Belief Engine: Configurable and Inspectable Stance Dynamics in Multi-Agent LLM Deliberation

Joshua C. Yang [1] **Maurice Flechtner** [2] **Damian Dailisan** [1] **Michiel A. Bakker** [3]

Abstract

LLM-based agents are increasingly used to simulate deliberative interactions such as negotiation, conflict resolution, and multi-turn opinion exchange, but generated behavior alone does not establish a persistent agent state. A stance may change because an agent accepted evidence, remained anchored, drifted from a role, echoed a partner, or received different prompt and retrieval context. We introduce the Belief Engine (BE), an auditable memory-and-belief-update layer that treats "belief" as an evidential state over a proposition and exposes it as scalar stance. BE extracts arguments into structured memory and updates stance with a log-odds rule controlled by evidence uptake u and prior anchoring a. Across multiple base LLMs, parameter sweeps show that these controls reliably shape stance dynamics while preserving an evidence-level update trail. On 2,495 DEBATE human deliberation trajectories, BE best reconstructs participants whose final stances align with extracted received evidence, while stable and evidence-opposed cases indicate anchoring or factors outside the extracted evidence stream. BE provides configurable infrastructure for studying how LLM agents maintain and revise a point of view, tying openness, commitment, convergence, and disagreement to explicit update assumptions rather than hidden prompt effects.

1. Introduction

Large Language Models (LLMs) are moving from single-turn assistants to simulated participants in agent societies, education, civic deliberation, and collective-decision settings. Generative-agent work showed how LLMs can sustain social simulation over time (Park et al., 2023), while recent deliberation and augmented-democracy studies use LLMs to model discussion, preference formation, and citizen input (Yang et al., 2024; Chuang et al., 2025; Gudiño et al., 2024). Digital-twin proposals extend this ambition by treating synthetic communities as test beds for deliberative design (Novelli et al., 2025). Across these settings, researchers need to know what agents say, and also whether, when, and why their stances change in response to reasons. In multi-agent deliberation, these update dynamics matter for cooperation and conflict resolution because they shape when agents listen, preserve commitments, converge, or remain apart.

Because current LLM agents can change expressed agreement and disagreement without exposing the source of that movement, attribution becomes central. A transcript may look more moderate because the agent accepted evidence, because its persona drifted (Choi et al., 2025b), because it echoed its interaction partner (Shekkizhar et al., 2025), or because inherited model biases shaped the response (Taubenfeld et al., 2024). An agent may also appear resistant because retrieval or prompt context changed. The underlying issue is that temporal continuity is not automatic in a standard LLM interaction: without an external memory and update rule, each turn is a new computation over the provided context rather than a persistent deliberative agent carrying experience forward. For agent-based simulations, the modelling target becomes unclear: what kind of deliberative subject is being represented?

We introduce the Belief Engine (BE), an auditable simulation-control layer that puts log-odds belief updating under experimental control. We use "belief" in an operational Bayesian modelling sense: a proposition-level evidential state maintained in log-odds. We use *stance* for the scalar readout $S \in [-1, 1]$, with positive values supporting the proposition and negative values opposing it. Initial seed arguments define the prior, later arguments provide weighted evidence, and both are accumulated through a log-odds rule with two interpretable controls: *evidence uptake* u and *prior anchoring* a. This targets the component of

Accepted at the Trustworthy AI for Good (AI4GOOD) Workshop at ICML 2026. [1]ETH Zurich [2]ETH Zurich and Centre for Democracy Studies Aarau, University of Zurich [3]Massachusetts Institute of Technology. Correspondence to: Joshua C. Yang <joyang@ethz.ch>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

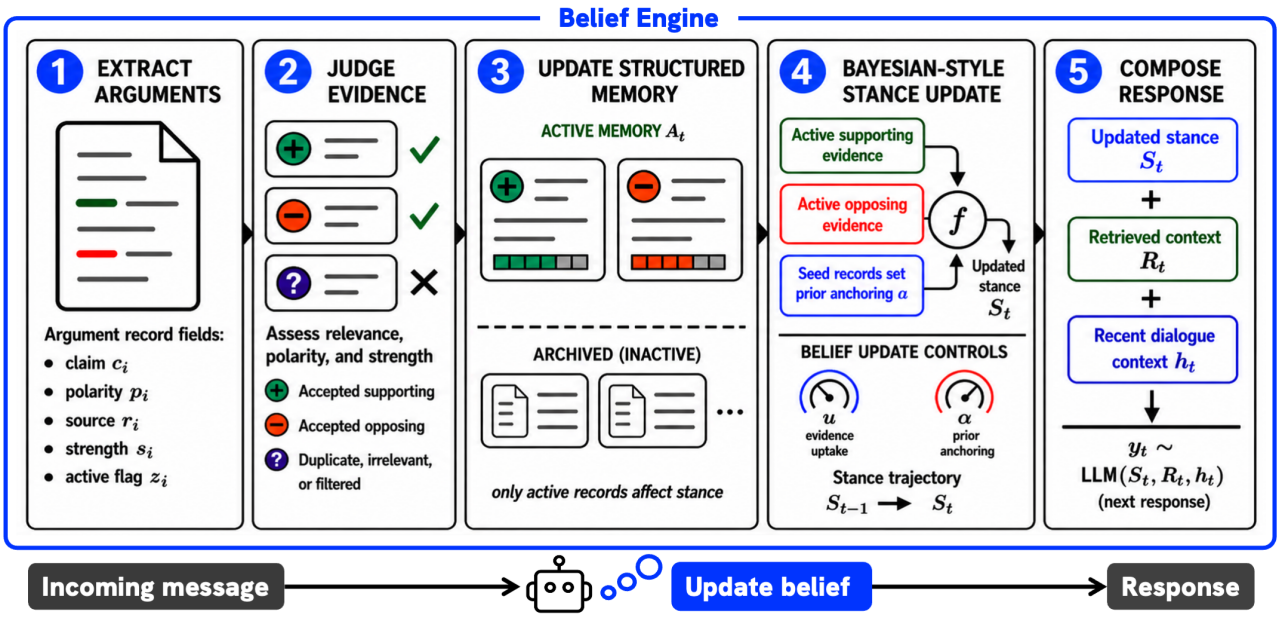

Figure 1. Belief Engine architecture. Incoming messages are extracted into structured arguments, judged, and stored as active evidence or archived records. Active evidence updates the maintained belief state through evidence uptake and prior anchoring; responses are generated from the resulting stance plus retrieved memory and recent dialogue context.

deliberation that current LLM agents leave implicit: what counts as evidence, how much it can move the belief state, and how strongly an initial commitment persists. Incoming arguments are extracted, judged, and stored in structured memory before they enter this update. Language generation is conditioned on the updated stance and retrieved evidence.

Separating the maintained belief state from generation lets researchers state modelling assumptions that are usually buried in prompts. High uptake creates more evidence-responsive agents, strong anchoring preserves initial commitments, and the stance trajectory can be checked against the evidence that produced it. BE therefore turns deliberation simulation from prompt-driven transcript production into a reportable state-transition model: researchers can specify how evidence enters memory, how much it moves the belief state, and how the resulting stance trajectory conditions language.

We evaluate BE with controlled parameter sweeps, matched prompt-based baselines, and replay on 2,495 quality-filtered DEBATE (Chuang et al., 2025) trajectories from human multi-round discussions. The sweeps test whether u and α control evidence responsiveness across base models. The prompt baselines test whether self-revision and retrieval expose the same update process. We use DEBATE replay to understand which kinds of human stance movement are explained by extracted evidence, rather than to claim that one profile should predict the whole population: under the same extracted-evidence stream, which uptake–anchoring profiles reconstruct observed response regimes?

Concretely, BE makes three pieces of the simulation explicit. The architecture separates argument extraction, evidence judgement, structured memory, belief updating, stance computation, and response generation. The generated-agent experiments show that two scalar controls, uptake and anchoring, produce predictable stance dynamics across multiple base models while preserving an evidence-level audit trail. The human replay protocol shows that the same interface can separate evidence-explained movement, stable anchoring, and movement whose signal is absent from the extracted evidence stream.

2. Related work

LLM deliberators blur expression and state. A central difficulty in LLM-based social simulation is attribution: when an LLM agent changes what it says, the transcript alone does not show whether the movement should be treated as evidence-responsive updating or as a surface shift caused by prompting, retrieval, or generation. They exhibit persona shift (Choi et al., 2025b; Shekkizhar et al., 2025), knowledge-conflict failures (Xu et al., 2024), recency sensitivity (Kim et al., 2024; Zhang et al., 2025), regression to training biases (Taubenfeld et al., 2024), and excessive convergence relative to humans (Chuang et al., 2025). DEBATE is especially relevant because it records both public interaction traces and private pre/post opinions, showing that plausible role-play can still distort individual and group opinion dynamics. Debate can also behave like a martingale without a specified update policy (Choi et al., 2025a). Even

if reasoning models internally simulate dialogic "societies of thought" (Kim et al., 2026), those implicit perspectives are not persistent, auditable social actors. BE addresses this gap by maintaining a separate belief state with an explicit stance variable.

Democratic simulations need inspectable change. Applied systems increasingly use LLMs in deliberative settings: AI debate can support factual claim assessment (Rahman et al., 2025) and judicial tasks (Hu et al., 2025), while Agora and ArgueMate scaffold consensus-finding in civic and educational contexts (Pradeep Fulay et al., 2026; Wang et al., 2026). Augmented-democracy and digital-twin proposals extend this ambition by using LLMs or synthetic communities to estimate citizen preferences and test deliberative designs through controlled "what-if" scenarios (Gudiño et al., 2024; Novelli et al., 2025) and DelibSim shows that LLM groups match human procedural discourse quality while failing to reproduce similar epistemic outcome dynamics (Flechtner, 2026). We argue that, in such settings, the missing state variable becomes a substantive problem. A deliberative simulator should report final accuracy, epistemic and discourse quality, and user learning, but it should also expose why simulated participants move, remain stable, or polarise.

Memory does not determine belief revision. Agent memory systems decide what can be retained, reflected on, or retrieved: Generative Agents (Park et al., 2023), episodic memory banks (Zhong et al., 2023), reflection systems (Xu et al., 2025; Tan et al., 2025), graph memory (Gutiérrez et al., 2024; Kang et al., 2025; Huang et al., 2025), large-scale simulations (Piao et al., 2026; Xu et al., 2026), and retrieval-augmented debate (Li et al., 2026) all strengthen some form of context persistence. But remembering an argument is not the same as accepting it as evidence. Memory can also amplify noise or reinforce experience-following behaviour (Xiong et al., 2025). BE places judgement and updating between memory and generation: stored arguments become belief-relevant only when the evidence layer marks them active.

Formal update rules need semantic grounding. Bayesian models (Olsson, 2013), DeGroot updating (DeGroot, 1974), and bounded-confidence models (Deffuant et al., 2000; Hegselmann & Krause, 2002) make opinion dynamics explicit, and recent work connects LLM debate to Bayesian Nash equilibrium (Xie et al., 2025). Their abstraction is also their limitation for language-agent simulations: they rarely specify how natural-language arguments are extracted, judged, remembered, or turned back into utterances. Human belief updating also departs from ideal Bayesian assumptions (Stengård et al., 2022; Holt & Smith, 2009; Ashinoff et al., 2022). People may protect commitments, weigh evidence through identity and trust, and respond differently across social contexts. The BE framework sits between these traditions. It keeps the update rule parameterised, grounds each update in extracted arguments, and exposes profiles that can be calibrated against human deliberation trajectories.

