop (1.144). We re-generated the ported arm under two further seeds on the engine as it stands today and got 1.908 and 1.464: against the original engine that is −0.157 (inconclusive) and −0.601 (resolves), and had we stopped there the headline would have been reported as seed-fragile. It is not. The engine is production code, and between the July run and this replication it absorbed changes from later work in this program (handouts as standing material, opening rationales, a decoder-noise change, and the production port itself). Re-generating new seeds from a checkout of the July commit — the commit whose engine code is byte-identical to the one the original run used — gives 1.044 and 1.238 against the original 1.144, a between-seed standard deviation of 0.10, and contrasts against the original engine of −1.020 and −0.826 that resolve as the original −0.921 did. So the headline is seed-robust on the engine it was measured on, and the August numbers measure something else: the production engine has drifted back toward the over-convergence the loop fixed (change rate 0.71 against 0.60 on the worse seed, wrong-to-correct conversion and rescue both up, conversations five messages longer), and on its own two seeds it is now resolvably *worse* than the bare round-robin chassis with the loop (+0.657 [+0.234, +1.065]), which in July it tied. We report that as a finding rather than a nuisance, because it is the paper's own warning made concrete: a validated inner loop does not stay validated through unrelated engineering, and an instrument like this one is what detects the regression. The original engine's loop no longer exists in the codebase, so that side of the pair remains one generation per group.

A second, disjoint set of 50 groups. The reviewer also noted that every interval here resamples one fixed stratified subset. We drew a second 50-group subset from the remaining 450 by the same stratified procedure with a different seed and ran the sticky-loop and instruction arms on deepseek there. The levels are not comparable across subsets, since each subset has its own real statistics (the loop scores 0.506 on subset 2 against 1.100 on the pilot), but the contrast is: -0.297 $[-0.773, +0.212]$, the pilot's sign and about the pilot's pooled magnitude, on one seed and not resolving. Between-subset variation of the contrast is therefore of the same order as between-seed variation, and neither reverses it.

A path-diversified initialisation does not beat the loop anywhere. The reviewer asked for a head-to-head with a DynaDebate-style intervention. DynaDebate's first component, dynamic path generation, initialises agents with diverse reasoning paths so that a debate does not start homogeneous; on our instrument each agent's starting answer is a real member's and cannot be changed, so we diversified what can be: a generator call writes, for every member of a group, a distinct everyday rationale that arrives at that member's real selection, with no two members sharing an argument or a style, and that rationale replaces the mechanical misconception template in the persona. Everything else is the sticky-loop chassis. On deepseek the arm scores 1.378 against the loop's 1.100 and the instruction arm's 1.693 ($+0.278$ $[-0.273, +0.843]$ and -0.315 $[-0.912, +0.332]$; neither resolves). On Qwen 9B it scores 0.561 against 0.599 (-0.037 $[-0.202, +0.172]$). On Llama 3.1 8B it scores 0.785 against 0.449 ($+0.335$ $[-0.002, +0.553]$), one draw short of resolving as worse. Diversity of rationale, then, adds nothing to the loop on two families and costs on the third, which is consistent with what §2 argued: the homogeneity that path generation breaks is not this instrument's failure, because the members' own misconceptions already differ. DynaDebate's second component, the process-centric audit of peer reasoning, is the verify arm below.

A verification-first gate is not more portable; on this task it is a teacher. The reviewer asked whether a more formal update rule — PABU's progress-gated belief or SAVeR's audit-before-commit — might travel where the soft loop did not. We built the prompt-level version: the same sticky loop, but before any belief may change the agent must list what the chat has claimed against its cards and say, claim by claim, whether the claim actually shows one of its cards cannot break the rule or a card it left out can, and may change only on a claim that does. No task logic is supplied. Run where the loop failed to transfer, on the same groups and seeds, it makes things worse. At Qwen 9B the audit arm scores 0.703 against the loop's 0.599 and the instruction arm's 0.265, and its wrong-to-correct conversion is 0.350 against the loop's 0.146 and the solo control's 0.212: agents made to audit the card-level claims work out the textbook answer in the audit. On Llama 3.1 8B it scores 0.641, between the loop's 0.449 and the instruction arm's 0.680. Every contrast is inconclusive ($+0.437$ $[-0.138, +0.837]$ against instruction at 9B; -0.039 $[-0.407, +0.191]$ on Llama), and the point estimates do not favour it anywhere. The reason is the one §9.3 gives for the task sheet: on a task whose structure is the knowledge being tested, asking an agent to verify claims about that structure is asking it to solve the task, so a formal defeat check is scaffolding rather than a gate. Whether such a rule is portable on a task with evidence orthogonal to its logic is exactly the question the companion paper's corpus was chosen to allow, and we do not answer it here.

