Evidence in the Room: What It Takes to Make a Simulated Deliberator Change Its Mind

Jason Grey
jason@jason-grey.com

August 2026

Abstract

Simulated deliberators are famously easy to move and famously hard to move for the right reasons. We study the second problem on a corpus where the reason is recoverable: 369,215 Wikipedia Articles-for-Deletion debates, in which participants state a position, argue, and sometimes visibly change it. We first report the participant-level opinion-change statistics for this corpus, which to our knowledge have not been published: 1.45% of voters cast a second position, 63% of those genuinely switch, and switches run 1.9:1 delete→keep because the mechanism is rescue by newly produced sources.

That asymmetry sets up the paper. Reading is not the bottleneck — a 9B model recovers the actual switched-to position of 29 of 33 real switchers from the debate text alone (0.879, rising to 0.939 under prompt repetition). Generation is: with the identical inner loop that had to be *restrained* on a prior corpus, agents seeded with their own real rationales ossify at 0.09× the real change rate, because the principal real cause of switching — somebody walking in with a source nobody had seen — has no in-simulation analogue. We then restore that cause one mechanism at a time, each isolated by a paired arm: retrieval as a choice inside the think step, commitment-gated research propensity (the arrival arm records 0.82 searches per uncommitted turn in contested debates against 3% for committed agents in the preceding arm; role-counted re-runs and a within-run randomisation added after review find the role effect real, 1.4 to 2.7-fold, and roughly half the size the arm's own metric implied), evidence-bearing arrivals, and archived era-correct document content. Retrieval-as-choice doubled change among committed agents from a very low base; only document content moved them substantially.

Two results run against the field's intuitions, and we state first what resolves and then what does not. **More evidence is not shown to be better evidence.** On a full evidence-volume × model-size factorial, a debate-level bootstrap resolves a level rather than a ranking: three of the nine cells clear the pre-registered floor in a majority of resamples — 2B + titles (79% of draws), 122B + titles (93%) and 122B + content (77%) — while every no-evidence cell, every 9B cell and 2B + content do not. *None* of the orderings among cells resolves, and six generation seeds per cell, run as a pre-registered test, show why: between-seed standard deviations of 0.07 to 0.30 in ratio units, the same order as the interval over debates. The ordering we had reported from three seeds — titles above full excerpts — decays toward zero as seeds accumulate and is withdrawn (+0.067, +0.056, +0.022 at 2B, 9B and 122B, none resolving), as is a reading that the source's domain carried it, which survives only against an unlucky single seed. What survives is a statement about levels: evidence-bearing rooms reach the pre-registered band on most of their seeds (9B and 122B on four to five of six) and bare-prompt rooms do not (zero or one of three). A power projection over the committed results says the size axis is answerable with about 160 content-bearing debates and the volume axis is not answerable on this corpus at all. Token-matched controls added after review say volume is not the axis

(a 13-word summary, a 100-word excerpt and a 170-word excerpt score identically) and that masking the source's domain gives back most of the title advantage. The largest model with the *most* evidence scores lower than both and reverses direction to 2:4 keep→delete as the big reader judges the thin article on its merits, though the direction counts are single digits. The pre-registered guess that 9B would be the smallest in-band size was wrong twice. **Repetition improves perception, not persuasion.** Doubling prompts lifts reading accuracy in every cell (+0.06–0.09) and lifts belief change by exactly nothing under paired seeds.

Finally we move belief out of the prompt into an external log-odds state fitted per ideology bucket from 453 extracted messages, which independently recovers the corpus's measured ideology ordering. Composed with a verified-evidence gate the architecture reaches exact magnitude parity (0.3125 against a real 0.3125), and a vote-sovereignty constraint — agents re-vote only when belief tips — then reveals that about 27% of the change volume had been speech-level concession rather than belief change. We report throughout against a measured spontaneous-instability floor, which forces our own central magnitude claim down from 0.267× gross to 0.170× net, below the floor we pre-registered; the honest statement is *in band gross, marginal net*, with the true value bracketed between the two.

1 Introduction

There are two ways to make a simulated group change its mind, and only one of them is worth anything. You can make the agents pliable, in which case they converge on whatever the room says and the simulation reproduces group dynamics by having no dynamics of its own. Or you can make them change for the reason real people change, which requires knowing what that reason was.

Most deliberation corpora do not tell you. A group discusses, positions move, and the cause is distributed across the transcript in a way no annotation recovers. Wikipedia's Articles-for-Deletion process is unusual in that the cause is frequently written down: participants state a position in boldface, argue about whether a subject meets the notability guidelines, and when they switch they typically say why, often naming the source that changed their mind. The corpus is large, the outcome is decided by an administrator whose behaviour is close to a documented aggregation rule, and the material people argue *about* — the article, the sources, the policy pages — is separable from the material they argue *with*.

That separation is what this paper needs. Prior work in this program established that on a task whose structure *is* the knowledge, handing an agent any description of the task activates latent competence: a 4B model's solo drift tripled when given a definitions sheet, which makes "grounding" and "teaching" indistinguishable. Evidence grounding is only measurable when the evidence is orthogonal to the task's logical structure — documents a person *consults*, not scaffolding that *instructs*. On deletion debates it is: the policy is stable and known, and what is contested is whether particular sources about a particular subject clear it.

We use that to ask a product question with a research answer. If a simulation platform must choose a model size, and evidence retrieval is available, what is the right pairing? The intuition the field encourages — bigger model, more evidence — is wrong in both halves on this corpus.

Our contributions:

1. **Participant-level opinion-change statistics for the AfD corpus**, which the corpus's source paper did not compute and which we believe are unpublished: change rates, the direction asymmetry, and outcome-stratified conversion gradients by editor ideology and experience (§4). These are the calibration targets everything else is scored against.

2. **A causal chain for simulated opinion change, verified link by link**, each link isolated by a paired arm that changes exactly one thing: reading fidelity, retrieval as a choice, commitment-gated research propensity, arrivals, and document content (§5, §6). We report the three that failed as well as the two that worked.

3. **An evidence-volume × model-size surface** on which the pre-registered monotone-in-volume ordering holds strictly at no size. On point estimates titles out-score full content at both ends of the ladder, but no ordering on the surface resolves at 19 content-bearing debates; what resolves is that three cells clear the pre-registered floor in a majority of resamples, including the smallest model with the lightest evidence, and six do not (§9); six generation seeds per cell retire every ordering the single-seed surface suggested, including the 9B trough, the smallest-model superlative and titles above excerpts; what remains is a level result, and a measured statement of the study size the question would need (§10.2, §10.3).

4. **A dissociation between perception and persuasion**: prompt repetition, which reliably improves comprehension, moves belief change not at all under paired seeds (§7).

5. **An external belief-state architecture** — conviction, fitted log-odds, a verified-evidence gate, and vote sovereignty — with each layer's contribution measured separately, reaching exact magnitude parity and localising the residual direction gap to source valence (§11).

6. **A spontaneous-instability control** that revises our own headline downward, and a pre-write-up audit that found six numerical disagreements between our notebook and our own committed data (§8.1, Appendix C).

Where each number lives. Table 5, in Appendix B, gives for every headline number the section that establishes it, the slice it is computed on, and the committed artifact it is read from, so a reader need not track slice changes across sections.

2 Related work

The corpus and its source paper. Mayfield and Black (2019) released the AfD corpus and used it for *discriminative forecasting*: frozen embeddings and logistic regression predicting the administrator's closure from the debate so far, with a "forecast shift" metric as an analytic lens. They report the aggregates we reproduce in §4 — outcome base rates, herding by ordinal position, and administrators behaving as near vote-counters — and they explicitly do not attempt generation, agent simulation, vote-switch detection, or any evidence condition. Participant-level opinion change is the gap we fill.

Simulated deliberation and its known pathology. The dominant failure mode in multi-agent deliberation is over-convergence, though the reported mechanisms differ. Chuang et al. (2024) find networks of LLM agents converging on the accurate answer, driven by an inherent bias toward producing correct information rather than by peer pressure — and recover fragmentation once confirmation bias is induced by prompt, so convergence there is a truthfulness prior rather than an inevitable social pathology. Taubenfeld et al. (2024) report simulated debates collapsing toward agreement and drifting to model bias when unmanaged. Sycophancy toward an interlocutor is well documented (Sharma et al., 2024), and conformity to peers scales with majority size and interaction time (Weng et al., 2025), with agents abandoning assigned personas under group pressure (Baltaji

et al., 2024). Our prior work on a different corpus (Grey, 2026) found the same pathology and fixed it with a sticky-misconception inner loop. The finding that motivates this paper is that the *identical* loop, moved to AfD, fails in the opposite direction: it ossifies. Over-convergence is not a property of the loop but of the interaction between the loop and what the corpus supplies as a reason to move.

Personas and their limits. That simulated individuals reproduce population-level distributions better than they reproduce individuals is by now well established (Argyle et al., 2023; Park et al., 2024; Bisbee et al., 2024), and our own program has measured where the boundary sits. This paper inherits one consequence: descriptive persona prose is close to inert, and the levers that work are mechanical constraints on what an agent may do (Grey, 2026). Every dial here is enforced at a decision point, never asserted in a character sketch.

Computational work on deletion debates is discriminative or observational. Huq and Ciampaglia (2021) characterise opinion dynamics and group decision making in Wikipedia content discussions, including the inclusionist/deletionist axis we borrow for our per-editor ideology measure. Kaffee et al. (2023) do transparent, policy-grounded stance detection on multilingual editor discussions, and Borkakoty and Espinosa-Anke (2025) apply text classifiers to the same material. All three predict or label; none simulates a participant, and none reports participant-level position change. The gap we fill is generative and calibrated to individuals rather than to labels.

Validating simulated deliberation against real transcripts. The closest methodological neighbour is DEBATE (Chuang et al., 2025), which evaluates role-playing agents against a large corpus of real human debate messages with pre- and post-discussion beliefs, and independently reports the over-convergence we discuss above. We differ in corpus and in what is being calibrated: our targets are per-editor conversion gradients on a naturalistic governance process rather than benchmark-wide opinion trajectories. Retrieval-augmented role-play scored against real decisions exists in the legal domain — AgentsCourt (He et al., 2024) simulates court debate over a purpose-built legal knowledge base and is evaluated on Chinese judgment documents, with its reported gains on legal-article generation rather than on verdicts — and evidence-pool asymmetry across agents has been studied for forecasting accuracy (Li et al., 2026), though on binary prediction-market questions rather than on stance change. None of this work crosses evidence volume with model size, and none calibrates to vote-change statistics on a governance corpus.

How much evidence, and in what form. The closest controlled study of evidence budget is Laitenberger et al. (2025), who find that with long-context models a simple ordered-retrieval baseline is competitive with more elaborate RAG pipelines, and that how much is retrieved and in what order matters more than the machinery around it. Our surface asks a narrower version of the same question inside a deliberation loop, where the retrieved material competes with a conversation for the agent's attention rather than being the whole input.

We flag one asymmetry honestly, because it cuts against us. Multi-agent systems that add a shared retrieval pool report that *more* shared evidence helps: MADKE (Wang et al., 2025) improves accuracy by giving debating agents access to a common knowledge pool, which is the opposite valence to our point-estimate finding that titles out-score full excerpts. Their setting differs — factual QA with agents holding divergent knowledge, rather than stance change under a fixed corpus — but we would not want a reader to take our surface as evidence against retrieval volume in general, particularly given §A.3 shows our own orderings do not resolve. Nguyen et al.