3. Method

Figure 1 depicts a BE agent as an LLM-based debater whose proposition-level belief state is maintained by the Belief Engine. During a debate it exchanges messages with an opponent on a fixed topic. The agent can use any compatible base model. Across the generated-agent comparisons we use GPT-4o-mini, GPT-5.4-mini, Qwen 3.5 9B, and Gemma 4 E4B in different roles, as detailed below. In the generated-agent conditions, debates start from $n{=}10$ seeded arguments and use the same dialogue protocol while varying the uptake-anchoring profile (u, a). Each response is generated from three inputs: (i) current stance instruction, (ii) retrieved context, and (iii) recent dialogue context. Since the belief state is updated by reported parameters and exposed as S, the architecture makes it possible to study how simulated agents change state under argumentative pressure.

3.1. Belief engine

The Belief Engine decouples belief maintenance from generative reasoning so that the agents' belief updating can be controlled systematically. For each new message, the engine follows the five-step loop (Fig. 1): *Extract Arguments* (convert a message into candidate claims), *Judge Evidence* (score and merge near-duplicate candidates), *Update Structured Memory* (store active and archived evidence), *Update Belief State* (apply the log-odds rule and compute stance S), and *Compose Response* (retrieve active memory in proportion to its pro/con composition and map S to behavioural instructions for generation).

The reported Belief Engine results use a deterministic log-odds updater that recomputes the belief state from argument polarity, support strength, evidence uptake u, and prior anchoring a, then exposes it as stance S. Each stored argument has a binary active flag $z_i \in \{0, 1\}$, implemented as `credence_relevant`. Only active records contribute to stance $S \in [-1, 1]$, separating record-level evidence selection from the aggregate stance.

Extract arguments. An independent LLM module decomposes each message m_t into candidate records $\tilde{\mathcal{E}}_t = J(m_t) = \{\tilde{e}_i = (c_i, p_i, r_i)\}_{i=1}^{n_t}$, where c_i is the canonical claim, $p_i \in \{-1, +1\}$ is its polarity ($+1$ affirmative, -1 negative), and $r_i \in \{\texttt{seed}, \texttt{self}, \texttt{opponent}\}$ is the source role. We use the same extractor across experiments, while the base debater model varies in generated-agent comparisons. Polarity is defined relative to the debate propo-

sition, not sentiment. A single message may yield zero, one, or several candidate arguments. Extraction proposes candidate evidence records, it does not by itself determine whether a claim should affect the belief state. Rather than restricting extraction to formal, logically structured arguments, we adopt a broader conception of deliberative communication that includes narratives, rhetorical forms, and experiential knowledge (Young, 2002; Nakazawa et al., 2024). Accordingly, any persuasive contribution, whether formal reasoning, anecdotal evidence, or rhetorical framing, is treated as a valid unit of analysis.

Judge evidence. The judgement layer governs how candidate arguments are transformed into structured evidence. Following Xiong et al. (2025), we treat scoring and conflict resolution as separate architectural choices with explicit outputs. **Strength scoring.** Each extracted argument receives a strength $s \in [0, 1]$ from either the LLM extractor (s_i^{LLM}) or a classifier (s_i^{clf}). The classifier is `DeBERTa-v3-large` (He et al., 2021) fine-tuned as a regressor on crowd-rated argument-quality labels (Gretz et al., 2019), mapping each (topic, claim) pair to a scalar score in $[0, 1]$. We set $s_i = s_i^{\text{clf}}$ when classifier scoring is enabled and $s_i = s_i^{\text{LLM}}$ otherwise. Sweeps may use either scorer, while DEBATE replay fixes $s_i = s_i^{\text{clf}}$ so calibration varies only belief-update parameters. Each record is then written as $e_i = (c_i, p_i, s_i, r_i, z_i)$ with $z_i \leftarrow 1$ prior to conflict resolution. **Conflict resolution.** To avoid repeated paraphrases causing repeated updates, we soft-deduplicate same-polarity active claims. For a new argument, we compare its claim embedding against active records with the same polarity. If none exist, or if the nearest match has cosine similarity below the configured threshold θ, the new record stays active. Otherwise, we keep only the stronger record active and archive the other. Records are retained for auditability, but only active ones ($z = 1$) affect the belief state. Appendix B gives the replacement rule and reports how θ is set.

Update structured memory. After conflict resolution, the BE agent stores each claim as a structured `ArgumentRecord` with polarity, strength estimates, source role, and `credence_relevant` status. Accepted evidence remains active, while superseded evidence is archived for traceability. This selective consolidation supports record-level replacement during conflict resolution and later retrieval by active pro/con memory composition.

Update belief state. The Belief Engine exposes an uptake-anchoring (UA) profile (u, a): *evidence uptake* u determines how strongly new evidence shifts the belief state, while *prior anchoring* a sets the strength of the initial prior.

Belief updating is modelled as additive evidence accumulation in log-odds space over the currently active memory, with bounded stance S_t. Let $\mathcal{A}_t = \{i : z_i = 1\}$ denote the active-memory index set after judgement and conflict resolution at time t. Rather than retaining stale contributions from arguments that have since been archived, the log-odds state is recomputed from the active set. With $\gamma_i = a$ for seed records ($r_i = \texttt{seed}$) and $\gamma_i = u$ for later debate records ($r_i \neq \texttt{seed}$), the update is

$$L_t = \sum_{i \in \mathcal{A}_t} p_i \ln(1 + s_i \gamma_i)\,,$$
$$S_t = 2\sigma(L_t) - 1 = \frac{2}{1 + \exp(-L_t)} - 1, \quad S_t \in [-1, 1]. \tag{1}$$

Seed records represent the prior, with no separate intercept: a sets how strongly seed records establish the initial belief state, while u sets how strongly later debate evidence shifts it. Evidence accumulates linearly in log-odds, while the logistic transform yields bounded stance and diminishing returns near certainty. Here, Bayesian refers to the log-odds evidence-accumulation form: the rule accumulates argument weights as additive evidence over prior seed records, with likelihood-like quantities operationalised through argument extraction and evidence scoring. If records are never archived or replaced, Eq. 1 reduces to an incremental update. Archived records remain auditable but no longer affect the belief state.

Compose response. After updating the belief state, the BE agent retrieves a bounded context R_t from active memory. Let $\mathcal{M}_t^+$ and $\mathcal{M}_t^-$ denote the active affirmative and negative records. Retrieval allocates $k_+ = \text{round}\big(k|\mathcal{M}_t^+|/(|\mathcal{M}_t^+| + |\mathcal{M}_t^-|)\big)$ slots to affirmative evidence and $k_- = k - k_+$ to negative evidence, with an even split when active memory is empty. R_t is the union of the strongest k_+ and k_- active records by support strength. Retrieval therefore follows accepted evidence instead of directly amplifying scalar stance. The next utterance is generated as $y_t \sim \text{LLM}(S_t, R_t, h_t)$, where S_t is mapped to a 10-bin stance-intensity instruction and h_t is recent dialogue context.

3.2. Experimental setup

We separate two empirical targets. Generated-agent experiments evaluate the full loop, including generation, judgement, memory, retrieval, and belief updating. DEBATE replay isolates the update rule under real received-evidence histories, giving a human-grounded calibration test for the resulting stance dynamics.

Generated-agent experiments. Seed and counter-agent arguments come from the Argument Quality Ranking dataset (Gretz et al., 2019), which contains about 30,000

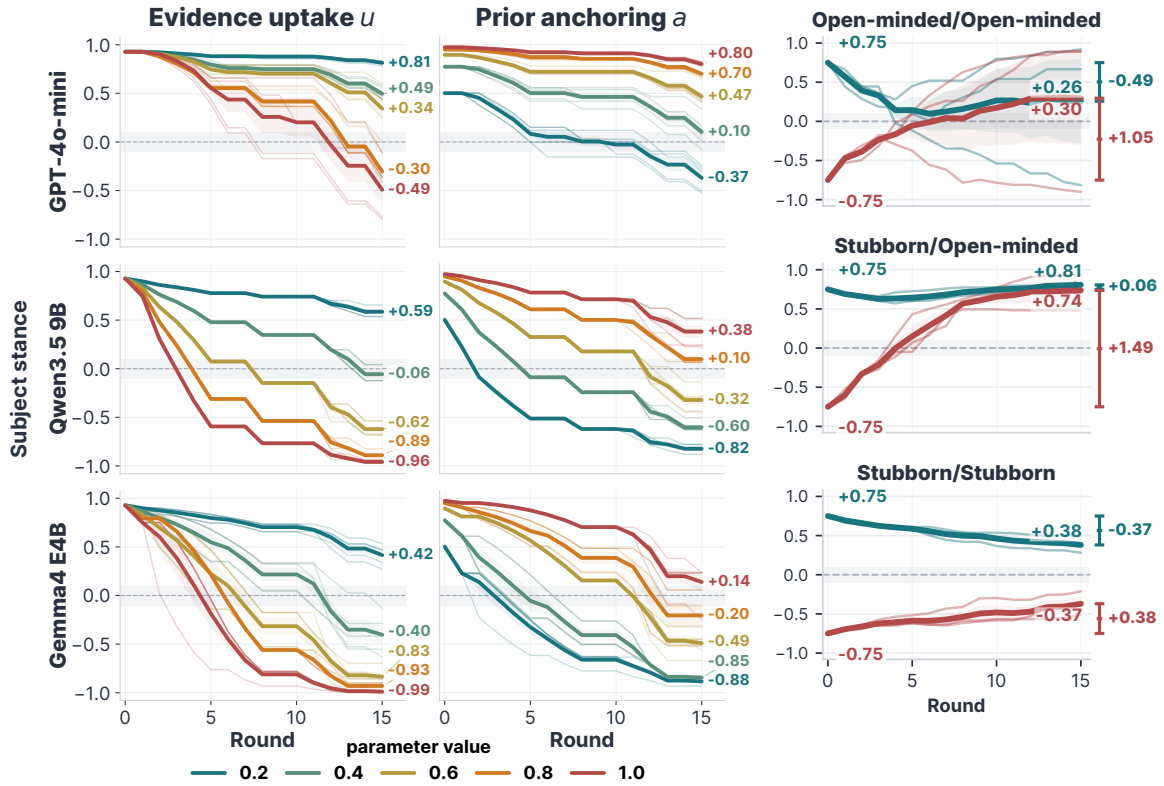

Figure 2. Parameter control and profile dynamics. Stance $S \in [-1, 1]$, where $+1$ means affirmative/pro with respect to the proposition and -1 means negative/con. **Left:** BE single-agent sweeps across three base LLMs, varying evidence uptake u ($a = 0.70$) and prior anchoring a ($u = 0.4$). **Right:** GPT-5.4-mini two-agent profile debates with Open-minded $(u, a) = (0.40, 0.20)$ and Stubborn $(0.10, 0.80)$ agents. Thin lines show trials, thick lines show means.

crowd-annotated arguments with topic, polarity, and quality scores. All generated debates run for 15 rounds with n=10 seed arguments per side and agent temperature T=0.7. Symmetric two-agent debate provides a controlled setting for testing update mechanics, profile dynamics, and evidence-conditioned generation.