What is left. A mechanism that resolves on one model, is excluded at that magnitude on four models of another family spanning a 500-fold parameter range, and has no working explanation

Configuration	comp. error	member match	change	rescue
2B mouth, 2B mind	0.501	0.173	0.667	0.273
2B mouth, 122B mind	0.434	0.333	0.308	0.212
122B mouth, 2B mind	0.466	0.237	0.673	0.212
122B mouth, 122B mind	0.568	0.346	0.173	0.151

Table 9: H11: which model speaks and which model thinks, on the same chassis with paired seeds.

for the difference. We would rather report that than the version of this paper that did not run the control. The practical consequence is a warning rather than a rule: an inner-loop intervention validated on one model should be revalidated before it is carried to another, and the revalidation is cheap when the target model runs locally.

9.2 Mouth and mind

If reading fidelity favours large models and realistic dynamics favour small ones, an agent might use both: a small model to speak and a large one to think and answer. Table 9 crosses the two.

Member match tracks the mind, as predicted. Change dynamics were predicted to track the mouth and *do not* — a 122B doing all the talking with a 2B thinking changes at 0.673, essentially the pure-2B rate. Belief updating happens in the think loop, so the mouth cannot carry the dynamics whatever it says. The mouth's contribution is real but marginal: a small mouth pulls a large mind's change rate toward real, and an articulate mouth makes a small mind's transcript easier to read.

The configuration space collapses accordingly. Speech can always be a cheap local model. The think model is the product slider: small mind for population-realistic dynamics, large mind for member-level fidelity. We note the mixed configuration is not the best cell on the instrument — the pure 397B generated arm scores 0.418 against its 0.434 — so the case for the split is cost, not peak fidelity. We note the point estimates here are single runs without intervals, and the two mixed cells differ by 0.031, which is well inside what §7 showed to be indistinguishable on this instrument — so we read the *direction* of the mouth/mind dissociation and not the ranking of the two mixed cells.

9.3 Grounding activates rather than substitutes

We had hoped supplied grounding could substitute for parametric knowledge, letting a small model behave like a right-sized mind. It does the opposite. The task sheet is short enough to quote in full, which is the cleanest answer to whether it leaks the solution:

> Reference sheet everyone in this study received: Each card has a letter on one side and a number on the other; you can only see one side of each card. A vowel is A, E, I, O, or U; every other letter is a consonant. Even numbers are 0, 2, 4, 6, 8; odd numbers are 1, 3, 5, 7, 9. The rule being checked: every card with a vowel on one side has an even number on the other side. At the end, each person submits the set of visible cards they would turn over to check the rule.

It contains no worked example, no cards, and no statement of which cards to turn over; it restates definitions and the rule under test. It is appended to the system prompt of every call in the grounded condition and nothing else changes. Giving it to a 4B model raised that model's *solo* textbook drift by 2.3× (wrong-to-correct 0.263 to 0.606, solo rescue 0.939). Restating the rule

woke dormant competence. The interaction is itself orderly: control drift moves −0.03, +0.09, +0.34 at 0.8B, 2B and 4B, so activation is proportional to latent knowledge.

Wording and placement, tested after review. A reviewer asked how sensitive the activation is to how the sheet is worded and where it sits. The stage-1 ladder runs locally in about a minute per size, so we ran four variants at 2B, 4B and 9B, plus a second bare seed (Table 15). At 4B, where the effect is largest, solo conversion is 0.263 and 0.285 on two bare seeds; the verbatim sheet in the system prompt gives 0.606, a paraphrase of it 0.679, the paraphrase in the user turn 0.635, and the verbatim sheet in the user turn 0.394. At 9B the bare seeds give 0.212 and 0.190, the paper's condition 0.650, and the three variants 0.382 to 0.467. So the activation is robust to both wording and placement: every grounded variant at 4B and 9B lies two to three times above the bare seed spread, and the spread *among* variants (up to 0.27) is smaller than the gap between any of them and bare. Which variant activates most is not stable across sizes — the paraphrase at 4B, the verbatim system-prompt sheet at 9B — so we read that ordering as noise. At 2B all five cells lie within 0.07 of one another, which is the "proportional to latent knowledge" reading again. What we did not test is the reviewer's other suggestion, a sheet the agent must *retrieve* rather than one that is always present; on this instrument that is the consult-on-demand design the follow-on paper uses, and we would expect it to reduce activation for the same reason evidence a participant chooses to read is not scaffolding.