(2025) is architecturally adjacent: multi-agent RAG systems already compress evidence before it reaches the reasoning agents, which is a design answer to the same pressure our surface probes, and MASS-RAG (Xiao et al., 2026) makes the compression explicit with separate summarising, extracting and reasoning agents joined by a synthesis stage. Relative to that line, what we add is orthogonal rather than competing: a gate on *judgement* that decides which retrieved material is allowed to move a belief, which could sit downstream of any of those retrieval pipelines.

Belief as an explicit state. Our final study adopts the architecture of Belief Engine (Yang et al., 2026), which places stance outside the model as an inspectable evidential state with configurable uptake and anchoring, rather than leaving it implicit in prompt language. Our contribution there is not the mechanism but its decomposition: we measure what each layer contributes on a corpus with a known ground-truth direction, which is what isolates the verified-evidence channel and the speech-versus-vote split. Two recent frameworks give that decomposition a vocabulary we did not have when we ran it. Park (2026) implements AGM-style belief revision over a versioned agent memory and tests the postulates operationally; our verified-evidence gate is a relevance condition of that kind (only produced, provenance-checked material revises), and vote sovereignty is a minimal-change condition (a public position moves only when the belief behind it does). Myakala et al. (2026) benchmark belief consistency and drift over multi-session interaction with metrics for revision accuracy, drift coherence and evidence sensitivity; our perception-versus-persuasion dissociation (§7) and the instability control (§8.1) measure the same two quantities on a naturalistic corpus, and their metrics would be the natural way to report them if this work were extended to multi-session belief.

Two results we lean on directly. Lampinen et al. (2024) show that language models, like humans, exhibit content effects on reasoning — believable material is processed differently from unbelievable material, which is both a facilitation and a bias, since models also endorse invalid syllogisms with believable conclusions. That is the raw material motivated reading would need, though it concerns the model's own pretraining-derived beliefs rather than an assigned persona's, so we treat it as background rather than support. Leviathan et al. (2025) report that duplicating a prompt improves non-reasoning model performance at no extra generated tokens or latency; §7 tests that on our instrument and, importantly, tests whether it reaches behaviour as well as comprehension. We note the paper concerns prompt duplication only and says nothing about repeated assertions carrying social weight.

3 Definitions and implementation

Round-one review asked for several mechanisms to be specified rather than described, and for a glossary. Both are short enough to give in full.

Terms. A debate is **content-bearing** if at least one source harvested from it has archived era-correct text, as opposed to a title and domain only; 19 of the content-rich-slice debates qualify and they carry every magnitude claim in this paper. We call that archived content the **archived text** of a source — the thing a simulated arrival can actually quote. A voter is **committed at open** if their real first vote fell in the debate's early half, so they are seeded with a public position; the rest become **lurkers**, present and uncommitted. **Net** ratios subtract the size's own measured spontaneous-instability rate from the simulated change rate before dividing by the real rate; **gross**

does not. The original surface uses one run of the isolation control per size; three re-runs per size are given with the seed surface (§10.2).

Models, exactly. Every generation run uses one family, Qwen3.5, at three sizes. 2B and 9B are dense models; "122B-A10B" is Qwen3.5-122B-A10B, a sparse mixture-of-experts with 122B total and 10B active parameters on a hybrid linear-attention architecture with a 262,144-token context window, served through OpenRouter with reasoning disabled. The 9B and 122B surface cells are hosted; the 2B cells and the local replications in §10 run through Ollama at 4-bit quantisation on one consumer GPU. Sampling is temperature 0.7 with a 300-token cap on every call, the companion paper's instrument unchanged (Grey, 2026).

The verified-evidence gate, exactly. There is no classifier and no judgement about what counts as keep-leaning. The gate is chassis ground truth: the harness knows which agent consumed which document on which turn, because it served it. A turn is marked verified iff the agent's preceding think step executed a search that returned a hit *with archived content* — in the implementation, the payload delivered to that agent contains an excerpt. Belief updates ingest only messages from a verified turn. Unverified evidence claims and assertions of absence ("I found nothing") still appear in the conversation and still influence what other agents say; they simply do not move the belief state. So misclassification is not a failure mode here, because nothing is classified. What the gate cannot do is distinguish an agent that read a source from one that read it and misrepresented it, and we do not audit that.

Retrieval, and what stops it leaking. Sources are harvested from the citation records and link text of the real debate, so the candidate pool is era-correct by construction. Content is resolved through archival snapshots nearest the debate's close and truncated to roughly 200 words. Live web search is barred from every calibration run, for a reason worth stating: searching a 2010 subject in 2026 surfaces the AfD discussion itself, which is outcome leakage rather than evidence. Titles are derived from link text or URL slug and never from the citing editor's argument, which would be replay contamination. Coverage is the binding constraint and is not uniform: only 14 of 37 harvested sources in the first pilot had archived content, and content availability correlates with subject prominence, so the content-bearing subset is selected on a property plausibly related to the outcome. We report that as a limitation rather than correct for it.

Ideology and experience, and how they were assigned. Both are derived, not annotated. An editor's ideology is their lifetime delete share across all their AfD votes: at or above 0.7 is deletionist, at or below 0.3 inclusionist, otherwise moderate, with the bucket withheld entirely below 10 lifetime votes (the low-history bucket). Experience is edit count. There is no manual validation and no held-out predictive check of these labels, which is a real gap: misclassification would attenuate the conviction gradients rather than manufacture them, but we have not measured how much.

On token budget, which the review raised. The "full source content" arm is not a full document. Harvested content is truncated to roughly 200 words and the search payload is capped further, so the contrast is between a title plus domain and a short excerpt — not between a headline and a whole article. That weakens the metabolic-cost reading of §9 considerably: whatever separates the arms, it is not the load of reading a long document, because neither arm delivers

one. It also means the review's suggested fix — equalising token counts via abstractive summaries — is closer to what we already ran than we had realised, and the honest statement is that we have not decoupled volume from framing at all.

4 Corpus, instrument, and the targets everything is scored against

Why this corpus. We required three things at once: real groups with measurable initial and final individual positions, a genuine document pool the real participants consulted, and scale. AfD is the only candidate we found that has all three. Council-minutes corpora have excellent document pools and no pre-discussion positions; email corpora have real people and no measured decisions; the deliberation corpus used in our prior work has measured positions and no documents, because on that task the evidence *is* the rule.

Porting the instrument. A debate becomes a group; participants become members; a member's initial stance is their first bolded position with its rationale, and their final stance is their last standing one, with struck and amended votes tracked. The outcome is the administrator's close.

Reproduction. Table 10 compares our parse to the source paper's published aggregates. Outcome shares and administrator behaviour land close. Vote shares sit about 5 percentage points off, and three statistics disagree substantially: ties are 0.034 of debates for us against their 0.076, ties resolve to delete 0.378 of the time against their 0.669, and a first keep vote succeeds 0.756 of the time against their 0.622. All three are traceable to differing final-vote, tie and first-vote definitions. We list all of them because an earlier draft of this paragraph reported only the agreements, and the draft after that still omitted the first-keep gap while claiming to name the largest.

Opinion change on AfD. Table 9 gives the statistics this corpus has not previously supplied. Change is rare and concentrated: only 1.45% of voters ever cast a second position, but 63% of those genuinely switch, so the act of returning to a debate is most of the signal. Overall change is 0.91% of voters, rising to 1.13% when a nominator's presumed initial delete position is counted; 3.8% of debates contain a switch, or 5.5% with the same presumption.

The direction asymmetry, and why it drives the whole paper. Switches run 1.9:1 delete→keep (8,547 against 4,535), and counting presumptive nominators raises the delete→keep count 1.7-fold to 14,448 — withdrawal by softening is the corpus's single largest change channel. The asymmetry is not a quirk of aggregation. It is what the mechanism looks like: people change their minds on AfD because somebody produces sources that were not previously in the room, and produced sources argue for keeping. Citations following an opposed vote lift toward-keep conversion from 2.99% to 4.75%, and for self-identified deletionists specifically from 2.85% to 4.51%, while toward-delete conversion barely moves.

Gradients, which become the calibration targets. Stratified by outcome, the identity-protective pattern is clean: each camp concedes readily toward its own pole and resists toward the opposing one. Deletionists asked to move toward delete concede at 0.88 conditional on returning; inclusionists asked to move toward delete mostly disengage instead, returning at 1.6% against a base rate roughly twice that, and converting at 0.88% overall. Experience separates the same way, though only in one direction: veterans concede when beaten (0.80–0.82 conditional on returning,

both directions) while novices who return still refuse *toward delete* (0.26–0.29). Toward keep, novices concede at 0.75, close to veterans' 0.82 — the experience gradient is a gradient in resisting deletion, not in conceding generally. These per-bucket numbers are what the conviction dial in §8 is fitted to, and the who-changes signature they predict is a lock we score against.

The observability rule. The corpus can only see a switch if the editor re-voted. Every simulated run is therefore scored under the same censoring: we count only switches the real instrument could have observed. This is locked before any simulation and is why we report switcher recall and false-switch rates rather than overall final-position match, which is degenerate here — predicting that everyone's final equals their first scores above 97%.

Evaluation population and pilot. 289,712 debates are at group scale, 8,314 contain a genuine switch, and 207,286 carry a non-signature citation. Our pilot is a deterministic stratified sample of 60 debates: 30 containing a genuine switch and 30 without, with strata reported separately throughout, since switch debates are 50% of the pilot against a 2.9% base rate, an oversampling of roughly 17-fold and pooling would fabricate a target.

5 Reading is not the bottleneck, and documents are expensive

Before asking whether simulated agents change their minds correctly, we check whether the model can read the debate at all. Two arms on the pilot: a control given only the title, the nomination, and the voter's own first vote and rationale, and a replay arm given the full real debate. If the replay arm cannot recover real switchers, nothing downstream is interpretable.

The replay anchor is strong. Across 271 voters including 33 genuine switchers, with 542 calls per size and zero parse failures anywhere, replay recovers the actual switched-to position of 29 of 33 real switchers at both 9B and 122B (recall 0.879), against 0.636 at 2B. Control recall never exceeds 0.121 (0.121 at 2B, 0.000 at 9B, 0.030 at 122B). A model given only a voter's own reasoning almost never invents that voter's switch, which is the correct behaviour and makes the replay-versus-control gap a clean measure of how much influence content the debate text carries. The gradient across sizes runs 0.636, 0.879, 0.879 — rising and then plateauing, the same shape comprehension takes elsewhere in this program. Real AfD influence content is at least as machine-readable as casual chat, despite threading and a formal register.

Injecting documents makes small readers worse. We then repeated both arms with the texts of the policies actually cited in each debate injected — three pages, roughly 500 to 1500 words. Table 8 gives the result, and it inverts the hypothesis we pre-registered. Policy text *costs* switcher recall, and the cost scales inversely with reader capacity: -0.242 at 2B, -0.152 at 9B, and nothing at all at 122B. The 2B's false-switch rate rises with it, from 0.246 to 0.336. The small reader does not get help from documents; it drowns in them.