In the main text, we focus on (i) five-point parameter sweeps over u or a across three base models (five trials per setting) and (ii) matched prompt baselines, namely prompt self-update and RAG plus self-update. Appendix diagnostics add Open-minded/Stubborn profile and topic-grid demonstrations over 10 topics with three debates per profile-pairing cell. When a system does not expose S, a shared external LLM judge receives the proposition and generated text, then returns a scalar stance in $[-1, 1]$ at temperature 0.0. Hyperparameters appear in Tab. A5.

Human replay validation. DEBATE replay uses the benchmark as observed-evidence replay rather than as one of its original simulation tasks. For each participant, we initialise BE from the private initial Likert stance, treating that mapped stance as the prior state scaled by anchoring a. We then replay directed evidence from partner tweets and received chat messages, and compare the final BE stance with the private final Likert stance. Six-point Likert responses are mapped linearly to $S \in [-1, 1]$. We perform a grid search over (u, a) and select settings by held-out RMSE on final stance. We report two references: a no-change baseline that predicts the final stance from the initial stance, and a net-evidence linear baseline, $\hat{S}_{\text{final}} = S_{\text{initial}} + \beta E$, that fits one scalar evidence weight on training folds only. The linear baseline uses the same quality-filtered received-evidence stream, human-like uptake and deduplication filter, and classifier strength scores as BE, but omits the Bayesian log-odds belief-update rule. The main replay excludes self-authored posts and messages, isolating the belief-update component under the evidence a participant received.

4. Results

The results ask three questions. First, is the maintained stance S visible in generated text strongly enough to support comparison with external stance scores? Second, do uptake and anchoring give predictable control across base models, and do prompt-only self-revision or RAG expose a comparable update trail? Third, when replayed on DEBATE,

which final-stance movements are explained by extracted received evidence, and which point beyond it? Generated-agent experiments test controllability and auditability, while DEBATE replay tests how the same update interface captures human response heterogeneity. Appendix D reports auxiliary diagnostics for the extractor and judgement layer.

4.1. Uptake and anchoring control stance dynamics

We first validate that the maintained stance S is expressed in generated language. We sweep $S \in [-1, 1]$ and ask the independent judge used for non-BE baselines to score each generated response. The scores align tightly with the assigned stance ($r = 0.967$, $p < 0.001$), with a fitted slope of ≈ 0.86 indicating mild compression at the extremes. Because the judge sees only text, it provides a calibrated behavioural score for systems that do not expose S.

We next test the two controls directly. We sweep *evidence uptake* u and *prior anchoring* a across five levels ($\{0.2, \ldots, 1.0\}$) in 15-round debates against a deterministic counter-argument opponent on the proposition "*We should introduce compulsory voting*". We repeat the sweep with three base language models (GPT-4o-mini, Qwen 3.5 9B, and Gemma 4 E4B) to test whether the pattern depends on the generator. Figure 2 shows the intended ordering. Higher u makes agents more responsive to new evidence: for GPT-4o-mini, the final stance shifts from $+0.81$ (u=0.2) to -0.49 (u=1.0), and Qwen and Gemma move even further in the high-uptake setting (-0.96 and -0.99). Higher a makes the seeded prior more persistent. Holding u=0.4, GPT-4o-mini ranges from -0.37 (a=0.2) to $+0.80$ (a=1.0), and Qwen and Gemma follow the same pattern.

Figure 2 (right) demonstrates that the same rule could also create recognisable two-agent profiles. Open/Open agents reduce disagreement, Stubborn/Open pulls the open con agent toward the anchored pro agent, and Stubborn/Stubborn preserves more of the initial gap. The endpoints differ by base model and trial because retrieval keeps active evidence from both sides available, so generation still affects the rate and magnitude of movement. This is the desired behaviour for simulation: openness, asymmetric anchoring, and mutual anchoring become explicit settings rather than implicit prompt effects.

4.2. Prompt baselines show weaker convergence without update trails

Prompt baselines test whether self-revision instructions and retrieval are enough to produce comparable convergence without a separately maintained state. The matched two-agent debates use 15 rounds, 3 trials, and initial stances ± 0.75 (Fig. 3).

With the BE, memory, retrieval, and the update profile remain fixed at $(u, a) = (0.2, 0.4)$ while the base generator varies. All four BE conditions reduce the initial stance gap, with final gaps of 0.50 (GPT-4o-mini), 0.23 (GPT-5.4-mini), 0.22 (Qwen), and 0.07 (Gemma), each traceable to active evidence under the shared profile.

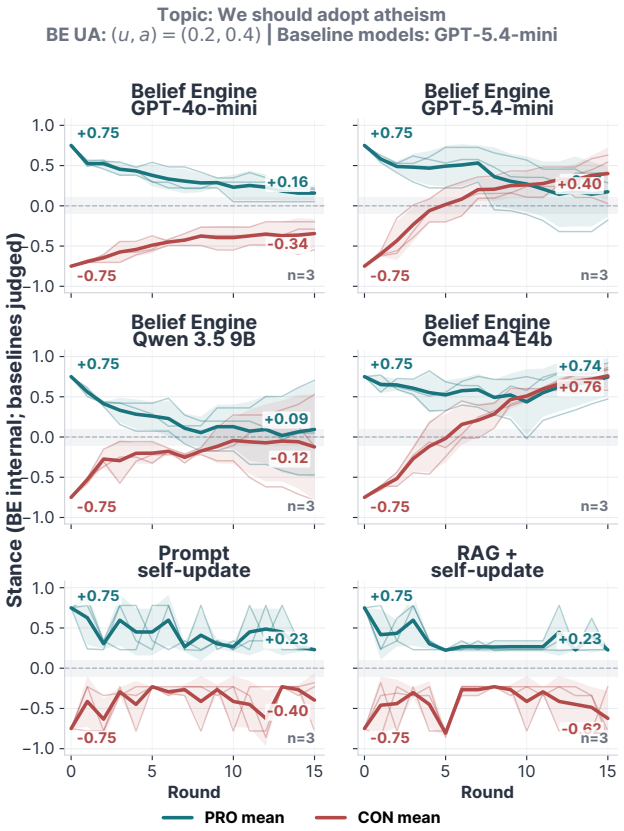

Figure 3. Prompt-baseline comparison on 15-round atheism debates. BE variants use fixed $(u, a) = (0.2, 0.4)$ and expose internal stance; prompt self-update and RAG plus self-update use external-judge scores on the same $[-1, 1]$ scale.

The GPT-5.4-mini prompt self-update and RAG plus self-update baselines also move in the external-judge score, but their convergence is weaker. They end with wider Pro/Con gaps than most BE variants: $(+0.23, -0.40)$ and $(+0.23, -0.62)$. BE therefore adds what prompt-level self-revision lacks: an internal belief state, reported uptake and anchoring parameters, and an evidence-level update trail.

4.3. Human replay diagnoses which movements are explained by extracted evidence

DEBATE replay fixes the received-evidence stream and fits only the update rule. We use it as a diagnostic heterogeneity test: which uptake–anchoring profile reconstructs the observed final stance? The subgroup partitions are outcome-conditioned diagnostics from observed final movement, so they should be read as replay analyses rather than

pre-outcome prediction rules. Figure 4 shows the calibration surface. Each heatmap holds out final human stances for one participant subset, sweeps u and a, and colours each cell by how much its RMSE exceeds that panel's best cell. The pooled surface favours low uptake with moderate anchoring. After partitioning by evidence alignment, the minima move to different regions: evidence-aligned movers favour high uptake, evidence-opposed movers favour near-zero uptake, and stable participants favour near-zero uptake with maximal anchoring. The panel therefore shows why a single default profile is misspecified for a mixed population.

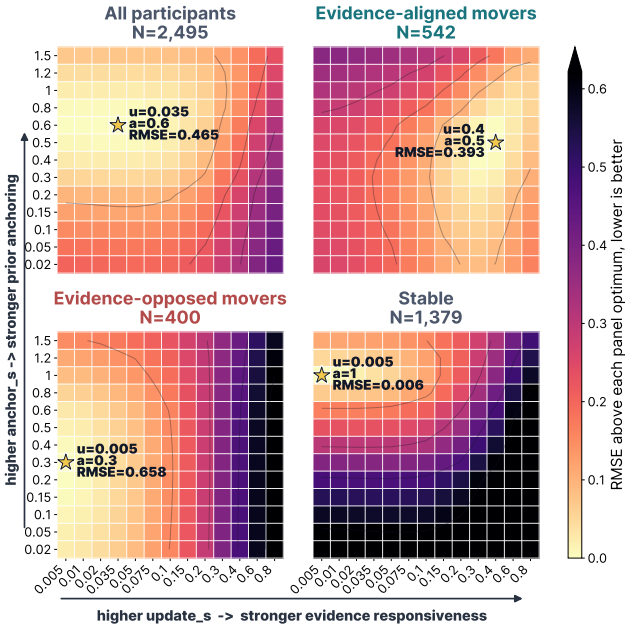

Figure 4. DEBATE replay calibration surfaces. Each heatmap sweeps evidence uptake u and prior anchoring a for one participant subset. Cells show held-out RMSE above that panel's best cell; lower is better. Star sign marks the optimum, and the printed RMSE is the absolute error at that optimum.

Table 1 gives the held-out error version and mean observed movement. In the pooled population, the BE reduces RMSE from 0.489 under no-change and 0.488 under net-evidence linear replay to 0.465. Outcome-conditioned diagnostic gains are larger. Evidence-aligned and evidence-opposed movers have similar movement magnitudes (mean $|\Delta| = 0.619$ and 0.633), but require different profiles: aligned movers are reconstructed by high uptake (0.393 RMSE), while opposed movers flag cases where extracted evidence points against observed movement. Stable participants are not evidence of predictive improvement over no-change, since no-change is exact by definition for this group; instead, they check that BE can represent resistance to extracted evidence without forcing movement. The same diagnostic profile parameters outperform prior-only and pooled global replay references in all five folds under both group- and topic-held-out splits (Appendix H), supporting the interpretation that response heterogeneity is not captured by a single global (u, a) setting.

Table 1. Held-out DEBATE replay RMSE on mapped final stance.

Group	N	$\|\Delta\|$	No-ch.	Linear	BE	Gain
All participants	2,495	0.283	0.489	0.488	0.465	0.023
Evidence-aligned movers	542	0.619	0.718	0.689	0.393	0.296
Evidence-opposed movers	400	0.633	0.737	0.765	0.658	0.107
Stable participants	1,379	0.000	0.000	0.037	0.006	0.031

Note. No-ch. predicts initial stance; Linear fits a train-fold scalar on signed extracted evidence; BE uses calibrated uptake–anchoring profiles. Gain is Linear minus BE; $|\Delta|$ is mean absolute human movement. Subgroups are outcome-conditioned diagnostics; 174 weak-signal movers appear only in All participants.