The consequence is a precondition rather than a defeat. On a task whose structure *is* the knowledge being tested, grounding and teaching are indistinguishable, so evidence-grounding effects are only measurable when the evidence is orthogonal to the task's logical structure — material a participant *consults* rather than scaffolding that *instructs*. Establishing that separation is a corpus requirement, and it is the requirement the follow-on work in this program is built on.

10 Discussion

One metric cannot carry a simulation. §6 is the sharpest case in this paper. On composite error the orchestrated engine and the round-robin chassis are indistinguishable; on how the transcripts actually read they are not close, and the engine is the only arm that reaches human-level vocabulary diversity between speakers. A reader deciding whether a simulated focus group is usable will notice the second thing first. We built this paper on one number, and it took a second number to see that.

Mechanisms beat descriptions, as far as we can tell. The one intervention that resolves constrains what an agent may do at a decision point: a think step that must reason from a stated misconception, conceding only on a defeat condition. The interventions that describe who the agent is did not resolve, and persona prose made speech *more* uniform rather than less — though §7 is clear that failing to resolve at 50 groups is not the same as failing to work. The memory dials belong with the descriptions rather than the mechanisms here: our own appendix records that the frozen setting is the best *memory* configuration measured and not an improvement over having none.

Over-convergence is not a property of an engine. The same inner loop that had to be restrained here ossifies on a corpus where change is driven by evidence arriving from outside the room. What differs is whether the corpus supplies a reason to move that the simulation can

reproduce. A number tuned to fix over-convergence on one task is not a property of the engine and should not be carried to another.

Neither scale nor a portable mechanism. Across a 500-fold parameter range, generated deliberation quality is flat above the capability floor, so scale is not the lever. But the mechanism is not portable either: §9.1 finds nothing of its magnitude at four sizes of a second family, and the tidy explanation we proposed for that — fit the loop to the model's default failure mode — is exactly what the ladder refuted. We are left with a warning rather than a design rule: validate an inner-loop intervention on the model you intend to ship, because ours did not survive the move, and we cannot yet say what predicts whether it will. Reach for a large model when member-level reading fidelity is the requirement, which the mouth/mind split shows is a separate dial.

Ignorance as a design material. On tasks with a known answer, the model's competence is a liability, and small models supply the needed ignorance for free. The obvious counter — ground the small model instead — fails, because grounding activates the very knowledge you were trying not to have.

11 Limitations

Seed variance is measured on both families, with one arm left single-seed. Our intervals resample groups, so they answer how much a result depends on which 50 groups we drew. §9.1 adds three-seed replicates on the local Qwen rungs and on deepseek's sticky-loop and instruction arms, and two further seeds of the engine with the ported loop. Between-seed spread is below the group interval everywhere but not negligible on deepseek's instruction arm (SD 0.26). The original engine, the worse half of the headline pair, remains one generation per group because its loop no longer exists in the codebase; the headline contrast is therefore one seed of the original engine against replicated seeds of the ported one, on the July engine code. The current production engine does not reproduce the ported result (§9.1).

The partial exception is the *resample* seed. We re-ran every separating pairwise contrast under seven different bootstrap matrices, which is why §7 can report one within-band separation as robust (7 of 7) and one as fragile (4 of 7). That measures stability of the interval procedure, not of the generations underneath it, and we do not want the two confused.

Fifty groups. Seventeen of the 28 arm pairs do not separate, including all five pre-registered contrasts against the baseline. Some of those are surely real differences we are not powered to see, and nothing here should be read as evidence that persona grounding or style coupling does *not* help. Conversely, of the 11 that do separate, two survive Holm correction and one is seed-fragile.

One task, and a peculiar one. Wason has a verifiable correct answer and is famous enough that models know it. Both features drive our central confound. Whether the ignorance dividend exists on tasks without a canonical answer is untested and we would not assume it.