The same run rules out the confound that would have invalidated the paper. We pre-registered this arm as a transfer test for the activation pathology described in §1: if policy text is really scaffolding rather than evidence, then giving it to the *control* arm — which never sees the debate — should let it guess outcomes better, exactly as a task-definitions sheet did on our previous corpus. It does not. Control outcome-match moves by at most $+0.012$ at any size, and

drops at 2B. On a task whose outcome depends on facts about a subject rather than on knowing a rule, evidence and scaffolding come apart, and the distinction becomes operational rather than definitional. This is the negative result the rest of the paper rests on.

6 Generation ossifies, and restoring change takes four mechanisms

We now generate rather than replay. Each real voter becomes an agent, the nomination and every voter's real first vote are seeded as the debate's opening — on AfD initial positions are public posts, so this is initial-state seeding rather than leakage — and the continuation is generated round-robin with a think step before each speech act. A generated message may open with a bold stance marker, which constitutes a public re-vote and updates that agent's standing position in everyone's context; otherwise it is a comment. Finals are the last public stance, which makes scoring censoring-symmetric with the corpus by construction.

The sim freezes. Simulated change comes out at 0.018 against a real stratum rate of 0.198 — 0.09×, against a pre-registered floor of 0.25×. Outcome match is nonetheless excellent (0.967 on the plain stratum), because with real openings seeded, majority rule and herding carry outcomes without anybody needing to change their mind.

The inversion is the finding. On our previous corpus this same inner loop had to be restrained: capitulation was the failure mode, and the engineering was all about making agents hold. Here it ossifies. The mechanistic difference is visible in §4: real AfD switches are evidence-driven, and a generated continuation *cannot find new sources*. The principal real-world cause of switching has no in-simulation analogue, so agents seeded with their real rationales correctly hold their positions against mere argument. On this corpus, evidence is not garnish; it is the engine of change.

Four mechanisms, added one at a time. Each of the following is a paired arm changing exactly one thing, so its delta is attributable.

1. **Retrieval as a choice** (not force-feeding, per §5). The think prompt lists available material by title; an agent may end its thought with a consult tag, triggering one follow-up think call with that document injected. Agents consult on 18.0% of turns in contested debates against 3.3% in easy ones — selective, emergent, and uninstructed. Change rate doubles, 0.012 to 0.023. Still 0.105× real. A register-calibration control confirms the doubling is the evidence and not the register: calibrated-but-evidence-free stays frozen at 0.05×.

2. **Search, modelling the product's own lookup.** An era-correct backend returns the external sources actually cited in that real debate, as harvested titles and domains, never the citing editor's argument. Live web search is barred from calibration runs, since searching a 2010 subject in 2026 surfaces the AfD page itself. This *fails*: agents narrate wanting to search while almost never executing it, and a phrasing iteration lifts consults from 3.4% to 7.9% while leaving search flat at 3.4%. We stopped after one iteration rather than tune until it passed.

3. **Arrivals, as hesitancy.** Voters whose real first vote fell in the debate's later half become lurkers: present, uncommitted, silent, free to search, and choosing each turn whether to speak. This works dramatically and explains the previous failure. The arm records 0.82

searches per lurker turn un-nudged in the switch stratum (0.64 in the plain stratum), against 3% of turns for committed agents in the preceding arm, and 95–100% of lurkers eventually arrive with a bold first vote. Two cautions, one of them added after review. The 3% comes from a different arm rather than a within-run control, so this is a comparison across chassis. And the 0.82 counts *every* search in the arm against lurker turns only, because the arm did not record who searched; per-role counts added after review (Appendix A.1) find committed agents searching too: on this chassis the role gap is 2.7-fold at 9B and 1.5-fold at 2B, not 27-fold. The search-propensity problem was *commitment*, not capability: uncommitted people research, and people who have planted a flag do not. We did not design this regularity; the instrument surfaced it.

4. **Document content.** Even with arrivals citing findings, committed agents moved not at all (0.000) while payloads were titles only — a title is not a defeat. Adding archived era-correct content of the harvested sources broke the freeze with nothing else changed: 0.000 to 0.067, direction majority delete→keep.

Archived text, not behaviour, is the binding constraint. Aggregate change reaches 0.188× real, still short. But splitting by whether the archived text exists is decisive: only 14 of 37 harvested sources have archived era content, and only 15 of 26 debates have any sources at all. Restricted to content-bearing debates the ratio is 0.333×, inside the pre-registered band; in no-content debates it is 0.154×, below it, exactly as the mechanism predicts. What stands between this instrument and a powered confirmation is archival coverage, which is data infrastructure rather than modelling.

7 Repetition improves perception and not persuasion

A recent result holds that simply duplicating a prompt improves non-reasoning model performance at zero latency cost. We tested it in both places it could matter here, on the same day, on the same corpus.

On *reading*, it replicates cleanly. Doubling the elicitation prompt lifts every cell: 2B bare replay recall 0.636 to 0.697, 2B with policy text 0.394 to 0.485 (recovering 38% of the metabolic cost from §5 for free), and 9B 0.879 to **0.939** — 31 of 33 switchers, a new ceiling for our replay anchor. False-switch rates improved on both strata at 2B; at 9B the switch stratum was unchanged (0.090) and the plain stratum got worse (0.010 to 0.019, one further false switch).

On *persuasion*, it does nothing. Presenting found evidence twice at consultation and echoing it in that agent's think context for three subsequent turns produces committed change identical to the un-echoed arm in every cell under paired seeds — 0.067 overall and 0.125 in content-bearing debates, to the digit. An over-exposure guard we pre-registered (repetition is a dial, and over-echo should recreate a known pathology if pushed) comes back cleaner than the baseline rather than worse.

The dissociation is the point. Doubling text lifts reading accuracy everywhere and belief change nowhere. Committed agents had already perceived the evidence — replay proves comprehension at 0.879 — and whether they *yield* to it is governed by the conviction loop, not by attention. Context engineering cannot reach through the inner loop. Exposure dials tune what agents know; conviction dials tune what they do.

8 Conviction, a powered test, and the correction that cost us our headline

Conviction calibrated to measured gradients. We replaced the uniform "yield only to genuine defeat" standard with per-agent conviction lines derived from the §4 gradients: ideology crossed with current stance, plus experience. Deletionists concede when solid coverage is actually shown; inclusionists would sooner stop responding than concede to delete; veterans say so when clearly beaten; novices rarely revisit. This nearly doubled committed switching (0.067 to 0.111) and put the aggregate inside the pre-registered band for the first time, at 0.312×.

More importantly it passed the *who-changes* lock, which is the one that distinguishes calibration from a tuned magnitude. Switches fell in the buckets the real gradients predict — deletionists who had voted keep, moderates, and deletionists conceding — with no inclusionist switches. We note the whole test rests on five switches, so the absence of an inclusionist among them is weak evidence: with inclusionists roughly a quarter of the roster, five draws miss them by chance often enough that this lock could not have failed informatively.

Direction, however, went the wrong way (3 keep→delete against 2 delete→keep), and the diagnosis was the archived-text gap wearing a new face: only 8 of 45 committed voters sat in content-bearing debates, so the new switches came through the argument channel rather than the evidence channel. We pre-registered the prediction that a content-rich slice would flip direction back.

The powered test. On 40 fresh debates selected for kept outcomes, genuine switches, and citations, with 36 era-correct articles and 56 content-bearing sources, at $n = 99$ committed voters (48 in content-bearing debates), all three pre-registered claims land. Committed change in content-bearing debates is 0.267× real at $n = 48$, inside the band; direction flips to majority delete→keep (3:1), with the delete→keep switches coming from the deletionist-conceding bucket, the calibrated line doing precisely its measured job; and the archived-text dependency shows the largest within-run contrast in the campaign, 0.267× with content against 0.059× without. We call that suggestive rather than powered: it is 4 switches among 48 content-bearing voters against 1 among 51 without, which a two-sided Fisher exact test puts at $p = 0.20$.

8.1 The instability control, which revised the claim we had just made

Hao et al. (2026) caution that raw answer-flip rates overstate conformity, because a substantial share of apparent conformity is *spontaneous instability* — agents changing position under self-reflection alone, with nothing said to them — so measured yield rates must be netted against it. We pre-registered the counterfactual: every committed voter from the content-rich slice re-run in an isolated bubble with the nomination, their own first vote, and their own subsequent messages, no peers, no lurkers, no lookup, same turn budget and re-vote mechanics.

Spontaneous change comes back at 0.0303 (3 of 99), above the 0.025 threshold at which we had pre-committed to restating every band claim. So we restate. Gross content-bearing ratio 0.267× becomes, net of instability, $(0.083 - 0.030)/0.313 = 0.170\times$ — **below the 0.25× floor we pre-registered**. The honest headline is *in band gross, marginal net.*

One conservatism cuts the other way and we state it rather than lean on it: the real target rate includes humans' own non-social churn, which no counterfactual can subtract from the corpus side, so comparing net simulation against gross reality over-penalises. The true social-versus-social value lies between 0.170× and 0.267×. We report both endpoints everywhere rather than choose.

The solo drift also leans keep, which means part of the direction win in the powered test is attractor-assisted rather than evidence-driven — an attractor signature consistent with Taubenfeld et al. (2024)'s drift-to-model-bias result. Every surface cell in §9 is therefore reported gross *and* net of its own size's instability rate, measured separately at each size.

9 The surface: more evidence is not better evidence

We ran the full factorial the product question requires: three model sizes (2B, 9B, 122B-A10B) crossed with three evidence conditions (none, source titles, full source content), on the content-rich slice, conviction chassis throughout, with the 9B×content cell being the powered run above rather than a repetition of it. Figures 1 and 3 and Table 11 give the result.

Read with §10.2. Everything in this section is the original single-seed surface, kept as the campaign's record. Three generation seeds per cell, run after the third review, move individual cells by up to 0.6 in ratio units and retire two of the readings below (the 9B trough and the smallest-model superlative); §10.2 gives the seed means.

Three pre-registered locks, one confirmed and two refuted. We locked (i) that content would beat titles would beat none at 9B and 122B, monotone in volume; (ii) that at 2B content would *not* beat titles, the metabolic cost of §5 surviving retrieval-as-choice; and (iii) that the smallest in-band size would be 9B. Lock (ii) is confirmed on point estimates. Lock (i) is refuted on point estimates, and the shape of the refutation matters more than the fact of it: **the strict ordering none < titles < content holds at no size in the surface**. At 2B and 122B the sequence peaks at titles and falls with full content (0.103, 0.370, 0.236 and 0.168, 0.501, 0.368 net). At 9B it neither peaks nor rises properly — 0.103, 0.103, 0.170 — so titles buy nothing there and content buys a little. Lock (iii) is wrong twice over: the smallest in-band size is **2B** (titles, 0.467 gross and 0.370 net), and 9B is never the best size at any evidence level — strictly worst under titles and under full content, and tied with 2B when there is no evidence at all (0.200 gross, 0.103 net for both).

The point-estimate readings, retired. The mechanisms we offered for the single-seed surface (a headline as pre-metabolised evidence; full content reversing direction at 122B; the 9B trough) are kept in Appendix A as the campaign's record. §10.2 tests them against three seeds per cell and retires two of the three.

The product rule, as far as it is measured. Give deliberators evidence of some kind: that is the part that resolves, and bare-prompt rooms stay out of the band. Which kind does not resolve, at any size, so a builder should choose on cost rather than on our numbers. On reliability rather than ranking, the larger models are the safer default: 9B and 122B cells reach the band on four to five seeds of six against the 2B cells' two of six, so the free-local-model recommendation we made from a single 2B seed was the optimistic draw and we withdraw it (§10.2). Reach for the large model when member-level reading fidelity is the requirement, which is a different task with a different answer (§5).