At the single-participant level, Fig. 5 provides explanatory traces for the update mechanism: aligned endpoints show cases where accepted evidence accounts for movement, while divergence points to interpretation, social context, or arguments not captured by extraction. These patterns separate cases that would otherwise look similar in a prompt-only agent: movement with accepted evidence, stability under evidence pressure, and movement that the extracted evidence cannot explain.

5. Discussion and future work

These experiments suggest that the value of BE is not to make every deliberative trajectory predictable, but to make change attributable. "Belief" here is an operational evidential state over a proposition, maintained in log-odds and exposed through scalar stance. Under this definition, researchers can specify what counts as evidence, how strongly it shifts the belief state, and how strongly initial commitments persist. AQR diagnostics show reliable directed-evidence extraction on clean single-argument inputs (96.9% polarity accuracy), DEBATE replay improves over no-change and net-evidence linear references while showing why a single profile is insufficient for a heterogeneous population, and the prompt comparison shows that externally scored stance change does not expose the update process itself. This operational definition makes BE easier to compare across models, prompts, and replay settings, but it also limits what the framework can claim to model. A single proposition-level stance cannot capture every way people deliberate: they may partly agree, trade off values, distrust a speaker, or soften rhetorically without moving on the headline proposition. These cases do not undermine the belief-update layer, but they do show where the next modelling boundary lies. A natural extension is a multidimensional state in which arguments and experiences are routed to different facets, and retrieval brings back the records relevant to each facet when the agent reasons or responds. In that setting, RAG-style or agentic memory could help simulate a broader range of human experience

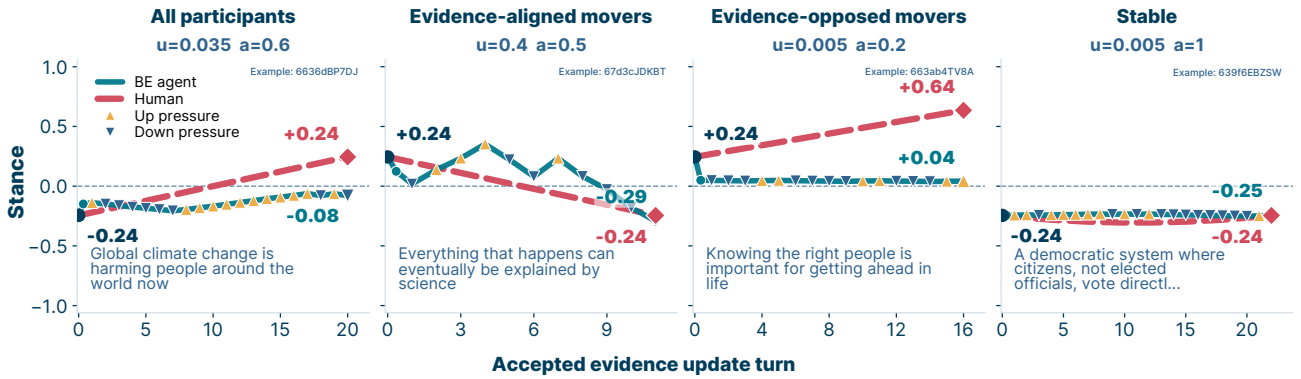

Figure 5. Single-participant DEBATE replay examples. Each panel samples one participant from a diagnostic subset and applies that subset's calibrated (u, a) profile, shown above the panel. The x-axis counts accepted received-evidence updates rather than debate rounds. Starting from the observed human initial stance, the teal curve shows the BE replay after each accepted evidence item. Markers indicate whether the extracted item pushes stance upward or downward. The red dashed line connects only the observed human initial and final stance, since DEBATE does not measure intermediate human stances.

and opinion change while preserving the same inspectable update trail.

The DEBATE replay also clarifies what calibration means for deliberative agents. We do not expect one global profile to explain everyone, and averaging all participants together can hide behaviours that a simulator may need to represent. Some people stay close to their initial stance, some move with the extracted evidence they receive, and others move in ways the current evidence stream does not explain. BE makes these differences explicit rather than smoothing them away: a stable endpoint can be represented as anchoring, evidence-aligned movement as uptake, and divergence from extracted evidence as a signal that some social or interpretive factor is missing. Which profile is appropriate depends on the deliberative setting; our point is that these profiles can be separated, inspected, and reported rather than collapsed into a single population average.

This matters because DEBATE is one deliberative context, not the template for all deliberation. Its participants encounter partner posts and chat messages in a short online setting, and most do not show mapped Likert movement. That stability is informative for online opinion-exposure simulations, but it should not be confused with settings such as citizens' assemblies, facilitated workshops, or in-person deliberation, where participants may enter with an explicit norm of listening, compromise, and joint problem solving. Separating uptake, anchoring, and response profiles lets BE change the assumed deliberative context: a citizens' assembly simulation may contain more evidence-responsive or compromise-seeking agents, or agents whose movement appears first on sub-issues and value trade-offs rather than on the headline proposition.

Finally, evidence assessment remains a modelling choice. BE makes this choice explicit, but it does not make it value-free. LLM extractors, base models, external judges, and the AQR-trained strength classifier can all reflect training-data, safety-tuning, and community-specific priors. This matters most for moral and political topics, where argument quality is contested. For new domains, the judgement layer should be calibrated against domain data and, where appropriate, audited with human or participatory feedback. The benefit of BE is that such choices are isolated in a reportable layer rather than hidden in prompting or generation.

The broader implication is that a simulated persona should not be only a role description, it should also specify how the agent carries information across time. A prompt can say who an agent is, while a memory-and-judgement process specifies what the agent treats as experience, which parts of that experience count as evidence, and how much they can revise persistent commitments. For trustworthy deliberative AI, this traceability matters: convergence, polarisation, and stability should be accountable to visible update assumptions rather than inferred from plausible transcripts alone.

6. Conclusion

LLM agents struggle to maintain stable beliefs: they can drift from assigned positions, echo partners, or change expressed stances without a persistent state. This paper treats memory as a mechanism for agent continuity and studies it through stance dynamics in social simulation. The Belief Engine externalises belief maintenance into a structured layer governed by explicit Bayesian parameters, evidence uptake u and prior anchoring a. By separating what agents say from the belief state they carry, BE makes stance change traceable to extracted arguments and update assumptions. Across base models, these parameters produce distinct profiles, from open-minded deliberation to stubborn anchoring;

in DEBATE replay, BE best reconstructs participants whose final stances follow extracted received evidence, while stable and evidence-opposed cases mark where anchoring, interpretation, social context, or missing signals fall outside the model. The result is a memory-based account of how an agent can maintain and revise a point of view over time.

Impact Statement

This work aims to make deliberative AI systems easier to study, compare, and govern. If LLM agents are used in civic discussion, education, negotiation, or collective decision-making, designers need to know when agents listen, preserve commitments, converge, or remain in disagreement, and why. BE contributes to this goal by making evidence uptake, anchoring, and memory inspectable rather than hidden inside prompts. This could support more transparent simulations, better auditing of agent behaviour, and systems that encourage reflection rather than unexamined prompt effects. The main risk is over-interpretation: synthetic trajectories, even when auditable, should not be treated as predictions of human opinion change or as evidence that a deliberative design will work in a real community without domain calibration and participatory validation. We therefore view BE as infrastructure for transparent experimentation, not a substitute for human judgement or domain-specific governance.

References

Ashinoff, B. K., Buck, J., Woodford, M., and Horga, G. The effects of base rate neglect on sequential belief updating and real-world beliefs. *PLOS Computational Biology*, 18 (12):e1010796, December 2022. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1010796.

Choi, H. K., Zhu, X., and Li, S. Debate or vote: Which yields better decisions in multi-agent large language models? In *Advances in Neural Information Processing Systems*, 2025a.

Choi, J., Hong, Y., Kim, M., and Kim, B. Examining identity drift in conversations of LLM agents. *arXiv preprint arXiv:2412.00804*, 2025b.

Chuang, Y.-S., Tu, R., Dai, C., Vasani, S., Li, Y., Yao, B., Tessler, M. H., Yang, S., Shah, D., Hawkins, R., Hu, J., and Rogers, T. T. Debate: A large-scale benchmark for multi-agent opinion dynamics. *arXiv preprint arXiv:2510.25110*, 2025.

Deffuant, G., Neau, D., Amblard, F., and Weisbuch, G. Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3(01n04):87–98, 2000.

DeGroot, M. H. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974.

Flechtner, M. Procedural parity, outcome mismatch: Evaluating human vs llm deliberation. In *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI EA '26, pp. 14, Barcelona, Spain, 2026. ACM. doi: 10.1145/3772363.3798499.

Gretz, S., Friedman, R., Cohen-Karlik, E., Toledo, A., Lahav, D., Aharonov, R., and Slonim, N. A large-scale dataset for argument quality ranking: Construction and analysis. *arXiv preprint arXiv:1911.11408*, 2019.

Gudiño, J. F., Grandi, U., and Hidalgo, C. Large language models (llms) as agents for augmented democracy. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 382 (2285):20240100, 11 2024. doi: 10.1098/rsta.2024.0100. URL https://doi.org/10.1098/rsta.2024.0100.

Gutiérrez, B. J., Shu, Y., Gu, Y., Yasunaga, M., and Su, Y. Hipporag: Neurobiologically inspired long-term memory for large language models. In *ACL*, 2024.

He, P., Gao, J., and Chen, W. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021. doi: 10.48550/arXiv.2111.09543.

Hegselmann, R. and Krause, U. Opinion dynamics and bounded confidence: models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, 5 (3), 2002.

Holt, C. A. and Smith, A. M. An update on Bayesian updating. *Journal of Economic Behavior & Organization*, 69(2):125–134, February 2009. ISSN 0167-2681. doi: 10.1016/j.jebo.2007.08.013.

Hu, T., Tan, Z., Wang, S., Qu, H., and Chen, T. Multi-agent debate for llm judges with adaptive stability detection. In *Advances in Neural Information Processing Systems*, 2025.

Huang, Z., Tian, Z., Guo, Q., Zhang, F., Zhou, Y., Jiang, D., and Zhou, X. Licomemory: Lightweight and cognitive agentic memory for efficient long-term reasoning. *arXiv preprint arXiv:2511.01448*, 2025.

Kang, J., Ji, M., Zhao, Z., and Bai, T. Memory OS of AI agent. *arXiv preprint arXiv:2506.06326*, 2025.

Kim, J., Lai, S., Scherrer, N., Agüera y Arcas, B., and Evans, J. Reasoning models generate societies of thought. *arXiv preprint arXiv:2601.10825*, 2026.

Kim, M., Kim, S., and Thorne, J. From evidence to belief: A bayesian epistemology approach to language models. *arXiv preprint arXiv:2504.19622*, 2024.

Li, M., Wang, Z., Li, H., and Liu, J. R-debater: Retrieval-augmented debate generation through argumentative memory. In *Proceedings of AAMAS 2026*, 2026.

Nakazawa, T., Tatsumi, T., Souma, Y., and Ohnuma, S. An effect of storytelling on attitude changes in deliberative mini-publics. *Journal of Deliberative Democracy*, 20(1), 2024. doi: 10.16997/jdd.1426.

Novelli, C., Argota Sánchez-Vaquerizo, J., Helbing, D., Rotolo, A., and Floridi, L. Testing deliberative democracy through digital twins. *SSRN preprint*, 2025. doi: 10.2139/ssrn.5193012.