The mechanism is demonstrated on one model only. We ran the control at four sizes of a second family and on a third (§9.1) and found nothing of the magnitude we measured on deepseek; at Qwen 9B the sign reverses under three seeds. The -0.921 headline should therefore be read as a result about `deepseek-chat-v3.1`. We do not know what distinguishes it: baseline capitulation

was our candidate and the ladder rules it out; the diagnostics locate the divergence downstream of the loop, at the elicited final, without explaining it; and three families that differ in architecture, data and tuning at once cannot separate those.

One model family for the ladder. Small models in the family are distilled from large ones, which may compress behavioural differences across sizes and bias against finding a curve. A flat result is therefore weaker evidence than a curve would have been. A second family was pre-registered as the robustness check and has not been run.

Spontaneous instability is netted, with a caveat. §8 nets every arm against the solo control. That control is a single elicitation rather than Hao et al. (2026)'s repeated self-reflection, so it is a floor on spontaneous change rather than a ceiling; and the real rate is gross, since human non-social churn cannot be subtracted from the corpus side. Both conservatisms push the nets further negative than a like-for-like comparison would.

Anonymous members. DeliData members carry no attributes, so nothing here speaks to simulating identified individuals; that boundary is studied separately in this program.

12 Ethics

The corpus is public, released CC-BY-4.0, and its participants are anonymous by construction. We report only aggregate statistics.

The dual-use surface is narrow. The mechanism we identify makes simulated agents *harder* to move rather than easier, and the paper's practical recommendation is to spend on resistance rather than on model capability. We note one asymmetry worth stating plainly: a technique that makes simulated people hold positions realistically also makes a simulated focus group a more convincing artifact, and we would caution against treating any of these fidelity numbers as licence to substitute simulated deliberation for consulting actual people. The best member-level match we measured for generated conversation is 0.385, against 0.827 for replaying a real one — a point estimate on an instrument where arms do not separate, but the gap is large enough to survive that caveat and is the right thing to quote when someone proposes the substitution.

A Do the agents sound like different people? (full)

H5 was pre-registered to raise style heterogeneity and appeared to lower it, on the strength of a length coefficient of variation falling from 0.18 to 0.15 against a real 0.35. That is a thin proxy: two agents writing identical prose at different lengths score as heterogeneous. Since the complaint it stands in for is "they all sound the same", we measured that directly on transcripts we had already committed. Table 10 reports four quantities per arm — vocabulary breadth, how similar two agents in a room are to each other, how much a speaker repeats itself, and how reliably a classifier can identify which member wrote a held-out message from that member's other messages alone.

The agents are not indistinguishable. They are over-distinguishable. Every generated arm except plain round-robin is *easier* to attribute than real humans: lifts of +0.16 to +0.25 against a real +0.06. A classifier identifies the author of a simulated message more reliably than

Arm	TTR	overlap	self-rep.	attribution lift
real humans, same 50 rooms	0.678 [0.65, 0.70]	0.464 [0.43, 0.50]	0.083 [0.07, 0.10]	+0.060 [0.02, 0.10]
round-robin, no inner loop	0.390* [0.37, 0.41]	0.799* [0.78, 0.82]	0.268* [0.25, 0.29]	+0.018 [-0.02, 0.06]
sticky inner loop	0.358* [0.34, 0.38]	0.765* [0.75, 0.78]	0.310* [0.29, 0.33]	+0.211* [0.16, 0.26]
+ persona prose	0.327* [0.31, 0.34]	0.825* [0.81, 0.84]	0.356* [0.34, 0.38]	+0.250* [0.20, 0.30]
+ style coupling	0.415* [0.40, 0.43]	0.725* [0.70, 0.75]	0.297* [0.28, 0.32]	+0.174* [0.13, 0.21]
+ calibrated conviction	0.351* [0.33, 0.37]	0.764* [0.75, 0.78]	0.327* [0.30, 0.35]	+0.250* [0.20, 0.30]
+ real openings	0.441* [0.42, 0.46]	0.717* [0.69, 0.74]	0.230* [0.21, 0.25]	+0.161* [0.12, 0.20]
orchestrated engine	0.643 [0.60, 0.68]	0.429 [0.40, 0.46]	0.138* [0.11, 0.16]	+0.201* [0.14, 0.25]

Table 10: Voice metrics on committed transcripts, with room-level bootstrap intervals (2,000 draws, shared matrix). TTR is type-token ratio (vocabulary breadth, higher is more varied). Overlap is mean pairwise similarity between members of a room (lower is more distinct). Self-repetition is the share of a speaker's word bigrams that repeat. Lift is author-attribution accuracy minus chance. Attribution uses cosine nearest-centroid over character trigrams, leave-one-message-out. *: the paired contrast against real humans excludes zero.

of a real one. Whatever is wrong with these transcripts, it is not that the speakers blur together, and our original reading of H5 was mistaken about the direction.