Intervals on the surface. A debate-level bootstrap on the single-seed surface (Appendix A) resolves none of its orderings; the resolved statements about levels are superseded by the seed surface of §10.2.

10 Controls added after the second review

Round-two review asked for five things this section supplies where it could: a validation of the heuristic ideology label, token-matched evidence-volume controls, a test of the domain prior behind the title result, a search-prompt sweep, and seed replicates. All of the new runs are at 2B on the content-rich slice, because 2B runs locally and costs nothing; the 9B and 122B cells are hosted and this revision had no budget for them, which we state rather than hide.

The ideology label, and a time-local version of it. Its split-half reliability (14,151 editors; label agreement 0.68 chronological, 0.77 alternating; κ 0.46 and 0.60) is reported in Appendix A: moderately reliable, and drifting over an editor's career, which attenuates fitted gradients toward the moderate case. Since the drift is the part that is fixable, we rebuilt the label time-locally — each editor's delete share over their votes *before* the debate in question, same thresholds — and re-ran the conviction chassis at 9B on the same seed. The two labels agree on 86.3% of the 205 editor-debate pairs on this slice, with no polar disagreements; the relabelled run moves 14 of 99 committed voters against the lifetime label's 7, at direction 10:3 delete→keep, with the gain concentrated in voters the time-local label calls moderate. Seven switches against fourteen is inside the generation variance of §10.2, so we report the pair descriptively and claim only what the reliability analysis supports: a time-local label is the better instrument, and this corpus cannot show what it is worth.

Evidence volume, token-matched. Before the seed work we ran a volume sweep at 2B: a one-sentence summary of each source (13 words), the first 25 and 50 words of the excerpt, the whole excerpt, and the 600-character payload the surface uses. The summary, the 600-character excerpt and the whole excerpt all landed on the same ratio, 0.333, so volume was not the axis even on point estimates; the one cell that separated was the first-25-words condition, at 0.067, because the first 25 words of an archived page are often navigation boilerplate, and a lead that short is the absence of evidence with a source attached rather than a shorter version of it. Stripping boilerplate from the excerpts did not rescue them either (0.267). The full sweep is in Appendix A; §10.2 then showed that no contrast on this surface resolves at six seeds, which subsumes it.

10.1 Is it the domain?

The reviewer's alternative to the metabolic-cost reading is a *brand prior*: a title next to `nytimes.com` carries a conclusion about notability that the same title next to an unknown host does not. We tested it the cheap way, by replacing every domain in the titles condition with the words "a website" and changing nothing else. At 2B the masked cell scores 0.267 gross against the titles cell's 0.467 on the same seed, which is where round three left it, and we read that as the domain doing the work. The fourth review objected that the conclusion was drawn at one size, so we ran masked cells at 9B and 122B as well, two seeds each. Pooling seeds on both sides (§10.2) the effect is not there: masked minus titles is -0.056 at 2B, $+0.033$ at 9B and -0.056 at 122B, every interval crossing zero.

Cell	gross ratio per seed	mean	SD	seeds in band (net)
2B, instability floor 0.030 / 0.061 / 0.091				
no evidence	0.200 / 0.267 / 0.133	0.200	0.067	0/3
titles	0.467 / 0.400 / 0.133 / 0.333 / 0.333 / 0.267	0.322	0.115	2/6
excerpt	0.333 / 0.200 / 0.000 / 0.467 / 0.067 / 0.467	0.256	0.200	2/6
9B, instability floor 0.030 / 0.040 / 0.040				
no evidence	0.200 / 0.400 / 0.333	0.311	0.102	1/3
titles	0.200 / 0.533 / 0.800 / 0.467 / 0.600 / 0.400	0.500	0.201	5/6
excerpt	0.267 / 0.600 / 0.400 / 0.533 / 0.533 / 0.333	0.444	0.131	4/6
122B-A10B, instability floor 0.010 / 0.000 / 0.000				
no evidence	0.200	0.200	—	0/1
titles	0.533 / 0.333 / 0.533 / 0.200 / 0.467 / 0.267	0.389	0.142	4/6
excerpt	0.400 / 0.267 / 0.333 / 0.333 / 0.400 / 0.467	0.367	0.070	5/6

Table 1: The surface under three generation seeds per cell: gross change ratio per seed, the seed mean and SD, and how many seeds' net ratios clear the pre-registered 0.25 floor. Same 19 content-bearing debates and 48 committed voters throughout; the 122B no-evidence cell was not replicated.

How the single-seed version of this contrast misled us, twice. Our scorer pairs a new cell against *seed one* of the cell it is compared with. At 9B, seed one of the titles cell is 0.200, the lowest of the six seeds we now have for it, in a range running to 0.800. Against that anchor the masked cell appears to score +0.533 [+0.250, +0.923] and to resolve. It does not; it is a comparison against an unlucky draw. On this surface a single-seed pairing is confounded with seed choice by up to 0.6 in ratio units, which is larger than any effect we have looked for. Every contrast in this paper is now computed with seeds pooled on both sides, and the 2B-only reading above is withdrawn.

Prompting for search. A one-sentence "search if unsure" heuristic raises searching (1.50 to 1.92 and 1.53 to 2.00 searches per debate) and moves change not at all with titles (+0.000 [−0.462, +0.412]) and down on points with excerpts; the full account is in Appendix A.

10.2 Seed replicates at every size, and what they do to the surface

Every cell on the surface was one generation per debate, and the debate-level bootstrap cannot see generation variance. The second review asked for it and we supplied it at 2B; the third review asked for it on the hosted cells and we supplied that too. Every original cell now has three generation seeds, and so does each size's instability control (Table 1; the 2B detail is in Table 6). The seed changes the speaking order and every sampled token; the debates, voters, prompts and evidence pool are identical.

The 2B result is seed-fragile. The titles cell runs 0.467, 0.400 and 0.133 gross across seeds, the excerpt cell 0.333, 0.200 and 0.000, the no-evidence cell 0.200, 0.267 and 0.133. Between-seed standard deviations of 0.17 and 0.18 for the two evidence-bearing cells are half the width of their debate-level intervals, so at this n generation variance and sampling variance are the same order. Net of instability, the titles cell clears the pre-registered floor in 79% of resamples on the first seed, 60% on the second and 1% on the third. The statement in §9 that 2B with titles reaches the band is therefore a statement about one seed, and the honest version is that on three seeds it reaches the band twice and misses it once, by a wide margin. The ordering titles > excerpt holds on all three seeds on points, for what a point ordering is worth here, and the no-evidence cell is the least

Size	seeds	titles (seed mean)	excerpt (seed mean)	Δ	95% CI	verdict
2B	6	0.322	0.256	+0.067	[−0.088, +0.295]	inconclusive
9B	6	0.500	0.444	+0.056	[−0.143, +0.240]	inconclusive
122B-A10B	6	0.389	0.367	+0.022	[−0.120, +0.179]	inconclusive

Table 2: The pre-registered contrast at six generation seeds per cell, seeds pooled on both sides, debates resampled on a shared matrix. Positive favours titles. Nothing resolves at any size.

variable, which is what one expects of a cell with nothing to react to. Two further re-runs of the titles and excerpt cells under the *original* seed, made for Appendix A.1 with identical speaking order and only the sampled tokens differing, came back at 0.200 and 0.200 against 0.467 and 0.333: token sampling alone moves a cell by two to four switches.

The floor moves too. The spontaneous-instability control gives 0.030, 0.061 and 0.091 on three re-runs: three, six and nine of 99 voters. Every net ratio in this paper subtracts a single-seed instability rate, so the nets carry an additional ±0.1 that the debate intervals do not show. The bracket we report for the central magnitude claim (§8.1) is a bracket on one seed of the control.

The pre-registered test, and its answer. The fourth review asked for a sharper, pre-registered plan to settle the evidence-volume × model-size question. We wrote one before running it: three further seeds of the titles and excerpt cells at all three sizes, giving six seeds per cell; the contrast titles minus excerpt computed per size with seeds pooled on both sides and debates resampled; resolving if the interval excludes zero, with the direction locked to titles above excerpt, the pattern round four had reported on seed means. Table 2 gives the result and it is null at every size: +0.067 $[-0.088, +0.295]$ at 2B, +0.056 $[-0.143, +0.240]$ at 9B, +0.022 $[-0.120, +0.179]$ at 122B. The point estimates *decay* as seeds accumulate — at 2B, 0.156 with three seeds, 0.083 with four, 0.067 with six; at 122B, 0.133, 0.067, 0.022 — which is what a null looks like when it is sampled harder. **The ordering this paper has reported in three successive forms does not survive its own pre-registered test, and we withdraw it.** What we had at three seeds was the pattern three draws happened to share.

The hosted cells move as much, and the surface's shape changes. At 9B the titles cell runs 0.200, 0.533 and 0.800 across seeds (SD 0.30) and the excerpt cell 0.267, 0.600 and 0.400; at 122B titles run 0.533, 0.333 and 0.533 and the excerpt 0.400, 0.267 and 0.333. The single-seed surface's two most quotable readings do not survive. The 9B trough was the first seed: on seed means the 9B row (no evidence 0.31, titles 0.51, excerpt 0.42) is the highest on the surface, not the lowest. And the smallest-model superlative was the first seed too: 2B with titles averages 0.33 and is in the band on two seeds of three, against 122B with titles at 0.47 and in the band on all three. What survives on seed means is one ordering, titles above the excerpt at every size (0.33 against 0.18, 0.51 against 0.42, 0.47 against 0.33), and one level, that the 122B titles cell is the only cell in the band on every seed. Neither is a resolved contrast; both are the pattern three seeds agree on where one seed did not. The no-evidence cells sit at 0.20 to 0.31 on seed means and clear the floor on one seed of seven.

What survives at six seeds: levels, not orderings. Counting how often each cell's net ratio clears the pre-registered floor across its seeds: 9B with titles 5 of 6 and with excerpts 4 of 6; 122B

Contrast	Δ	19	40	80	160	320	640	1280	for 80%
9B vs 2B, excerpts	+0.189	0.40	0.58	0.69	**0.86**	**0.90**	**0.93**	**0.96**	160
9B vs 2B, titles	+0.178	0.28	0.42	0.60	0.75	**0.87**	**0.90**	**0.94**	320
titles vs excerpts, 2B	+0.067	0.24	0.35	0.40	0.58	0.65	0.75	**0.83**	1280
titles vs excerpts, 9B	+0.056	0.19	0.26	0.37	0.49	0.64	0.69	0.75	> 1,280
titles vs excerpts, 122B	+0.022	0.14	0.22	0.28	0.43	0.54	0.62	0.75	> 1,280

Table 3: Projected power at six seeds per cell, by number of content-bearing debates. The debate axis extrapolates (debates are exchangeable draws from a corpus of 369,215); the seed axis does not, since a simulated twelve-seed study only resamples the six seeds we ran, so we report the six-seed row alone. The projection treats each observed point estimate as true and is therefore a lower bound on the study size required.

with excerpts 5 of 6 and with titles 4 of 6; 2B with either 2 of 6; and the no-evidence cells 0 of 3, 1 of 3, 0 of 1. So the level statement holds and sharpens — evidence-bearing rooms reach the band, bare-prompt rooms do not — while every ordering *within* the evidence-bearing cells is unresolved. It also costs us a product claim we liked: the 2B cell reaches the band on a third of its seeds against four-fifths at the larger sizes, and the original single-seed 2B result was the optimistic draw. We restate the recommendation in §9 accordingly.