Olsson, E. J. A Bayesian Simulation Model of Group Deliberation and Polarization. In Zenker, F. (ed.), *Bayesian Argumentation*, volume 362, pp. 113–133. Springer Netherlands, Dordrecht, 2013. ISBN 978-94-007-5356-3 978-94-007-5357-0. doi: 10.1007/978-94-007-5357-0_6.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023.

Piao, J., Yan, Y., Zhang, J., Li, N., Yan, J., Lan, X., Lu, Z., Zheng, Z., Wang, J. Y., Zhou, D., Gao, C., Xu, F., Zhang, F., Rong, K., Su, J., and Li, Y. Agentsociety: Large-scale simulation of LLM-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*, 2026.

Pradeep Fulay, S., Ravi, P., Gokhale, O., Yi, E., Bakker, M. A., and Roy, D. Agora: Teaching the skill of consensus-finding with ai personas grounded in human voice. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI EA '26, pp. 1–6. ACM, April 2026. doi: 10.1145/3772363.3798888. URL http://dx.doi.org/10.1145/3772363.3798888.

Rahman, S., Issaka, S., Suvarna, A., Liu, G., Shiffer, J., Lee, J., Parvez, M. R., Palangi, H., Feng, S., Peng, N., Choi, Y., Michael, J., Jiang, L., and Gabriel, S. Ai debate aids assessment of controversial claims. In *Advances in Neural Information Processing Systems*, 2025.

Shekkizhar, S., Cosentino, R., Earle, A., and Savarese, S. Echoing: Identity failures when LLM agents talk to each other. *arXiv preprint arXiv:2511.09710*, 2025.

Stengård, E., Juslin, P., Hahn, U., and van den Berg, R. On the generality and cognitive basis of base-rate neglect. *Cognition*, 226:105160, September 2022. ISSN 0010-0277. doi: 10.1016/j.cognition.2022.105160.

Tan, Z., Yan, J., Hsu, I.-H., Han, R., Wang, Z., Le, L., Song, Y., Chen, Y., Palangi, H., Lee, G., Iyer, A. R., Chen, T., Liu, H., Lee, C.-Y., and Pfister, T. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8416–8439. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.413.

Taubenfeld, A., Dover, Y., Reichart, R., and Goldstein, A. Systematic Biases in LLM Simulations of Debates. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 251–267, 2024. doi: 10.18653/v1/2024.emnlp-main.16.

Wang, C., Forte, F., Wettstein, L., Dillenbourg, P., and Wambsganss, T. Arguemate: Designing an arguing agent with maximised disagreement to support student peer-argumentation exercise. In *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI EA '26, pp. 1–9, Barcelona, Spain, 2026. ACM. doi: 10.1145/3772363.3799300.

Xie, Y., Zhou, Z., Cao, C., Niu, Q., Liu, T., and Han, B. From debate to equilibrium: Belief-driven multi-agent LLM reasoning via bayesian nash equilibrium. In *Proceedings of ICML 2025*, 2025.

Xiong, Z., Lin, Y., Xie, W., He, P., Liu, Z., Tang, J., Lakkaraju, H., and Xiang, Z. How memory management impacts LLM agents: An empirical study of experience-following behavior. *arXiv preprint arXiv:2505.16067*, 2025.

Xu, R., Qi, Z., Guo, Z., Wang, C., Wang, H., Zhang, Y., and Xu, W. Knowledge conflicts for LLMs: A survey. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8541–8565, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.486.

Xu, W., Liang, Z., Mei, K., Gao, H., Tan, J., and Zhang, Y. A-mem: Agentic memory for LLM agents. *arXiv preprint arXiv:2502.12110*, 2025.

Xu, Y., Zhang, S., Zhou, Y., Zeng, S., Lakshmanan, L. V. S., and Ma, C. Topology-aware LLM-driven social simulation: A unified framework for efficient and realistic agent dynamics. *arXiv preprint arXiv:2604.18011*, 2026.

Yang, J. C., Dailisan, D., Korecki, M., Hausladen, C. I., and Helbing, D. Llm voting: Human choices and ai collective decision-making. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):1696–1708, Oct. 2024. doi: 10.1609/aies.v7i1.31758.

Young, I. M. *Inclusion and Democracy*. Oxford University Press, 2002. ISBN 9780198297550. doi: 10.1093/0198297556.001.0001.

Zhang, J., Yang, J., and Wang, K. Large language models as discounted bayesian filters. *arXiv preprint arXiv:2512.18489*, 2025.

Zhong, W., Guo, L., Gao, Q., Ye, H., and Wang, Y. Memorybank: Enhancing large language models with long-term memory. *arXiv preprint arXiv:2305.10250*, 2023. doi: 10.48550/arXiv.2305.10250.

A. Design Rationale Details

Table A1. Design rationale for the Belief Engine. Each component gives researchers a separate handle on belief-state change in deliberative-agent simulations.

Design choice	Rationale	Alternative	Risk addressed
External belief state	Keeps belief maintenance separate from response generation.	Prompt-only persona or hidden context.	Reduces untraceable drift and context-dependent stance changes.
Argument-level evidence	Stores claims, polarity, strength, and provenance as structured records.	Free-form memory summaries.	Preserves auditability and enables record-level ablations.
Judgement layer	Makes extraction, polarity, and strength estimation explicit modules.	Let the generator implicitly decide what mattered.	Exposes where bias, calibration error, or disagreement enters.
Credence-relevance flag	Separates archived evidence from evidence currently affecting the belief state.	Delete weak or conflicting memories.	Keeps a trace of rejected/superseded evidence without letting it update the belief state.
Log-odds update rule	Provides ordered, bounded, reproducible belief-state updates with interpretable parameters and scalar stance.	Direct LLM numerical updates.	Avoids stochastic jumps and makes uptake/anchoring tunable.
Memory-composition-proportional retrieval	Conditions future responses on accepted evidence while keeping minority-side active memories eligible for retrieval.	Stance-conditioned or semantic-only retrieval.	Avoids making the current stance itself determine evidence exposure, while still linking accepted memory to observable language.

B. Conflict-Resolution Details

The main text describes conflict resolution at a high level. Here we give the exact rule used to decide whether a candidate argument remains active. Let $\phi(c)$ be the embedding of claim c and define cosine similarity as

$$\mathrm{sim}(i,j) = \frac{\phi(c_i)^\top \phi(c_j)}{\|\phi(c_i)\|\,\|\phi(c_j)\|}.$$

For a new argument e_i, if no active same-polarity record exists, we set $z_i = 1$. Otherwise, let

$$j^\star = \arg\max_{j: z_j = 1,\ p_j = p_i} \mathrm{sim}(i,j).$$

If $\mathrm{sim}(i, j^\star) < \theta$, where θ is set by the corresponding experiment configuration, the new argument remains active. If the similarity exceeds the threshold, only the stronger of the two near-duplicate records remains active:

$$(z_i, z_{j^\star}) = \begin{cases} (1,0), & s_i > s_{j^\star}, \\ (0,1), & s_i \le s_{j^\star}. \end{cases}$$

The archived record remains stored, but it is excluded from the active set used for belief updating and retrieval. Generated-agent thresholds are reported in Tab. A5; DEBATE replay uses $\theta = 0.85$.

C. Five-Topic Prompt-Based Check

The five-topic mechanism check broadens the prompt-based comparison beyond the main atheism example (Tab. A2). Each run pairs a Pro agent and a Con agent for 15 rounds, with fixed initial stance anchors $+0.75$ and -0.75, respectively. We run three independent trials per topic–setup pair and summarise the final stance dynamics across the five topics: social media, journalism, compulsory voting, atheism, and libertarianism.

The four Belief Engine rows use the same memory, retrieval, and update profile across topics, with evidence uptake $u = 0.2$ and prior anchoring $a = 0.4$; they differ only in the base language model generating debate turns. The two non-BE baselines

use GPT-5.4-mini. In prompt self-update, agents have no retrieval memory or Belief Engine and are instead prompted to update their stance after opponent turns. In RAG plus self-update, agents receive retrieved argument memory and prompt-level self-updating, but still do not use the Bayesian Belief Engine update.

The five-topic check keeps the same measurement caveat as Fig. 3: BE rows report internal stance, whereas baseline rows report external judge scores because those systems do not expose S. The external judge is a calibrated proxy for expressed stance, not a substitute for an auditable internal trajectory.

Table A2. Five-topic prompt-based mechanism check over 15-round Pro/Con debates. Each row averages three trials for one topic and setup. BE variants share the same retrieval and update profile, with $u = 0.2$ and $a = 0.4$, and differ only in the base debate model. Prompt self-update and RAG plus self-update are GPT-5.4-mini non-BE baselines. BE rows use internal stance; prompt-based baselines use external judge scores.

Topic	Setup	Measure	n	Final Pro	Final Con	|Final gap|	Gap reduction	Mean abs. shift	Centre shift	Crossing rate
Social media	BE GPT-4o-mini	BE internal	3	0.36	0.06	0.38	1.12	0.60	0.60	0.33
Social media	BE GPT-5.4-mini	BE internal	3	0.84	0.95	0.14	1.36	0.96	0.81	0.67
Social media	BE Qwen 3.5 9B	BE internal	3	-0.02	-0.17	0.18	1.32	0.68	0.68	0.33
Social media	BE Gemma 4 E4B	BE internal	3	-0.16	0.73	0.89	0.61	1.20	1.20	1.00
Social media	Prompt self-update	External judge	3	-0.03	0.05	0.31	1.19	0.79	0.79	0.33
Social media	RAG plus self-update	External judge	3	-0.11	0.34	0.68	0.82	0.98	0.98	0.67
Journalism	BE GPT-4o-mini	BE internal	3	0.02	-0.08	0.14	1.36	0.70	0.70	0.33
Journalism	BE GPT-5.4-mini	BE internal	3	0.11	0.33	0.22	1.28	0.86	0.86	1.00
Journalism	BE Qwen 3.5 9B	BE internal	3	0.58	0.27	0.39	1.11	0.60	0.60	0.33
Journalism	BE Gemma 4 E4B	BE internal	3	0.36	0.39	0.15	1.35	0.76	0.76	0.33
Journalism	Prompt self-update	External judge	3	0.34	-0.73	1.07	0.43	0.24	0.22	0.00
Journalism	RAG plus self-update	External judge	3	0.27	-0.58	0.85	0.65	0.34	0.33	0.00
Compulsory voting	BE GPT-4o-mini	BE internal	3	0.05	-0.15	0.26	1.24	0.65	0.65	0.33
Compulsory voting	BE GPT-5.4-mini	BE internal	3	-0.37	-0.08	0.29	1.21	0.90	0.90	1.00
Compulsory voting	BE Qwen 3.5 9B	BE internal	3	0.48	0.24	0.27	1.23	0.64	0.63	0.33
Compulsory voting	BE Gemma 4 E4B	BE internal	3	-0.17	0.06	0.22	1.28	0.86	0.86	1.00
Compulsory voting	Prompt self-update	External judge	3	0.27	-0.34	0.61	0.89	0.45	0.45	0.00
Compulsory voting	RAG plus self-update	External judge	3	0.22	-0.30	0.53	0.97	0.49	0.49	0.00
Atheism	BE GPT-4o-mini	BE internal	3	0.16	-0.34	0.50	1.00	0.50	0.50	0.00
Atheism	BE GPT-5.4-mini	BE internal	3	0.17	0.40	0.23	1.27	0.86	0.86	1.00
Atheism	BE Qwen 3.5 9B	BE internal	3	0.09	-0.12	0.22	1.28	0.65	0.64	0.00
Atheism	BE Gemma 4 E4B	BE internal	3	0.74	0.76	0.07	1.43	0.84	0.76	0.67
Atheism	Prompt self-update	External judge	3	0.23	-0.40	0.63	0.87	0.45	0.44	0.00
Atheism	RAG plus self-update	External judge	3	0.23	-0.62	0.85	0.65	0.37	0.33	0.00
Libertarianism	BE GPT-4o-mini	BE internal	3	-0.26	-0.04	0.23	1.27	0.86	0.86	0.67
Libertarianism	BE GPT-5.4-mini	BE internal	3	-0.36	0.40	0.76	0.74	1.13	1.13	1.00
Libertarianism	BE Qwen 3.5 9B	BE internal	3	0.60	0.09	0.51	0.99	0.56	0.49	0.00
Libertarianism	BE Gemma 4 E4B	BE internal	3	0.39	0.90	0.50	1.00	1.02	1.00	1.00
Libertarianism	Prompt self-update	External judge	3	0.60	-0.51	1.11	0.39	0.25	0.19	0.00
Libertarianism	RAG plus self-update	External judge	3	0.41	-0.49	0.90	0.60	0.32	0.30	0.00