The defect is vocabulary poverty and repetition. Agents use roughly half the vocabulary breadth of real participants (0.33–0.44 against 0.71) and repeat themselves three to four times as often (0.23–0.36 against 0.07). They are easy to tell apart precisely *because* each is stuck in its own narrow rut. H5's verdict survives — persona prose is still the worst arm, with the highest cross-agent overlap (0.825) and the most self-repetition (0.356) — but the mechanism behind it was misread, because length variance cannot see either quantity.

Orchestration matters here, and it is the largest effect in this section. The orchestrated engine reaches a cross-agent overlap of 0.429 against a real 0.430, while every round-robin arm sits between 0.72 and 0.83. Its vocabulary breadth is 0.643 against a real 0.709, where the others manage 0.33 to 0.44. This is the same engine that §7 shows is *not* distinguishable from the round-robin chassis on aggregate change statistics. So orchestration buys something real that the fidelity metric this paper is built on cannot see, and a paper that reported only composite error would have concluded the engine does not matter.

With intervals, added after review. The first draft reported these as descriptive statistics. Resampling rooms (2,000 draws, one shared matrix, real humans scored on the same 50 rooms) gives the intervals in Table 10. Every generated arm differs from real on vocabulary, overlap and self-repetition at the 95% level, with one exception that is the point of the previous paragraph: the orchestrated engine does not separate from real humans on vocabulary (-0.035 $[-0.077, +0.010]$) or on overlap (-0.035 $[-0.083, +0.011]$), though it still repeats itself more ($+0.055$ $[+0.028, +0.083]$). Against the sticky round-robin chassis the engine's advantage resolves on all three: vocabulary $+0.285$ $[+0.247, +0.324]$, overlap -0.336 $[-0.373, -0.298]$, self-repetition -0.172 $[-0.198, -0.147]$. H5's verdict also survives: persona prose is worse than the chassis it was added to on vocabulary (-0.031 $[-0.055, -0.007]$), overlap ($+0.060$ $[+0.043, +0.077]$) and self-repetition ($+0.046$ $[+0.018, +0.075]$). Over-distinguishability resolves for every sticky-loop

arm (lift +0.16 to +0.25 against real +0.06 [+0.02, +0.10]) and not for plain round-robin (+0.018 [−0.018, +0.055]); between generated arms, lift mostly does not separate.

Attribution and repetition are mechanically linked — a repetitive speaker is easy to attribute — which is exactly why we report overlap and self-repetition alongside, and why we read the three together rather than any one alone.

B Tables and figure moved from the main text

The main text summarises each of these; they are here in full so the compact build keeps the closing sections inside the reviewer's window.