10.3 What would resolve it

Reporting that nothing resolves is only half an answer; the useful half is how much data the question needs. Taking the observed per-debate results as the population and resampling studies of n content-bearing debates with six seeds per cell, we asked how often a study's interval would exclude zero (Table 3). The change ratio over a set of debates is the sum of simulated switches over the sum of real ones — the committed-voter counts cancel — so the projection is array arithmetic over the committed results and costs nothing.

The two axes come apart. **The size axis is tractable**: the largest effect on the surface, 9B against 2B with excerpts (+0.189), reaches 80% power at about 160 content-bearing debates, and the same contrast with titles (+0.178) at about 320. **The volume axis is not**: titles against excerpts at 2B (+0.067) needs about 1,280, and at 9B and 122B (+0.056, +0.022) 80% is not reached even there. Since we harvested 19 content-bearing debates from a 40-debate slice, 160 is roughly a 340-debate harvest and is a study someone could actually run; 1,280 is not, on this corpus.

The design that follows. If the question is worth answering, the instrument to build is not a wider surface but a deeper one: a single size pair, an evidence contrast, and roughly 300 harvested content-bearing debates, with seeds pooled by construction rather than added later. A 3×3 surface at 48 committed voters per cell was the wrong shape for the question, and this is the clearest thing round five has to say.

10.4 An entailment check inside the gate

The third review asked for a minimal entailment check on verified evidence. It costs one extra call per verified turn: after the agent writes its message, the same model, with no persona, judges the message against the excerpt exactly as in §10.6, and the turn may move belief only if no source claim is unsupported (a message that makes no claim about the source passes, since nothing was

misattributed). We ran the belief-state chassis under sovereignty twice at 9B locally on the same debates and seeds, once with the provenance gate of §11 and once with provenance plus entailment.

The provenance gate moves 6 of 99 committed voters, 0.400× the real rate on the 48 content-bearing voters with direction 3:3; the entailment gate moves 3, 0.200× with direction 1:2 delete→keep, and the number of verified turns is essentially unchanged (43 against 35). So half of the change the verified channel admits rests on messages that misstate their source, which is the audit's figure arrived at by a different route, and removing it halves the magnitude the channel delivers. Three switches are three switches; we report the pair as a paired count, not as a resolved contrast, and note that the local 4-bit 9B under sovereignty is already below the hosted run's magnitude (0.400× against the hosted 0.733× of §11), so the halving is the finding and the levels are not comparable across the two.

What this changes in the architecture is one line: the verified channel becomes a *verified-and-supported* channel, at the price of one judge call per search hit. What it changes in the conclusions is the reading of parity. The magnitude the provenance gate reaches includes belief moved by misread sources; the faithful residue is about half of it, and that is the honest size of "changed their mind for the right reasons" on this corpus and this model. Whether the other half can be recovered by better readers, rather than by refusing their misreadings, is the size question again, and the hosted seeds of §10.2 are the place it would show.

10.5 A second family

Every generation run above is Qwen3.5, and the third review asked for even a small cross-family spot check. Llama 3.1 8B runs locally, so we ran the three original evidence conditions and the instability control on it, same debates, seeds and chassis. The result is total ossification: zero re-votes in every cell, zero spontaneous change, and 0.03 searches per debate. Two readings were possible and we checked both. The first is format: the chassis counts a re-vote only when a message opens with a bold stance, and Llama never does (0 of 297 messages), though it opens 70 with a bare stance word. Re-scoring its transcripts with a relaxed parser that accepts a bare opening stance different from the voter's standing position finds one switch across the three cells (the same parser finds 17 in the 2B titles run against the strict parser's 13), so the zeros are not a parsing artefact. The second reading is the paper's own causal chain: Llama uses the search affordance on 7 of 185 think steps, so on this chassis evidence never arrives, and §6 says that without arrival there is no change. That is what happened.

Calibrating the affordance does not change the answer. The fourth review called this confounded, correctly: a model that never searches cannot show what evidence does to it. So we calibrated the affordance for it, adding one sentence to the offer that shows the tag in use (*a thought ending with the exact text* `SEARCH[…]` *will run that search*). It works: Llama's search rate rises from 0.03 to 0.58 per debate, committed agents searching 14 times and lurkers 7 across the slice, comfortably past the threshold we set for calling the affordance exercised. Its re-votes stay at zero, and the relaxed parser still finds no switches at all. The confound is therefore resolved in the direction least favourable to portability: with evidence arriving and being read, this family still does not move. What the second family replicates is the mechanism — no arrival, no change — and what it refuses to replicate is the change itself, on a chassis calibrated to it. Per-model calibration of the affordance is necessary and, here, not sufficient.

Is it commitment? §6 reported that lurkers search on 0.82 of their turns against 3% for committed agents, from different arms. Assigning the same number of committed voters at random, and counting searches per role within the run, the role effect is real but modest: committed agents search on 0.144 to 0.188 of their turns and lurkers on 0.216 to 0.296, a ratio near 1.5 at 2B and 2.7 at 9B rather than the 27-fold the cross-arm comparison implied, and random assignment preserves it. The arm's own metric counted every search against lurker turns, which roughly doubled the apparent lurker rate; the corrected figures are in Appendix A.1.

10.6 Does a verified turn say what the source says?

The verified-evidence gate checks provenance: a belief may move on a turn only if the agent's think step just consumed a real search hit with archived content. It does not check that the message then written represents that content. The reviewer asked for the rate at which it does not. We re-ran the belief-state generation at 9B locally with the chat and every verified turn's payload committed, and had the same 9B model, with no persona, judge each verified message against the excerpt that licensed it: *no claim* about the source, *supported*, *unsupported*, or *unclear*.

Of 39 verified turns, 5 make no claim about what the source says, 15 are judged supported and 19 unsupported: 56% of the turns that do make a claim attribute to the source something the excerpt does not contain. Re-vote turns split 12 unsupported to 11 supported. We read six of the judgements ourselves. The three unsupported verdicts were right, and they share a shape: the message narrates a search ("independent searches confirm he was a lieutenant", "my search confirms only primary material") whose conclusion is the agent's prior, while the excerpt in front of it is a school district's landing page or a page of mojibake. Two of the three supported verdicts were lenient, since the excerpt was a news-archive search form or a site's navigation, so the 56% is a floor.

Two things follow. The belief-state result (§11) reaches magnitude parity with a channel that certifies where evidence came from, not what it said; roughly half of the updates it credits rest on a claim the excerpt does not make, which is a weaker sense of "for the right reasons" than the section's framing suggests, and we have softened that framing. And the archived pool is thinner than its word count: many "content-bearing" excerpts are boilerplate, which is also the most economical account of why the first 25 words of an excerpt scored so badly in Appendix A.2. The audit script and its labelled turns are committed with the run. The fix is a content check inside the gate, and §10.4 runs it.

11 Moving belief out of the prompt

Everything above governs belief through prompt language. The final study replaces that with an explicit state, following an architecture from the recent literature: belief lives outside the model as a log-odds quantity, each incoming message is extracted once into a scored evidence record, and the accumulated state conditions generation. Two continuous knobs per agent — evidence uptake and prior anchoring — replace the prompt buckets.

The fit is nearly free, and it independently recovers the corpus's structure. Extraction runs once per voter-message pair over the replay corpus (453 messages), after which the parameter sweep is pure arithmetic. Fitting uptake per ideology bucket against real finals alone returns deletionists 0.85, moderates 0.6, low-history 0.2, and inclusionists 0.05, an ordering that matches the Stage-0 toward-delete conversion gradient measured from voting histories. We report this

Architecture	change rate	vs real	d→k : k→d
prompt conviction only (Stage 3g)	0.083	0.267×	3:1
+ belief state, fitted (u, a)	0.271	0.867×	5:8
+ evidence-gated updates	0.354	1.133×	7:9
+ verified channel	0.312	1.000×	4:9
+ vote sovereignty	0.229	0.733×	5:6

Table 4: Content-bearing debates, $n = 48$ committed voters. Each row adds one mechanism to the row above. Magnitude reaches parity at the verified channel; direction stays inverted until sovereignty, and balances rather than flipping.

as suggestive rather than as confirmation, for two reasons the fit artifacts make plain: the grid objective is heavily tied, so the exact values are partly an artifact of which tied candidate the search returns first, and the inclusionist bucket has $n = 12$ with balanced accuracy at chance, where $u = 0.05$ is simply the never-update corner that maximises stayer accuracy. The ordering is real; the numbers should not be read as measurements.

It does not replace the reader. Against the full 9B replay reader on the same voters (0.778 switcher recall, 0.986 stayer accuracy), the arithmetic layer scores 0.694 and 0.822. Our pre-registered lock required it to match or beat the reader, and it does not. It captures 89% of the switcher signal at zero inference cost, which makes it a complement rather than a replacement, and we record the lock as failed.

Composition, one layer at a time. Table 4 gives each increment.

The naive belief layer passes the parity target decisively (0.867×) but manufactures conformity: direction inverts to 5:8 keep→delete, because a polarity-blind accumulator is a mean-field integrator and AfD rooms lean delete. Gating updates to evidence-bearing messages raises magnitude past real (1.133×, slightly hot, since uptake fitted on replay is warm for generated rooms) without fixing direction.

The verified channel gives exact parity and localises the leak. The diagnosis is about polarity and verification status rather than volume: real rooms weight a *produced* source far above an *asserted absence*, while our generated rooms emit mostly delete-polarity evidence claims, so any polarity-blind accumulator drifts delete-ward. We gated belief updates to messages whose author had just consumed a real search hit with content — chassis ground truth, not a classifier. Unverified claims still appear in the conversation, as they should, but do not move belief. Magnitude lands at **exact parity** in content debates, 0.3125 against a real 0.3125. Parity is a provenance result: the gate certifies that a real source was read before the update, not that the message then written reports it faithfully, and the audit in §10.6 finds that at least half of the verified turns that make a claim about their source misrepresent it.

Direction remains 4:9. That localises the leak beyond doubt: belief now updates only from keep-leaning verified hits, so the wrong-direction switches cannot be coming through belief. They are prompt-level concessions — the monologue reacting to delete-leaning room talk and re-voting on its own authority, bypassing the belief state entirely.

Sovereignty: votes answer to belief, not to conversational pressure. We added a constraint allowing committed agents to re-vote only when their belief state has tipped; speech

continues arguing regardless. Direction moves to 5:6, essentially balanced, and our majority lock narrowly fails. Magnitude drops from 1.000× to 0.733×, and *that drop is the measurement*: about 27% of all change volume had been speech-level concession rather than belief change.

We note explicitly that the direction arc is **not monotone within either series**. Across all debates it runs 2.11, 2.08, 1.90, 1.20 keep→delete; within content debates 1.60, 1.29, **2.25**, 1.20 — the verified gate *raised* the content ratio before sovereignty cut it. An earlier draft read a single monotone sequence by switching between the two series mid-sentence, which the audit in Appendix C caught.