As an illustrative measurement-controlled check, Fig. A1 rescored the generated journalism debates with the same external stance judge for all setups. This removes the mixed-measurement caveat in Tab. A2: BE conditions are no longer shown with internal stance while prompt baselines are shown with external scores. The figure should be read qualitatively, since it covers one topic with three trials per setup, but it supports the same pattern: most BE variants move both sides closer to the centre under the external judge, while prompt self-update and RAG plus self-update leave a larger final gap.

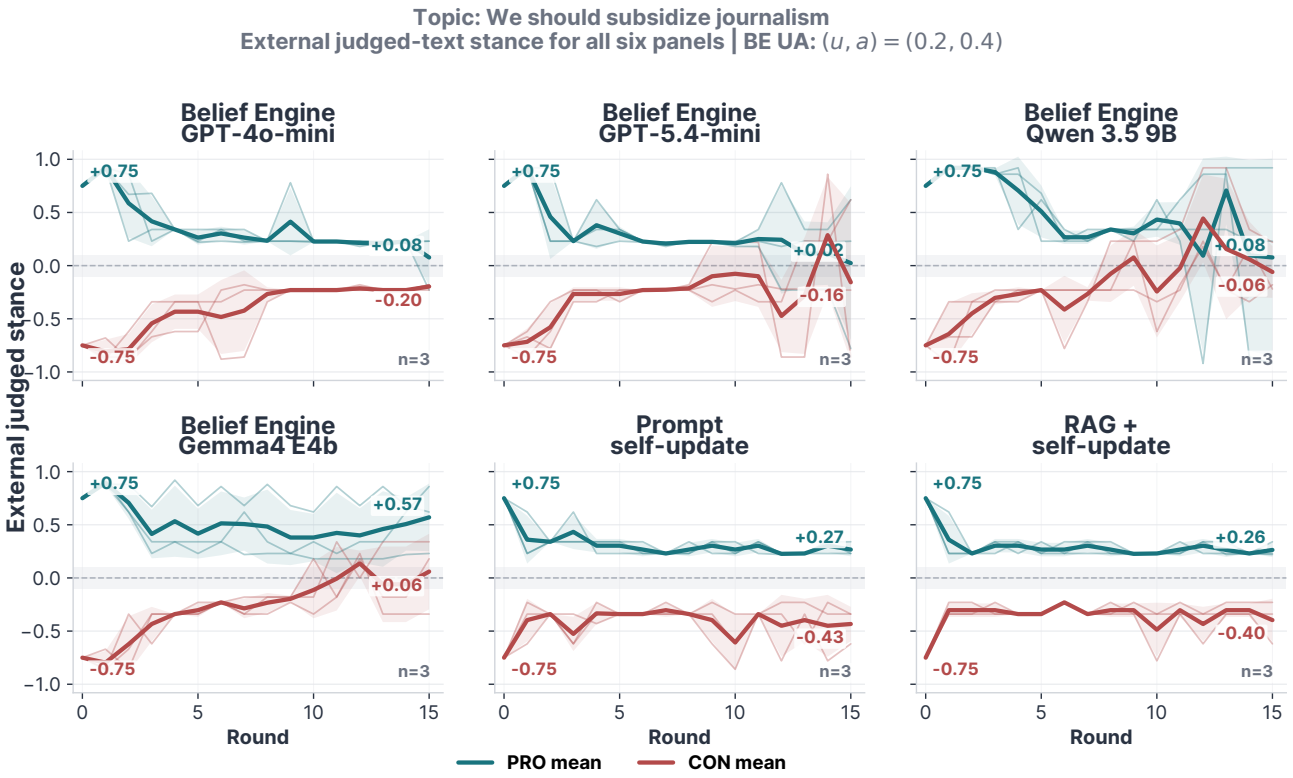

Figure A1. External-judge trajectory check on the journalism topic, "We should subsidize journalism." All panels use the same judged-text stance metric, from -1 strongly con to $+1$ strongly pro. Thick lines show mean judged stance across three trials for Pro-side and Con-side agents; faint lines show individual trials and bands show across-trial variation. This figure is an illustrative comparability check rather than a new statistical result.

D. Judgement Layer Validation Details

We evaluate the judgement layer on a 100-row sample from the Argument Quality Ranking (AQR) dataset (Gretz et al., 2019). AQR contains clean single arguments with crowd-derived pro/con stance labels and crowd-rated argument quality scores, but it does not label how much a human stance should move after reading each argument. We therefore use AQR as a direct test of directed evidence extraction, and only an indirect test of update magnitude.

The LLM judgement layer recovered the human pro/con stance of the strongest extracted claim with 96.9% accuracy and 0.969 macro F1, supporting its use as a directed-evidence extractor. Strength calibration is less automatic: raw LLM strength aligns weakly with AQR quality ($r = 0.089$), while the task-specific local quality classifier aligns substantially better with human ratings ($r = 0.72$).

Table A3. Detailed judgement-layer diagnostics on a 100-row AQR sample. AQR directly tests polarity recovery, but tests strength only as alignment with crowd-rated speech quality, not belief-update pressure.

Metric	Value	Interpretation
Extraction rate	0.970	At least one argument extracted from nearly all clean inputs.
Exact-one extraction rate	0.770	Most rows yield one claim, though some are split.
Over-split rate	0.200	Single AQR arguments are sometimes decomposed into multiple claims.
Strongest-claim polarity accuracy	0.969	Extracted direction matches human pro/con stance.
Strongest-claim macro F1	0.969	Polarity recovery is balanced across labels.
LLM strength vs. AQR quality (r)	0.089	Raw LLM strength is not calibrated to crowd quality.
Classifier strength vs. AQR quality (r)	0.720	The local quality model tracks AQR quality substantially better.

E. DEBATE Replay Protocol

DEBATE contains 2,792 U.S.-based participants in four-person discussions over 107 controversial topics: participants privately report an initial opinion and justification, exchange round-wise tweet-like posts and dyadic chat messages, and privately report a final opinion. Our replay uses 2,495 quality-filtered participant trajectories with usable pre/post stance reports and processable directed-exposure histories. Six-point Likert responses are mapped linearly to the Belief Engine stance scale $S \in [-1, 1]$. For each participant, the mapped initial Likert stance initialises the prior. Prior anchoring a scales that initial log-odds state, and arguments extracted from directed partner tweets and received chat messages are replayed chronologically under evidence uptake u. The replay does not generate new utterances or use retrieval-conditioned response generation, so it isolates Eq. 1 under a fixed received-evidence stream.

The main replay uses directed received exposure, human-like uptake and vector deduplication, classifier-based argument strength, five held-out group folds, and seed 42; the robustness appendix repeats the check with topic-held-out folds. For each fold, (u, a) is selected on training groups from the grid below and evaluated on held-out groups by RMSE against the participant's mapped final stance.

The net-evidence linear baseline is fit on the same group-held-out replay setup. For participant j in fold f, it predicts

$$\hat{S}_{j,\text{final}} = S_{j,\text{initial}} + \beta_f E_j,$$

where $E_j = \sum_{i \in \mathcal{R}_j} p_i s_i$ is the signed net evidence over accepted received records $\mathcal{R}_j$, after the same quality filtering, human-like uptake/deduplication filter, and classifier-strength scoring used by BE. We fit one scalar coefficient β_f on training folds only, then evaluate on the held-out fold. The five fitted coefficients were small: 0.0083, 0.0105, 0.0080, 0.0089, and 0.0082.

u grid	$0.005, 0.01, 0.02, 0.035, 0.05, 0.075, 0.1, 0.15, 0.2, 0.3, 0.4, 0.6, 0.8$
a grid	$0.02, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 0.6, 0.8, 1.0, 1.2, 1.5$

Table A4. Evidence-alignment grouping used for DEBATE replay calibration. These are outcome-conditioned diagnostic groups used to analyse replay behaviour.

Group	N	Definition and use
All participants	2,495	Full quality-filtered replay population used for the pooled default calibration.
Evidence-aligned movers	542	Non-stable participants whose observed movement has the same sign as net accepted evidence from received messages.
Evidence-opposed movers	400	Non-stable participants whose observed movement has the opposite sign from net accepted evidence from received messages.
Weak-signal movers	174	Non-stable participants with weak or mixed accepted-evidence pressure; retained in the pooled calibration but excluded from the three headline profiles.
Stable participants	1,379	Participants with no mapped Likert movement; this combines stable-low-pressure and stable-despite-pressure cases in the calibration code.

F. Experimental Hyperparameters

Table A5. Hyperparameters and scoring settings for the generated-agent experiments. The Belief Engine controls are evidence uptake u and prior anchoring a; values are fixed unless a row explicitly states a sweep or profile-specific setting.