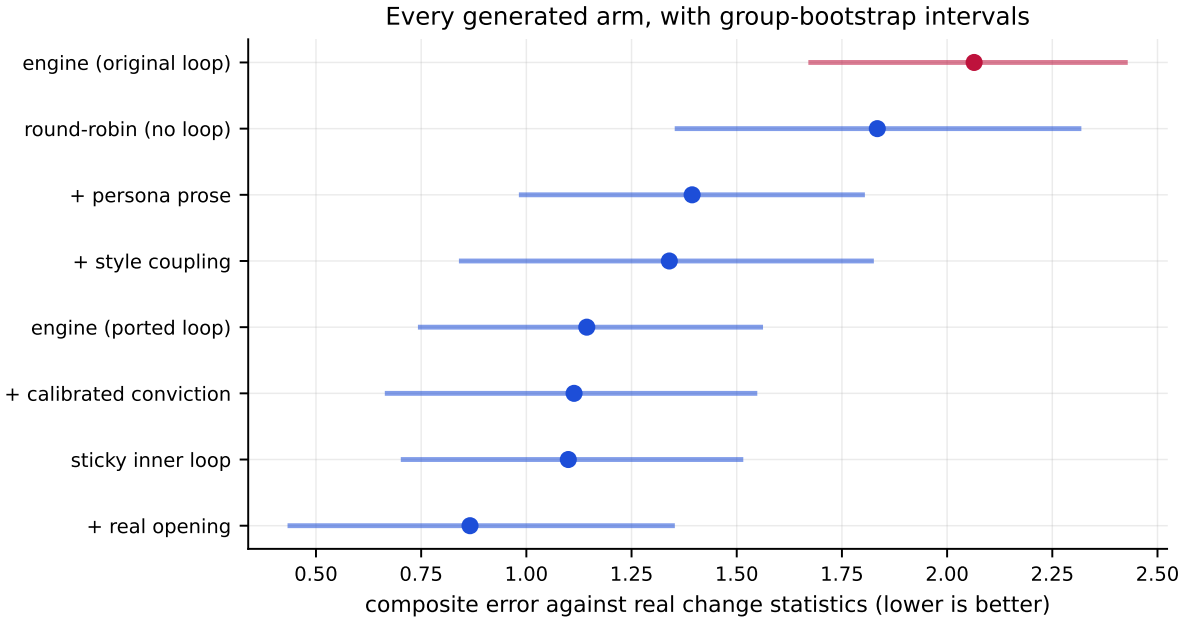

Figure 2: Every arm with its *marginal* interval. Marginal intervals overlap much more than paired contrasts do: the engine arm (red) shares almost all of its interval with round-robin and does not separate from it (+0.230 [−0.259, +0.704]), yet separates from six of the other seven arms on the paired grid. This figure shows spread; §7 decides separations.

C The audit

Before writing this paper we re-checked its two notebooks against their own committed run data: five agents, 510 claims, each discrepancy confirmed by us independently before correction. Seven were blocking, meaning a paper stating them would have stated something false. Three changed conclusions.

The first invented a research requirement. A notebook reported that every generated arm showed zero churn among members who started *correct*, against a real 21%, and concluded that realistic simulation needs a non-epistemic churn channel that no arm had. Recomputing from the committed results, the ported engine changes 4 of 19 such members — 0.2105, exactly the real rate. Only the round-robin think arms are at zero. The arm that met the requirement was reported

Pair (first − second)	Δ	95% CI	boot. p	Holm	BH
engine (original loop) vs engine (ported loop)	+0.921	[+0.429, +1.412]	0.001	yes	yes
engine (original loop) vs + real openings	+1.198	[+0.590, +1.753]	0.001	yes	yes
engine (original loop) vs sticky loop	+0.965	[+0.363, +1.558]	0.002	no	yes
+ real openings vs round-robin	−0.968	[−1.544, −0.404]	0.002	no	yes
+ calibrated conviction vs engine (original loop)	−0.951	[−1.544, −0.332]	0.003	no	yes
engine (ported loop) vs round-robin	−0.691	[−1.250, −0.136]	0.012	no	no
+ style coupling vs engine (original loop)	−0.724	[−1.313, −0.135]	0.018	no	no
engine (original loop) vs + persona prose	+0.670	[+0.120, +1.216]	0.018	no	no
round-robin vs sticky loop	+0.734	[+0.075, +1.406]	0.029	no	no
+ real openings vs + persona prose	−0.528	[−1.021, −0.039]	0.038	no	no
+ style coupling vs + real openings	+0.474	[+0.006, +0.922]	0.047	no	no

Table 11: The 11 unadjusted separations on the 28-pair grid, ordered by bootstrap p. "Holm" and "BH" say whether the pair survives family-wise ($\alpha = 0.05$) and false-discovery-rate (0.05) correction across all 28 pairs.

Composite variant	headline Δ	95% CI	verdict	pre-reg. resolving	grid resolving
the paper: floor 0.05, unweighted mean	−0.921	[−1.409, −0.428]	lower	1/6	11/28
same, real target resampled with the arm	−0.921	[−1.891, −0.411]	lower	1/6	11/28
floor 0.10	−0.911	[−1.395, −0.424]	lower	1/6	11/28
floor 0.20	−0.764	[−1.160, −0.356]	lower	1/6	11/28
no floor	−0.921	[−1.409, −0.428]	lower	1/6	11/28
each deviation in real-statistic SD units	−3.937	[−5.535, −2.271]	lower	1/6	13/28
max deviation (L_∞)	−1.167	[−2.129, −0.172]	lower	1/6	5/28
root-mean-square	−0.959	[−1.532, −0.370]	lower	1/6	9/28
change + rescue only	−0.688	[−1.161, −0.188]	lower	1/6	9/28
gain + lift only	−1.153	[−1.761, −0.556]	lower	1/6	9/28
leave out score gain	−0.900	[−1.383, −0.401]	lower	1/6	9/28
leave out change rate	−1.157	[−1.827, −0.501]	lower	1/6	10/28
leave out correct lift	−0.787	[−1.233, −0.316]	lower	1/6	12/28
leave out rescue rate	−0.839	[−1.240, −0.443]	lower	1/6	10/28

Table 12: The headline contrast under every definition of composite error we tried, with how many of the six pre-registered contrasts and of the 28 grid pairs resolve under each. Same 2,000-draw resample matrix throughout; the first row reproduces Table 4.