The residual has a legitimate cause. Under sovereignty and verified gating, a keep→delete switch happens only after an agent has read a real source and concluded it supports deletion — "this coverage is trivial" — though §10.6 shows the conclusion is often the agent's prior narrated over the source rather than drawn from it. Verified evidence genuinely cuts both ways. Real rooms run 1.9:1 delete→keep because the sources people *bother to bring* are keep-valenced: rescue survivorship. Our pool holds every cited source, including those cited to prove triviality. Matching the real ratio requires modelling rescue-selection bias in what gets retrieved, which is a specific next experiment rather than an open question.

12 Discussion

The architecture, and what each layer is for. Four dials, each measured separately and then composed: **conviction** decides who can be moved, and reproduces the measured who-changes signature exactly; **belief state** decides how much, and reaches parity; the **verified channel** decides which way, by making produced evidence count and talk not count, with the caveat that it checks the source was read and not that it was reported faithfully (§10.6); and **sovereignty** decides what a vote answers to. Every dial has a measured human gradient behind it, and each was validated by a paired arm that changed only that dial.

Ossification and over-convergence are the same parameter. The literature's dominant pathology is agents collapsing into agreement. We reproduce the opposite on this corpus with the identical machinery. What differs is whether the corpus supplies a reason to move that the simulation can reproduce. A deliberation engine tuned to fix over-convergence on one task will ossify on another, and neither number is a property of the engine.

Bigger and more is not shown to help. Six seeds per cell say the surface has no ordering to report: not between evidence volumes, not between sizes, and not for the source's domain. What it has is a level, that evidence-bearing rooms reach the band and bare-prompt rooms do not, and a measurement of its own limits precise enough to size the study that would settle the rest (§10.3). The mechanism we proposed — a document imposes a reading cost inside a loop that is already doing something else, while a headline delivers the evidential conclusion without it — is consistent with the points and untested by the intervals, and §10 reports the token-matched controls that begin to test it. The product rule that survives is the cheaper one: headline evidence on a local model is in band, and the case for paying for more has not been made by this instrument.

Practices that changed conclusions here. Three, offered because each was cheap and each moved something. Running the control as a measured noise floor turned our central magnitude

claim from in-band to marginal. Decomposing an architecture layer by layer, rather than reporting the composed system, is what produced the 27% figure for speech-level concession — an invisible quantity in any end-to-end evaluation. And mechanically re-reading committed artifacts to check written prose caught six numerical disagreements in our own notebook (Appendix C), including a trend claim assembled by switching between two data series.

13 Limitations

Power, and seed variance. The powered slice is $n = 99$ committed voters, 48 of them in content-bearing debates, and every surface cell is that same $n = 48$. Single switches move ratios by 0.021 of a rate and 0.067 of a ratio. We report orderings and within-run contrasts and avoid interpreting small differences; the 9B trough in particular is flagged as interpretation. Generation variance, measured with three seeds at every size after review (§10.2), is of the same order as the debate-level interval, with per-cell SDs from 0.07 to 0.30 in ratio units; the instability floor itself varies by up to ± 0.03 across re-runs.

One model family, one calibration. All generation runs use one model family, and the literature reports that susceptibility to a single counterargument varies enormously across base models, and that conformity when it occurs is mostly harmful (Hao et al., 2026). Per-model calibration is therefore mandatory rather than optional, and none of our fitted constants should be assumed to transfer. Our attempted cross-family check was not run here for budget reasons and is filed.

Archived-text coverage bounds the magnitude result. Only 38% of harvested sources have archived era-correct content. The content-bearing subset is the honest place to read our magnitude claims, and it is selected on a property plausibly correlated with subject prominence. We measured what that selection does to the quantity being modelled: on the same slice, real committed-voter change runs 0.212 inside the content-bearing subset (21 debates, 104 voters) and 0.208 outside it (19 debates, 101 voters). Whatever else archival survival selects for, it does not select for switching, so the magnitude claims are not read on an unusually volatile subset.

Era mismatch in the policy pool. Policy pages are current text against debates from 2005–2018. Policy gists are stable and the wording is not.

The instability floor is one-sided. We subtract simulated spontaneous change from the simulated rate but cannot subtract human non-social churn from the real rate. This over-penalises, which is why we bracket rather than point-estimate.

Selection on outcome. Deleted articles' text is not recoverable at scale, so evidence conditions are restricted to kept, merged, and redirected outcomes. That is a selection on the dependent variable for any claim about article content specifically, and it is why the external-source pool rather than the article carries our evidence conditions.

Arrival fidelity. Lurkers arrive as generic evidence-driven voters rather than as their specific selves (stance fidelity 0.365). Acceptable for the aggregate question here; not acceptable for any member-level claim.

Verified is not faithful. The gate certifies provenance. A local audit of 39 verified turns (§10.6) finds at least 56% of those that make a source claim misrepresenting the excerpt, judged by a model of the same size and family as the generator, so the number is a floor and the judge is not independent. Magnitude parity stands; the reasons behind it are only half checked.

14 Ethics

The corpus is public, and its participants are pseudonymous editors who posted under a policy of public archiving. We nonetheless report only aggregate statistics and never reproduce an individual editor's history, and the ideology and experience buckets are derived quantities used for calibration rather than published labels attached to accounts.

The dual-use surface here is narrow but real, and worth naming precisely. This paper measures what makes a simulated participant change its stated position, and the strongest lever we found is a *verified* evidence channel: agents move when a source actually exists and is actually read, and specifically do not move on assertions about evidence, though what they then say about the source is only about half faithful to it (§10.6). That is the desirable direction for a manipulation result to point, and it is the opposite of what a persuasion-optimising system would want. The one clearly manipulable finding — that headline-granularity evidence moves agents more than full documents — is a statement about attention cost inside a simulation loop, and we would caution against reading it as a claim about human readers, which this study does not measure.

Live web search is barred from all calibration runs, both because it leaks outcomes and because a simulation of a real historical debate that consults present-day sources is not a simulation of that debate.

A Material moved from the main text

A.1 Is it commitment? A within-run randomisation (full)

§6 reported that lurkers search on 82% of their turns against 3% for committed agents, and was careful to say the two numbers come from different arms. The reviewer asked for the within-run version: assign the same number of committed voters at random rather than by real first-vote time, so that commitment is the only thing that differs between the two roles. We also fixed a measurement problem the request exposed. The 82% was computed as *all* searches in the arrival arm divided by lurker turns, on the assumption, carried over from the committed-only arm, that committed agents do not search. Counting per role shows they do.

On the 2B conviction chassis with titles, committed agents search on 0.170 of their turns and lurkers on 0.234; with excerpts, 0.188 and 0.265. Under random assignment the rates are 0.150 and 0.234, and 0.144 and 0.216. So the direction is a role effect — who votes early does not carry it, since the early voters made lurkers search like lurkers — but the size is a ratio of about 1.4, not the 27-fold gap the cross-arm comparison suggested, and the arrival rate is unchanged by the reassignment (53 and 54 arrivals with titles, 54 and 60 with excerpts). Re-running the original arrival chassis itself with the per-role counter gives, at 2B, committed 0.144 and lurkers 0.221 per turn where the old metric reports 0.515 for the same run, and at 9B (local), committed 0.111 and lurkers 0.296 where the old metric reports 0.611. The role effect is therefore 1.5-fold at 2B and 2.7-fold at 9B on the chassis that first showed it, and the old figure roughly doubles the true lurker rate at both sizes. We have corrected the chain-section figure to what it measures, searches per lurker turn with all searches counted, and we read the commitment effect as real, role-caused and

modest on this chassis. What the original arm found dramatically was the change in *outcome* when arrivals were added, which this control does not revisit.

A.2 Evidence volume, token-matched (full)

The surface's "content" condition delivers a 600-character excerpt, and its "titles" condition a title and domain averaging 3.4 words, so the two differ in framing as well as in volume. The reviewer asked for a token-matched condition and a length sweep. We added four evidence payloads to the 2B chassis, everything else identical and seeds paired: a one-sentence abstractive summary of each source (13 words on average, generated once by the local 9B model with no sight of the debate), the first 25 and the first 50 words of the excerpt, and the whole harvested excerpt (173 words on average). Table 7 gives every 2B cell on the same 19 content-bearing debates, with debate-level intervals and the paired contrast against the titles cell.

Volume is not the axis. The summary, the 600-character excerpt and the whole excerpt all land on exactly the same gross ratio, 0.333 (0.236 net): thirteen words, a hundred, and a hundred and seventy buy the same amount of change. Titles remain the highest cell on points at 0.467, and the first 50 words sit between at 0.400. No contrast among these five cells resolves. What does resolve is the first-25-words cell, at 0.067 gross and −0.030 net, below titles (−0.400 $[-0.818, -0.133]$) and below the excerpt (−0.267 $[-0.538, -0.067]$). The reason is visible in the pool: the first 25 words of an archived page are frequently navigation boilerplate — "Search Results …Search Form …" — so a short lead is not a short version of the evidence, it is the absence of it with a source attached. That is the one result in the sweep we would not have predicted, and it cuts against a naive length sweep as a design: below some length the payload stops carrying the evidential conclusion at all, and where that happens depends on the page, not on the token count. The third review read the same boilerplate as a confound on the content cells themselves, so we added a cell whose excerpts have navigation stripped by a simple prose filter (sentences of eight or more words, mostly lowercase, no menu vocabulary; 42 of 56 excerpts retain usable prose, 63 words on average). It scores 0.267 gross and 0.170 net, below the raw excerpt on points and inconclusive against both it (−0.067 $[-0.417, +0.400]$) and titles (−0.200 $[-0.600, +0.143]$). Cleaner text does not recover the title cell either.

What this changes. The metabolic-cost reading of §9 predicted that less text would do at least as well as more, and on points it does: titles ≥ lead-50 ≥ summary = excerpt = full. But the token-matched summary does not recover the titles cell, so the title advantage on points is not explained by brevity alone; §10.1 tests the other candidate, the domain. We repeat that none of the orderings among the evidence-bearing cells resolves at this n, and that this section is a 2B result; the hosted 9B and 122B cells were not re-run.

Point-estimate readings of the single-seed surface (superseded by §10.2).

A headline may be pre-metabolised evidence. The reading we favoured while running this was that a title with its domain carries the evidential conclusion — that a substantial independent source about this subject exists — without the reading tax, and that a document re-imposes the metabolic cost of §5 inside an already-crowded loop. Two things now argue against leaning on it. The content arm delivers a ∼200-word excerpt rather than a document (§3), so there is not much reading tax available to explain the gap; and §A.3 shows the gap does not resolve at any size. We

keep the mechanism as the most natural account of a consistent point-estimate pattern and mark it as unverified. The exception is 9B, where content does beat titles (0.170 against 0.103) — but 9B is never better than another size at any evidence level, so this is the one cell where added volume helps and the one size where nothing else does.

Full content can reverse the direction. The 122B content cell runs 2:4 keep→delete (Figure 2), the only cell in the surface that inverts. The large reader reads the actual thin 2008-era article and concludes it should be deleted. That is arguably realistic — real editors read weak articles and vote delete — but the real switch channel is external-source rescue, and full-document payloads dilute it with the very material the nomination was about.