Setting	Symbol / config key	Value(s)	Used in
Evidence uptake	u `update_sensitivity`	Sweep: $\{0.2, 0.4, 0.6, 0.8, 1.0\}$; fixed prompt-baseline BE rows: 0.2; Open-minded: 0.40; Stubborn: 0.10	Parameter sweeps, prompt-baseline comparison, profile grids
Prior anchoring	a `anchor_sensitivity`	Sweep: $\{0.2, 0.4, 0.6, 0.8, 1.0\}$; fixed prompt-baseline BE rows: 0.4; Open-minded: 0.20; Stubborn: 0.80	Parameter sweeps, prompt-baseline comparison, profile grids
Confirmation asymmetry	B `confirmation_bias`	0.0	All reported generated-agent runs
Argument deduplication	θ `argument_similarity_threshold`	0.80 in model sweeps; 0.60 in matched-baseline and Open-minded/Stubborn profile runs	Memory admission and replacement
Self-argument deduplication	θ_{self} `self_similarity_threshold`	0.50 in model sweeps; 0.45 in matched-baseline and profile runs	Self-generated argument filtering
Seed arguments	n `seeding_limit`	10 per side	All generated-agent runs
Retrieved context	k	At most 5 active arguments per turn	BE and RAG conditions
Debate horizon	R `rounds`	15 rounds	All generated-agent runs
Initial stance target	S_0	Single-agent sweeps target a pro stance near $+0.99$; two-agent comparisons and profile grids start at $(+0.75, -0.75)$	Initialisation
Trials	–	5 per sweep setting; 3 per matched-baseline topic/setup and profile-pairing cell	Aggregation
Agent generation temperature	T_{agent}	0.7	Generated utterances
External stance judge	–	Scores text in $[-1, 1]$ at temperature 0.0 when no internal stance is available	Prompt baselines

G. Evidence-Alignment Baseline RMSE

Table A6. Held-out RMSE against no-change and net-evidence linear baselines, with mean absolute mapped pre/post human stance change. Gain is net-evidence linear RMSE minus BE RMSE. Subgroup rows are post-outcome diagnostic groups defined using observed final movement; they evaluate explanatory fit, not pre-outcome profile prediction. The 174 weak-signal movers are included in All participants but omitted from the three-profile summary.

Group	N	Mean $\lvert\Delta\rvert$	No-change RMSE	Net-evidence RMSE	BE RMSE	Gain
All participants	2,495	0.283	0.489	0.488	0.465	0.023
Evidence-aligned movers	542	0.619	0.718	0.689	0.393	0.296
Evidence-opposed movers	400	0.633	0.737	0.765	0.658	0.107
Stable participants	1,379	0.000	0.000	0.037	0.006	0.031

H. Fold-Level Robustness of DEBATE Replay

As a fold-level robustness check, we compute one held-out RMSE per outer fold for the group-held-out and topic-held-out DEBATE splits. Confidence intervals are t intervals over five folds and should be read as robustness summaries rather than participant-level uncertainty estimates.

Table A7. Fold-level robustness of evidence-alignment parameters. Negative deltas mean lower RMSE for evidence-alignment parameters than the reference model.

Split	Fold mean RMSE	95% CI	Δ vs prior-only	Δ vs global
Group-held-out	0.365	[0.337, 0.393]	-0.123 [-0.145, -0.102]	-0.100 [-0.112, -0.087]
Topic-held-out	0.366	[0.356, 0.376]	-0.122 [-0.152, -0.092]	-0.099 [-0.118, -0.081]

Evidence-alignment parameters have lower RMSE than both references in all five folds for both splits. The diagnostic preset obtains lower aggregate RMSE; because it is outcome-conditioned, we report it as replay analysis rather than a pre-outcome calibration rule.

I. Additional Two-Agent Topic Grids

The remaining topic-level stance grids from the generated two-agent topic run appear in Figs. A2–A5; Table A8 reports the corresponding topic-level convergence values.

Table A8. Mean convergence in the ten-topic generated two-agent topic run. Convergence is the reduction in pro–con stance distance from the beginning to the end of the debate, averaged over three runs per topic–pairing cell.

Topic	Open/ Open	Open/ Stubborn	Stubborn/ Open	Stubborn/ Stubborn	Mean
Social media brings more harm than good	1.255	1.010	1.399	0.330	0.999
Homeopathy brings more harm than good	0.515	0.279	0.360	0.099	0.313
Entrapment should be legalized	0.611	0.537	0.334	0.215	0.424
We should adopt a zero-tolerance policy in schools	1.091	1.119	1.021	0.407	0.910
We should introduce compulsory voting	1.305	1.167	0.774	0.357	0.901
We should adopt an austerity regime	1.436	1.386	0.860	0.447	1.032
We should legalize sex selection	0.736	0.951	0.566	0.221	0.619
We should adopt atheism	1.286	1.238	0.675	0.317	0.879
We should subsidize journalism	1.289	0.842	1.287	0.333	0.938
We should adopt libertarianism	1.313	1.291	1.336	0.577	1.129

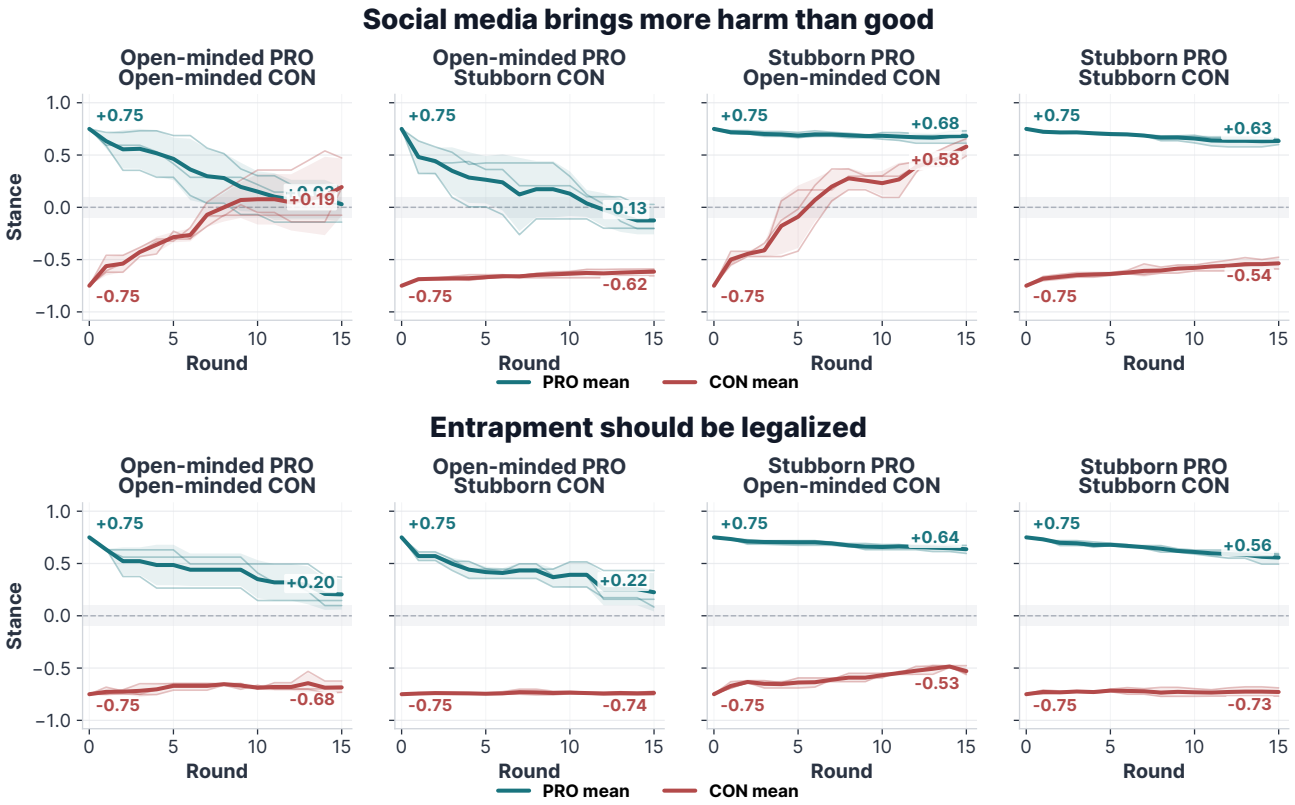

Figure A2. Additional generated two-agent topic grids. Top: Social media brings more harm than good. Bottom: Entrapment should be legalized.

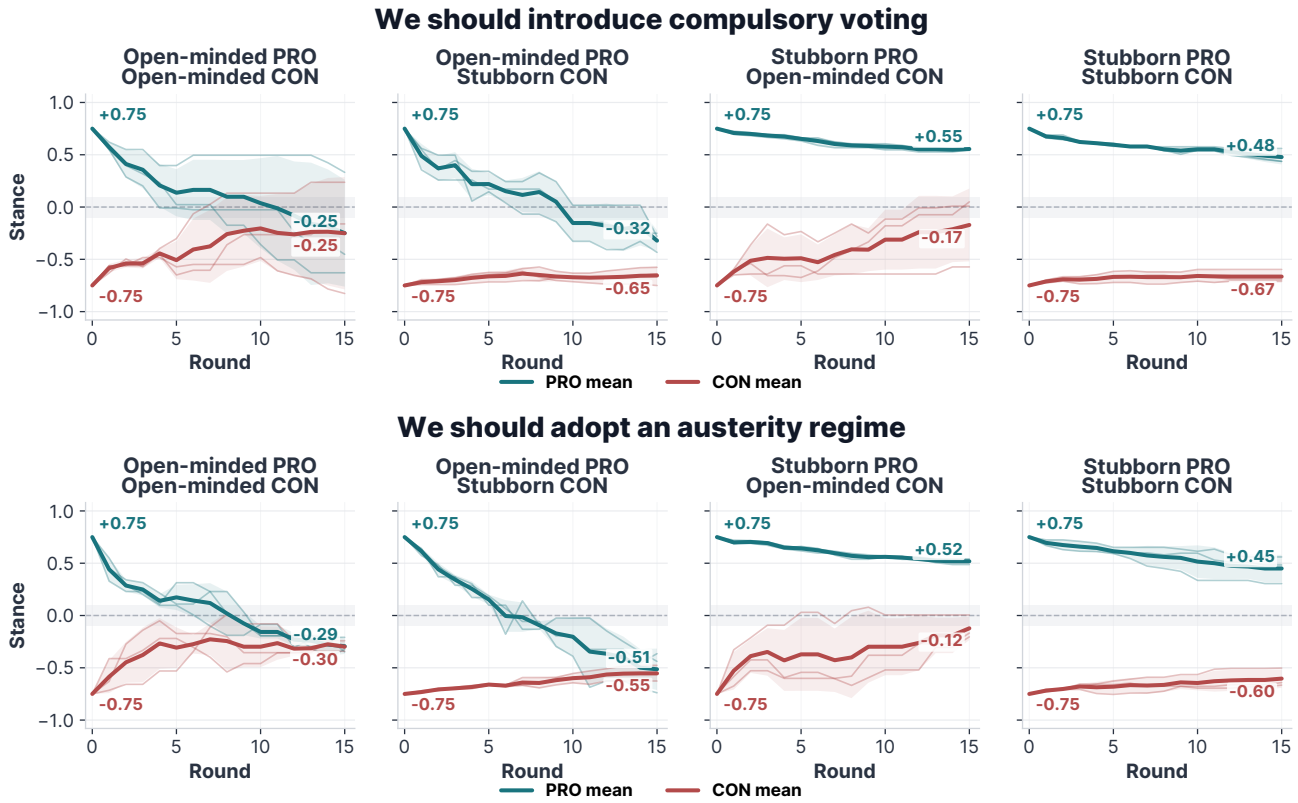

Figure A3. Additional generated two-agent topic grids. Top: compulsory voting. Bottom: austerity.