Arm	change	net of C	95% CI	w2c	net of C	95% CI
engine, original inner loop	0.821	+0.058	[+0.003, +0.112]	0.730	+0.157	[+0.016, +0.293]
round-robin, no inner loop	0.904	+0.141	[+0.092, +0.191]	0.650	+0.077	[−0.075, +0.227]
round-robin + sticky inner loop	0.590	−0.173	[−0.257, −0.093]	0.438	−0.135	[−0.270, +0.011]
+ fusion persona prose (H5)	0.705	−0.058	[−0.134, +0.013]	0.504	−0.069	[−0.206, +0.060]
+ quantitative style coupling (H6)	0.667	−0.096	[−0.204, +0.006]	0.547	−0.025	[−0.197, +0.145]
+ population-calibrated conviction (H7a)	0.615	−0.147	[−0.235, −0.065]	0.474	−0.099	[−0.248, +0.040]
+ real 5-message opening (H7b)	0.635	−0.128	[−0.209, −0.049]	0.423	−0.150	[−0.295, +0.000]
engine, ported inner loop	0.596	−0.167	[−0.260, −0.073]	0.453	−0.120	[−0.274, +0.038]
solo control C	0.763			0.573		
real groups	0.641			0.256		

Table 13: Change rate and wrong-to-correct conversion for every arm, gross and net of the solo control C (arm − C, paired group bootstrap). C is deepseek answering alone with two repeats pooled.

Model (sticky loop)	msg words	thought words	concession	odd in thought	in speech	in final	final tracks thou
deepseek think	29	49	0.090	0.83	0.71	0.42	0.66
Qwen 0.8b think	55	162	0.219	1.00	0.98	0.21	0.28
Qwen 2b think	23	113	0.085	0.94	0.73	0.30	0.43
Qwen 4b think	17	54	0.059	0.78	0.61	0.16	0.52
Qwen 9b-local think	28	56	0.077	0.82	0.66	0.13	0.46
Llama 8B think	32	64	0.296	0.57	0.62	0.27	0.85

Table 14: Sticky-loop diagnostics from committed artifacts, same chassis and 50 groups. "Odd in thought / speech / final" is the share of wrongly-seeded members whose initial selection lacked the odd-number card and whose private thoughts, public messages, or elicited final selection contain it. "Final tracks thought" is agreement between the final's inclusion of that card and the last private thought's mention of it.

Solo w2c	no sheet	verbatim, system prompt (the paper)	verbatim, user turn	paraphrase, system prompt	pa
2B	0.066 / 0.095	0.153 / 0.197	0.131	0.080	
4B	0.263 / 0.285	0.606 / 0.569	0.394	0.679	
9B	0.212 / 0.190	0.650	0.438	0.467	

Table 15: Solo wrong-to-correct conversion (the H10 activation measure) under four wordings and placements of the task sheet. Two values are two generation seeds. Same 50 groups, same elicitor.

four paragraphs below the claim that nothing met it. Churn here is a property of orchestration, not a missing module.

The second was a headline built on a number that was never measured. The ignorance dividend was reported as 17×, from a large-model baseline of roughly 0.75. That baseline exists in no artifact: the collection code hard-codes a null for that arm, so the column is empty. Recomputing with the project's own scoring functions gives 0.573 and a dividend of 13.0×. The tell is that $0.75/0.0441 = 17.0$ exactly — the ratio was derived from the invented baseline rather than measured.

The third was a superlative contradicted elsewhere in the same document, which is the pattern in most of the rest. A heading or summary line overstates what the body text below it correctly says, and the heading is the version that propagates into downstream documents. Three of the seven took that form, and it recurred in a companion paper in this program.

We report this for two reasons. It is the reason §7 exists, and the corrections all run in the direction of weaker claims, which is not the direction unreported errors usually run.

D Preregistration, locks, and amendments

Every stage locked hypotheses, arms, metrics and decision rules before any paid call. Verdicts are three-valued and, as of this paper, decided on intervals; where the original notebooks decided on point estimates, §7 restates them.