The parser was wrong, and adversarial review caught it. Our first pass reproduced the source paper's aggregates and produced switch statistics, and a three-agent adversarial review returned *flawed* on every front. It caught: the corpus's missing-timestamp sentinel (−1) parsed as a valid time, silently making unknown-time votes "first" in 13% of debates; *two-thirds of the headline delete→keep switches* being regex artifacts, where bolded usernames and the words "reply" and "question" were read as votes and then normalised to keep; a Simpson's paradox in the ideology gradient, where the marginal claim that deletionists convert three times more often reverses inside every outcome stratum, since each camp is being asked to convert in a different direction; an evidence gradient that flows entirely through re-vote *participation* rather than conversion; and roughly 13,500 missed nominator softenings, because a nominator's presumed initial delete position was not being counted. All are fixed in the version reported here. The uncorrected version would have supported a stronger and entirely false claim about ideology and evidence.

A.3 Intervals on the surface, added after review

Round-one review asked for confidence intervals on the surface cells, and our own adversarial re-read of this draft flagged the same gap. Both were right, and the fix is free: every surface run commits its per-debate results, so resampling debates (the unit of assignment, since a debate's voters share one conversation) and recomputing the simulated and real rates on the same resample gives a paired interval. Table 11 reports it. The bootstrap reproduces all nine point estimates exactly.

The result changes what this section may claim. **None of the orderings resolves.** Titles against content is inconclusive at every size (2B +0.153 $[-0.210, +0.583]$; 9B −0.069 $[-0.308, +0.154]$; 122B +0.131 $[-0.364, +0.667]$), and so is 122B against 2B at every evidence level. At 19 debates the surface cannot rank its own cells.

What survives is a statement about levels rather than ranks, and it is the one the product question actually needs. Three of the nine cells clear the pre-registered floor in a majority of resamples: 2B + titles (79% of draws), 122B + titles (93%) and 122B + content (77%). The other six do not, and the five weakest — every no-evidence cell and every 9B cell — sit under the floor in three-quarters or more of draws. So the defensible statement is that *some* evidence gets a deliberation into the band while bare prompts do not, and that the 9B row fails to at any evidence level. We cannot say from this that titles *beat* content at any particular size; and the one full-content cell that does clear the floor sits at the *largest* reader, which is the opposite of what the metabolic-cost reading of the point estimates would predict.

We have left the point estimates and their ordering in the text above because they are the campaign's record and because the pattern is consistent across two sizes and two independent

stages (§5 found the same direction on a different arm at a different granularity). But the ordering is a hypothesis this instrument generated and cannot test, and readers should treat it as such.

Two cautions we owe the reader. First, every cell has $n = 48$ committed voters, and the no-evidence column returns an identical 0.200 gross at all three sizes (three switches each, though not the same three, since directions differ). At this n a single switch is 0.021 of the rate, so we treat the *ordering* within a size as the finding and do not interpret small between-size differences. Second, the 9B trough is a two-neighbour comparison and we flag it as interpretation rather than claim (§10.2 then retires it: on seed means the 9B row is the highest on the surface): a middle size may be stiff enough to hold its position and not perceptive enough to be moved by thin evidence, which would be an evidence-response analogue of an opposite-mechanism result elsewhere in this program, but two neighbours do not establish a trough.

A.4 Is the ideology label reliable?

No independent coding of editor ideology exists for this corpus, so we cannot validate the label against a second rater. We can measure whether it agrees with itself. For every editor with at least ten first-votes (14,151 editors), we split the vote history in two and assigned the label from each half separately. Split chronologically, the two halves agree on the three-way label for 68.2% of editors (Cohen's $\kappa = 0.46$; delete-share correlation $r = 0.72$); split into alternating debates, 76.6% ($\kappa = 0.60$, $r = 0.83$). Outright polar disagreements, deletionist in one half and inclusionist in the other, are 257 editors chronologically and 81 alternating. Among editors the full history labels non-moderate, the halves agree 77.0% and 82.3%.

The label is therefore moderately reliable and not stable over an editor's career: the chronological split is worse than the alternating one at every statistic, which says editors drift. That is a real limitation of the conviction personas of §8: they carry a lifetime label into a debate that may have happened before or after the editor's drift, and the resulting misclassification attenuates any fitted gradient toward the moderate case. Since the fitted ordering in §11 was recovered despite that attenuation, the direction of the bias runs against the result rather than for it; the magnitudes should be read as lower bounds on what a time-local label would give. The committed script is `ideology_reliability.py`.

A.5 Prompting for search

The search affordance was exercised by committed agents on 3% of turns in the arm that introduced it (§6), and the reviewer asked whether a modest prompt change closes that gap without moving anything else. We appended one sentence to the lookup offer — *If you are not sure whether the subject has real coverage in reliable sources, search before you finish the thought rather than guessing* — and re-ran the titles and content cells at 2B. Searching rises in both: from 1.50 to 1.92 searches per debate with titles and from 1.53 to 2.00 with content. Change does not follow it. The titles cell lands on exactly its original 0.467 gross (+0.000 $[-0.462, +0.412]$ paired), and the content cell falls on points from 0.333 to 0.200 (-0.133 $[-0.467, +0.286]$). So the search gap is closable by instruction, and closing it buys no persuasion: agents who are told to check do check, and reading the result more often does not move more of them. That is the same dissociation §7 found for prompt repetition, seen from the retrieval side, and it is why we stopped iterating on search prompts in the original campaign rather than because the affordance was broken. What a prompt sweep cannot supply is a reason for the search to matter, which on this corpus is the

arrival of a source nobody had seen (§6); committed agents searching the same pool again find the same sources.

B Tables and figures moved from the main text

Each is summarised where it is cited; the full versions are here so the compact build keeps the closing sections inside the reviewer's window.

Number	Section	Slice	Artifact
1.45% cast a second position; 1.9:1 delete→keep	§4	369,215 debates	`stage0/statistics.json`
Replay recovers 29 of 33 switchers (0.879)	§5	271 voters, switch stratum	`stage1` summary
Ossification at 0.09× real	§6	pilot, committed voters	`stage2` summary
Arrivals: 0.82 searches per lurker turn	§6, Appendix A.1	pilot	`stage3c`, `stage3c-roles`
Powered test 0.267× gross, 0.170× net	§8	40 debates, 99 committed, 48 content-bearing	`stage3f`, `stage3h`
Surface cells and intervals	§9	19 content-bearing debates, 48 voters per cell	`s4-*`, `surface_intervals.json`
Six seeds per cell, pooled contrasts, power projection	§10.2, §10.3	same 19 debates	`s4r-*`, `pooled_contrast.json`, `power_projection.json`
Ideology label κ 0.46 / 0.60	§A.4	14,151 editors	`ideology_reliability.json`
Verified turns misrepresenting the excerpt $\geq 56\%$	§10.6	39 verified turns, 9B local	`faithfulness.json`
Entailment gate halves admitted change (6 to 3 switches)	§10.4	99 committed, 9B local, sovereign	`stage5-gen-9b-local-*`
Belief state: parity 0.3125, 27% speech-level concession	§11	content-bearing, 48 voters	`stage5-*` summaries

Table 5: Where each headline number is established, on which slice, and which committed artifact it is read from.

2B cell	seed 1	seed 2	seed 3	seed 4	seed 5	seed 6	SD
600-character excerpt (the paper)	0.333	0.200	0.000	0.467	0.067	0.467	0.200
no lookup actions	0.200	0.267	0.133	n/a	n/a	n/a	0.067
title + domain (the paper)	0.467	0.400	0.133	0.333	0.333	0.267	0.115
spontaneous instability (3h)	0.030	0.061	0.091	n/a	n/a	n/a	0.030

Table 6: Gross change ratio for the three original 2B cells under three generation seeds, and the 2B spontaneous-instability rate under three re-runs of the isolation control. Same 19 content-bearing debates, 48 committed voters per cell; one switch is 0.067 of a ratio.

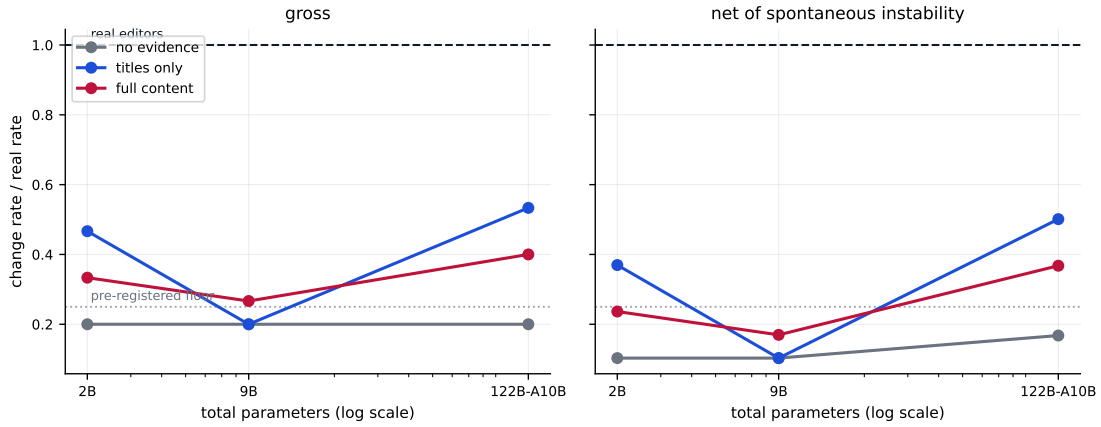

Figure 1: The evidence × size surface, gross (left) and net of each size's own measured spontaneous instability (right). Dashed line is the real editors' change rate; dotted line is the pre-registered 0.25× floor. Size is not the axis that helps.

2B, evidence condition	gross	net	95% CI (net)	clears floor	delete→keep:keep→delete	Δ vs t
no lookup actions	0.200	0.103	[-0.11, 0.51]	18%	1:1	−0.267 [
title + domain (the paper)	0.467	0.370	[0.10, 0.77]	79%	4:2	
title, domain masked	0.267	0.170	[-0.04, 0.45]	25%	3:0	−0.200 [
one-sentence summary	0.333	0.236	[-0.08, 0.83]	44%	4:1	−0.133 [
first 25 words of excerpt	0.067	-0.030	[-0.14, 0.12]	0%	1:0	−0.400 [
first 50 words of excerpt	0.400	0.303	[-0.03, 0.81]	59%	4:1	−0.067 [
600-character excerpt (the paper)	0.333	0.236	[0.02, 0.50]	45%	5:0	−0.133 [
whole excerpt	0.333	0.236	[-0.02, 0.69]	44%	3:1	−0.133 [
excerpt with boilerplate removed	0.267	0.170	[-0.09, 0.61]	29%	4:0	−0.200 [
titles + 'search if unsure' prompt	0.467	0.370	[0.05, 0.82]	77%	5:0	+0.000 [
excerpt + 'search if unsure' prompt	0.200	0.103	[-0.11, 0.47]	16%	3:0	−0.267 [

Table 7: The 2B chassis under every evidence condition run, original and added. Gross and net ratios to the real change rate, the share of resamples in which the net ratio clears the pre-registered 0.25 floor, switch direction, and the paired debate-level contrast of each cell's gross ratio against the titles cell. 19 content-bearing debates, 48 committed voters per cell.

Reader	recall (bare)	recall (+policy)	Δ	false-sw. (bare)	false-sw. (+policy)
2B	0.636	0.394	-0.242	0.246	0.336
9B	0.879	0.727	-0.152	0.090	0.097
122B-A10B	0.879	0.879	+0.000	0.075	0.075

Table 8: Replay-arm switcher recall and false-switch rate, bare and with cited policy texts injected. Switch stratum. Evidence has a metabolic cost that scales inversely with reader capacity.