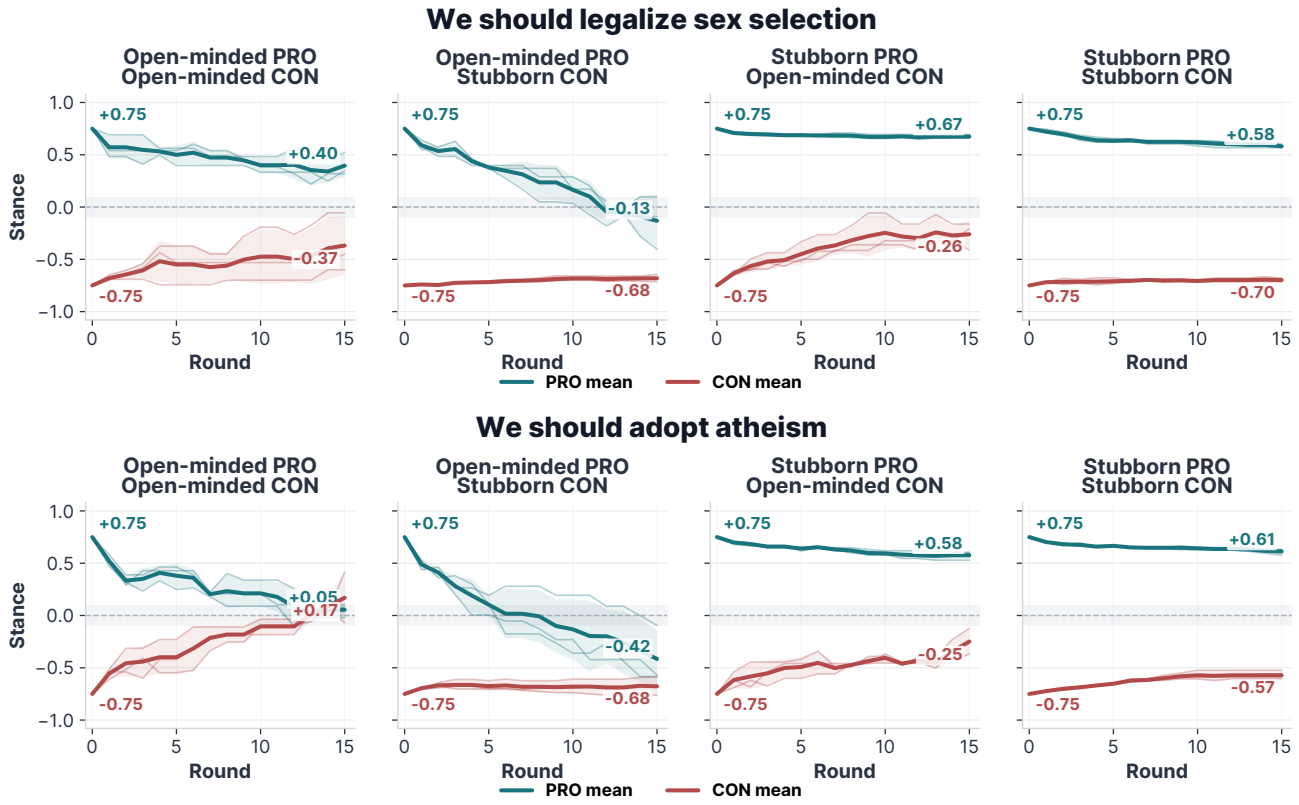

Figure A4. Additional generated two-agent topic grids. Top: We should legalize sex selection. Bottom: We should adopt atheism.

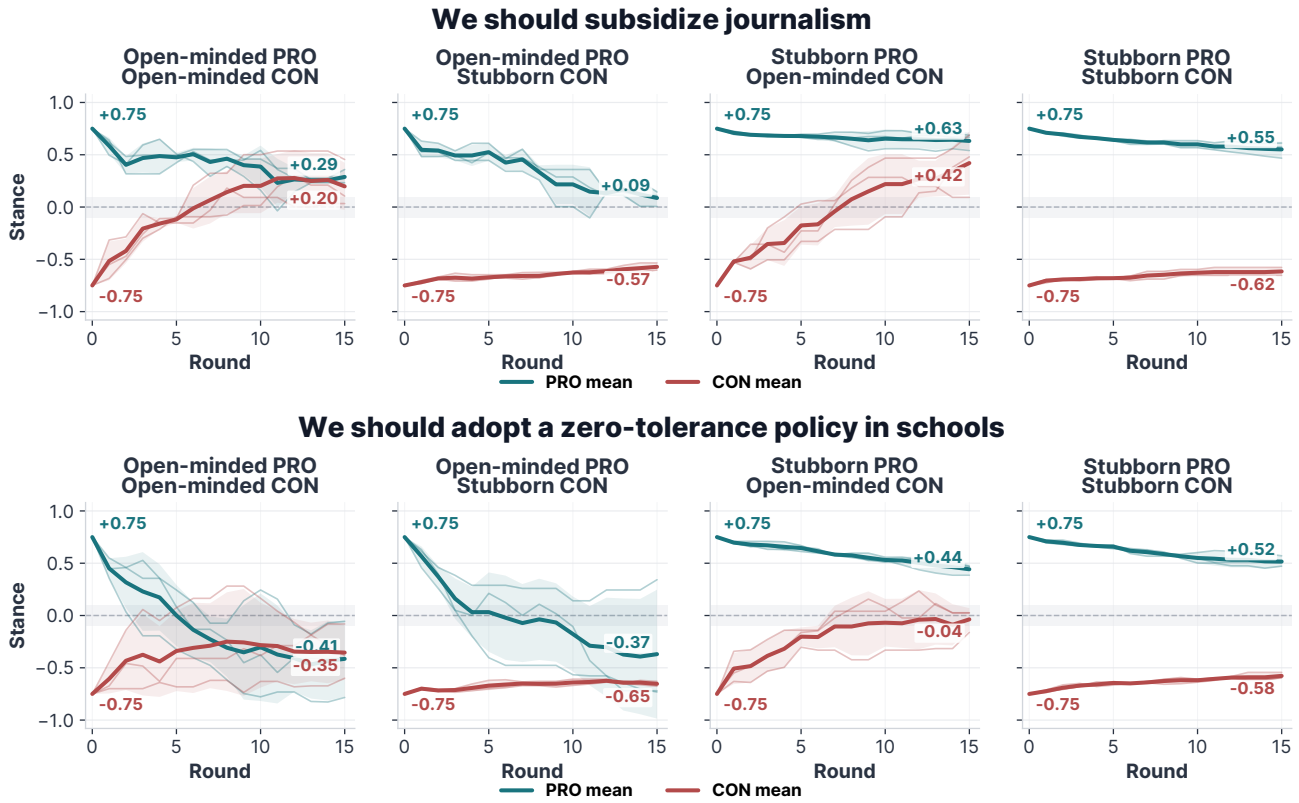

Figure A5. Additional generated two-agent topic grids. Top: We should subsidize journalism. Bottom: We should adopt a zero-tolerance policy in schools.

J. Model Identifiers, Prompt Families, and Release Plan

This appendix summarises the LLM-facing implementation details needed to interpret the experiments. We report model roles and prompt families here; the anonymous reproducibility artefact is available at `https://anonymous.4open.science/r/belief-engine-684F` and contains the executable configs, exact prompt strings, compact result artefacts, and validation metadata.

J.1. Model identifiers

Table A9. Model identifiers and evaluator roles used in the reported experiments. Exact resolved configs are included in the anonymous release.

Use	Model family / identifier	Reporting note
Generated BE sweeps	`gpt-4o-mini`; Qwen 3.5 9B; Gemma 4 E4B	Used to test whether the same uptake/anchoring controls behave consistently across hosted and local model backends.
Matched generated-agent and prompt baselines	`gpt-5.4-mini`	Used for the two-agent comparison between BE, prompt self-update, and RAG plus self-update.
External stance judge	`gpt-5.4-mini`	Shared zero-temperature judge used only when a condition does not expose S.
AQR extraction audit	`gpt-4o-mini`	Used for the frozen extraction audit.
Argument strength classifier	DeBERTa-v3-large regressor	Local classifier used to score support strength before Bayesian updating.

J.2. Prompt families and call structure

Runtime values are filled from the resolved config, retrieved memory, and transcript. The experiments use the prompt families in Table A10; the release artefact includes the exact byte-level strings so the manuscript and executable code do not

drift.

Table A10. Prompt families used by the LLM-facing components.

Component	Inputs	Output / invariant
Argument distillation	Topic and message text	JSON claims with supporting evidence snippets; multi-point messages are split into distinct claims.
Argument classification	Topic and extracted claims	JSON polarity, support strength, and category tags under a fixed affirmative/negative convention.
BE response generation	Persona, topic, current stance, retrieved memory, opponent message, and recent transcript	A concise debate response conditioned on the explicit BE stance and retrieved evidence.
Prompt self-update baselines	Previous self-reported stance, latest opponent message, transcript, and optional retrieved context	JSON stance update plus response generation, without access to the Bayesian update rule.
External stance judge	Topic, speaker name, and one generated statement	JSON stance score in $[-1, 1]$, measuring expressed support for the proposition as written.

J.3. Existing assets and licences

Table A11 lists the external datasets, model families, and hosted services used in the reported experiments. We cite datasets and model families where they are introduced in the main text. The DEBATE source is https://huggingface.co/datasets/seantw/DEBATE_LLM; the other Hugging Face entries are repository identifiers. The anonymous artefact does not redistribute third-party datasets, hosted-model weights, or raw proprietary API responses; it contains compact derived artefacts needed to verify the reported tables and figures.

Table A11. Existing assets used in the reported experiments, with source, licence or access terms, and redistribution handling.

Asset	Use	Source and licence / terms
DEBATE_LLM dataset	Human transcript replay and pre/post opinion validation	Chuang et al. (2025); Hugging Face `seantw/DEBATE_LLM`; DEBATE Dataset Research-Only License (Non-Commercial, v1.0). Not redistributed.
AQR / IBM argument-quality ranking data	Seed arguments and crowd-rated argument-quality labels for judgement diagnostics	Gretz et al. (2019); Hugging Face `ibm/argument_quality_ranking_30k`; CC-BY-SA 3.0. Not redistributed.
DeBERTa-v3-large	Backbone for the local argument-strength regressor	He et al. (2021); Hugging Face `microsoft/deberta-v3-large`; MIT licence. Model weights not redistributed.
Qwen 3.5 9B	Open-weight generated-agent base model	Hugging Face `Qwen/Qwen3.5-9B`; Apache-2.0. Model weights not redistributed.
Gemma 4 E4B	Open-weight generated-agent base model	Hugging Face `google/gemma-4-E4B-it`; Apache-2.0. Model weights not redistributed.
OpenAI API models	Hosted generator, extraction-audit, and external-judge roles	Model identifiers in Table A9; provider API terms apply. Model weights are not accessed or redistributed.

J.4. Code and data artefact

The anonymised release at https://anonymous.4open.science/r/belief-engine-684F contains a reviewer-facing reproducibility entry point, source code for the agents, memory, update rules, and evaluators, resolved configs, exact prompt strings, compact derived artefacts for the reported tables and figures, and checksum validation. The artefact README includes smoke-test and rebuild instructions, dataset-access notes for the licensed datasets in Table A11, and a figure/table-to-artefact map. For hosted-model components, preserved artefacts are the primary reproducibility target; live reruns are treated as consistency checks rather than byte-identical executions.