Locks that failed. H3 was refuted on the original engine and supported only after the inner-loop fix. H4's specific optimum was refuted while its monotonicity prediction held — forgetting turned out to be load-bearing rather than a cost saving. H5's style-heterogeneity prediction was refuted with the sign inverted. H6 hit its surface targets and exposed a conflation between spoken concession and submitted belief. H7b's member-match threshold was missed. H8's interior

optimum was refuted in favour of its own pre-registered alternative. H9 splits: trend supported, monotone form falsified in every bootstrap draw. H10 was refuted with the mechanism inverted. H11's prediction that dynamics track the speaking model was refuted.

Amendment, resolved. The H4 grid as written names five configurations under an argmin rule, and the results list four. The omitted one is the no-memory-dials baseline, and it scores better than the selected winner on the tuning slice (0.623 against 0.700), so the rule applied literally selects no memory at all. The rule was written more broadly than intended — H4 asks which memory setting is best, and the baseline was a reference rather than a candidate — but the distinction does not matter, because the two slices disagree on the ordering: the baseline wins on the tuning slice and the memory configuration wins on the held-out slice (0.902 against 0.921). That is a tie, consistent with §7. We therefore report the frozen memory setting as the best *memory* configuration measured and not as an improvement over having none.

Known leakage. Conviction gradients for H7a were fit on 448 non-pilot groups, disjoint from the 50 scored here.

Reproducibility. Every prompt, the round-robin chassis, the orchestrated engine adapter, the scoring code, and every per-group result behind every number in this paper are committed in the research repository that accompanies it, together with the bootstrap, sensitivity, multiplicity, voice, netting and diagnostic scripts that produce the tables from those results. Per-metric per-arm deviations with intervals are in the committed `sensitivity.json`, so any reweighting a reader prefers can be computed without re-running a model. The repository is to be released with the paper.

References

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023. doi: 10.1017/pan.2023.2.

Blaž Bertalanič and Carolina Fortuna. The cost of consensus: Isolated self-correction prevails over unguided homogeneous multi-agent debate, 2026. URL `https://arxiv.org/abs/2605.00914`. Accepted at the ACM Conference on AI and Agentic Systems (CAIS) 2026.

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.

Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. Debate or vote: Which yields better decisions in multi-agent large language models? In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL `http://papers.nips.cc/paper_files/paper/2025/hash/934252acd87f254d5d4672fbde283bd2-Abstract-Conference.html`.

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.

Xiqi Hao, Zengqing Wu, Yu-Xuan Qiu, Chuan Xiao, Ruiqi Xu, Shuyuan Zheng, and Jianbin Qin. Not all flips are conformity: Decomposing stance convergence in multi-agent LLM debate, May 2026. URL `https://arxiv.org/abs/2606.00820`.

Haitao Jiang, Lin Ge, Hengrui Cai, and Rui Song. PABU: Progress-aware belief update for efficient LLM agents, 2026.

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.

Andrew K Lampinen, Ishita Dasgupta, Stephanie C Y Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, 3(7):pgae233, 2024. doi: 10.1093/pnasnexus/pgae233.

Zhenghao Li, Zhi Zheng, Wei Chen, Jielun Zhao, Yong Chen, Tong Xu, and Enhong Chen. Dynadebate: Breaking homogeneity in multi-agent debate with dynamic path generation, 2026. URL `https://arxiv.org/abs/2601.05746`.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, 2023. doi: 10.1145/3586183.3606763. arXiv:2304.03442.

Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. Generative agent simulations of 1,000 people, 2024. arXiv:2411.10109v1.

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

Haotian Wang, Xiyuan Du, Weijiang Yu, Qianglong Chen, Kun Zhu, Zheng Chu, Lian Yan, and Yi Guan. Learning to break: Knowledge-enhanced reasoning in multi-agent debate system, 2023. URL `https://arxiv.org/abs/2312.04854`.

Andrea Wynn, Harsh Satija, and Gillian Hadfield. Talk isn't always cheap: Understanding failure modes in multi-agent debate, 2025. URL `https://arxiv.org/abs/2509.05396`. ICML MAS Workshop 2025.

Wenhao Yuan, Chenchen Lin, Jian Chen, Jinfeng Xu, Xuehe Wang, and Edith Cheuk Han Ngai. Verify before you commit: Towards faithful reasoning in LLM agents via self-auditing. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026. arXiv:2604.08401.