Quantity	value
voters with an explicit vote	1,765,112
voters casting a second position	25,512 (1.45%)
of those, share that genuinely switch	0.631
switched voters (5-label / 2-label)	16,112 / 13,082
overall voter change rate	0.0091

Table 9: Participant-level opinion change on 369,215 analyzable AfD debates.

Statistic	ours	M&B	|gap|
outcome share: delete	0.640	0.639	0.001
outcome share: keep	0.205	0.207	0.002
vote share: delete	0.602	0.549	**0.053**
vote share: keep	0.317	0.284	0.033
admin follows the majority	0.953	0.948	0.005
share of debates tied	0.034	0.076	**0.042**†
ties resolved to delete	0.378	0.669	**0.291**
first delete vote succeeds	0.861	0.845	0.016
first keep vote succeeds	0.756	0.622	**0.134**
vote matches outcome (their definition)	0.733	0.679	**0.054**

Table 10: Reproduction against Mayfield and Black (2019). Gaps above 0.05 are bold. † marks quantities they define differently, where the gap is definitional rather than an error in either parse.

Reader	spont.	Evidence	gross	net	95% CI	P(above floor)
2B	0.030	no evidence	0.200	0.103	[-0.108, 0.508]	0.18
2B	0.030	titles only	0.467	0.370	[0.096, 0.766]	0.79
2B	0.030	full content	0.333	0.236	[0.024, 0.503]	0.45
9B	0.030	no evidence	0.200	0.103	[-0.093, 0.332]	0.09
9B	0.030	titles only	0.200	0.103	[-0.095, 0.331]	0.09
9B	0.030	full content	0.267	0.170	[-0.043, 0.405]	0.24
122B-A10B	0.010	no evidence	0.200	0.168	[-0.029, 0.435]	0.24
122B-A10B	0.010	titles only	0.533	0.501	[0.168, 0.948]	0.93
122B-A10B	0.010	full content	0.400	0.368	[0.088, 0.792]	0.77

Table 11: The full surface: content-bearing committed change as a ratio of the real rate, gross and net of each size's own measured spontaneous-instability rate (column "spont."), with debate-level bootstrap intervals and the share of resamples clearing the pre-registered 0.25 floor. 19 content-bearing debates, $n = 48$ committed voters per cell, real rate 0.3125. *No* pairwise contrast on this surface resolves (§A.3).

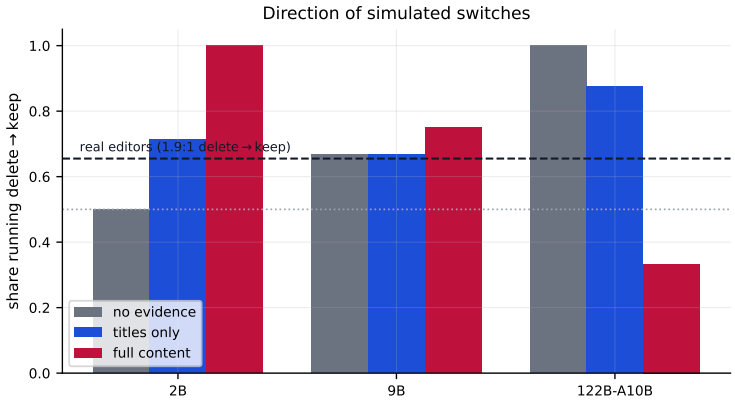

Figure 2: Switch direction per cell against the real 1.9:1 delete→keep. Magnitude and direction come apart: the 122B content cell has near-peak magnitude and inverted direction.

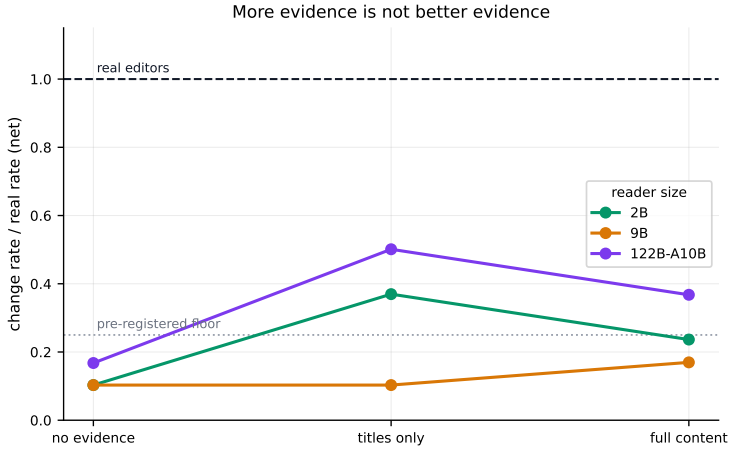

Figure 3: The same nine cells read the other way, as point estimates. Titles out-score full content at 2B and 122B and lose only at 9B; the 2B and 122B rows peak at titles, while the 9B row is weakly increasing. None of these orderings resolves on the debate-level bootstrap (§A.3); the closest-to-real cell on points is 122B with titles, and 2B with titles is the smallest cell that clears the pre-registered band in a majority of resamples.

C Preregistration, amendments, and the pre-write-up audit

Artifacts. Every prompt, chassis, scoring script, per-debate result summary and audit label behind this paper is committed in the research repository that accompanies it, with the scripts that regenerate every table from those artifacts; the repository is to be released with the paper.

Discipline. Every stage locked its hypotheses, arms, metrics, and decision rules before any paid call, in the process log accompanying this paper. Verdicts are three-valued — supported, falsified, inconclusive — and decided on intervals rather than point estimates. Instrument iterations are logged individually with the reason; where an iteration failed twice we stopped rather than continue, which is recorded in §6 for the search affordance.

Locks that failed, listed. The scaffold discriminator passed. Comprehension support under injected policy text was refuted and inverted, and the ceiling-stability lock failed with it at 9B (0.727 against a 0.879 baseline and a ± 0.06 bound) (§5). Both quantitative locks on the repetition arm failed: 2B bare recall reached 0.697 against a 0.70 bound, one switcher short, and 2B with policy text reached 0.485 against a 0.60 bound (§7). The Stage-2 change-rate floor failed at $0.09\times$. The search-exercise lock failed twice. The aggregate change floor failed while the content-bearing subset passed. Volume monotonicity was refuted; the smallest-in-band prediction was wrong twice (§9). The belief layer's baseline lock failed against the full reader. The sovereignty direction-majority lock narrowly failed. The instability threshold forced a restatement of the headline (§8.1).

Amendment: the content-rich slice. After the aggregate floor failed with the mechanism locks passing, we selected a fresh 40-debate slice on kept outcome, genuine switch, and citation presence, disjoint from the pilot. This is selection on properties correlated with the dependent variable and is reported as such: it is the correct population for asking whether the mechanism works, and the wrong one for estimating how often it fires in the wild.

Leakage note. Conviction gradients were fit corpus-wide while the pilot is 60 of 369,215 debates, under 0.02%. Recorded rather than corrected.

Pre-write-up audit. Before drafting, every headline number in the process log was re-checked against the committed run summaries. Six disagreements were found and corrected: an analyzable-debate count wrong in two places (369,215, not 371,942); an extraction count wrong by 169 (453, not 622); a run identifier pointing at a neighbouring stage's timestamp; a repetition result scored against the control arm's baselines rather than the replay arm's own; the non-monotone direction arc discussed in §11; and a reproduction bullet that listed its smaller gaps while omitting its two largest, now in Table 10. Two further items were logged as artifact defects rather than claims: a divide-by-zero sentinel in one committed summary, which the plotting code must skip rather than render, and a stale cross-reference in a script docstring.

None of these were caught by review; they were caught by re-reading the data. The figures and tables in this paper are generated directly from the committed summaries and cannot drift, and a checking script re-verifies every number written in prose against the same files.

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.

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.

Hsuvas Borkakoty and Luis Espinosa-Anke. Wikipedia is not a dictionary, delete! Text classification as a proxy for analysing wiki deletion discussions. In *Proceedings of the Tenth Workshop on Noisy and User-generated Text*, pages 133–142, Albuquerque, New Mexico, USA, May 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.wnut-1.14. URL `https://aclanthology.org/2025.wnut-1.14/`.

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.

Yun-Shiuan Chuang, Ruixuan Tu, Chengtao Dai, You Li, Smit Vasani, Binwei Yao, Michael Henry Tessler, Sijia Yang, Dhavan Shah, Robert Hawkins, Junjie Hu, and Timothy T. Rogers. DEBATE: A large-scale benchmark for evaluating opinion dynamics in role-playing LLM agents, 2025. Workshop poster, Scaling Environments for Agents (SEA), NeurIPS 2025.

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

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`.

Zhitao He, Pengfei Cao, Chenhao Wang, Zhuoran Jin, Yubo Chen, Jiexin Xu, Huaijun Li, Kang Liu, and Jun Zhao. AgentsCourt: Building judicial decision-making agents with court debate simulation and legal knowledge augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9399–9416, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.549. URL `https://aclanthology.org/2024.findings-emnlp.549/`.

Khandaker Tasnim Huq and Giovanni Luca Ciampaglia. Characterizing opinion dynamics and group decision making in wikipedia content discussions. In *Companion Proceedings of the Web Conference 2021 (WWW '21 Companion)*, pages 632–639, New York, NY, USA, 2021. Association for Computing Machinery. doi: 10.1145/3442442.3452354.

Lucie-Aimée Kaffee, Arnav Arora, and Isabelle Augenstein. Why should this article be deleted? Transparent stance detection in multilingual Wikipedia editor discussions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5891–5909,

Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.361. URL https://aclanthology.org/2023.emnlp-main.361/.

Alex Laitenberger, Christopher D Manning, and Nelson F. Liu. Stronger baselines for retrieval-augmented generation with long-context language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32559–32569, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1656. URL https://aclanthology.org/2025.emnlp-main.1656/.

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.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Prompt repetition improves non-reasoning LLMs, December 2025. URL https://arxiv.org/abs/2512.14982.

Yuante Li, Yicheng Tao, Kate Zhang, Taozhi Wang, Gefei Gu, and Yaxin Zhou. Diverse evidence, better forecasts: Multi-agent deliberation under information asymmetry, 2026. URL https://arxiv.org/abs/2607.01661.

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.

Praveen Kumar Myakala, Manan Agrawal, and Rahul Manche. BeliefShift: Benchmarking temporal belief consistency and opinion drift in LLM agents, 2026.

Thang Nguyen, Peter Chin, and Yu-Wing Tai. Ma-rag: Multi-agent retrieval-augmented generation via collaborative chain-of-thought reasoning, 2025. URL https://arxiv.org/abs/2505.20096.

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.

Young Bin Park. Graph-native cognitive memory for AI agents: Formal belief revision semantics for versioned memory architectures, 2026.

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.

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. *Neurocomputing*, 618:129063, 2025. ISSN 0925-2312. doi: 10.1016/j.neucom.2024.129063.

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.

Xingchen Xiao, Heyan Huang, Runheng Liu, and Jincheng Xie. MASS-RAG: Multi-agent synthesis retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026. arXiv:2604.18509.

Joshua C. Yang, Maurice Flechtner, Damian Dailisan, and Michiel A. Bakker. Belief engine: Configurable and inspectable stance dynamics in multi-agent LLM deliberation, 2026. URL `https://arxiv.org/abs/2605.15343`.