BELIEF ENGINE: BAYESIAN MEMORY FOR CONFIGURABLE OPINION DYNAMICS IN LLM AGENTS

Joshua C. Yang[1] **Damian Dailisan**[1] **Maurice Flechtner**[1,2]
[1]Computational Social Science, ETH Zurich
[2]Center for Democracy Aarau, University of Zurich
joyang@ethz.ch

ABSTRACT

Large Language Model (LLM) agents can debate fluently, but they do not reliably maintain beliefs across long interactions. This makes it difficult to use them for opinion-dynamics studies where trajectories must be stable, interpretable, and reproducible. We introduce the *Belief Engine*, a configurable belief architecture that externalises belief state and updates it from extracted arguments. The engine stores adjudicated evidence in memory and updates a bounded stance score using a simple Bayesian log-odds rule with tunable parameters controlling evidence sensitivity, anchoring, and asymmetric weighting. In controlled two-agent debates across topics, we show that Belief Engine produces stance trajectories that are smoother and more reproducible than LLM-based (Agentic) updating, and that its parameters provide monotonic control over persuadability and resistance. By separating what an agent says from how its beliefs are updated, the framework enables traceable and controllable opinion dynamics in LLM-agent simulations.

1 INTRODUCTION

The rise of Large Language Models (LLMs) holds the promise of transforming social simulation by replacing rigid rule-based agents with entities capable of natural deliberation: reasoning through arguments, synthesising evidence, and expressing nuanced positions. If realised, this would enable computational experiments on opinion formation, democratic deliberation, and collective decision-making at unprecedented scale and realism. Applications range from virtual politicians that maintain consistent policy positions across multiple debates to digital democracy platforms where AI agents represent diverse citizen perspectives in large-scale deliberations (Gudiño et al., 2024). Yet this promise rests on a critical assumption: that LLM agents can maintain stable, coherent beliefs across extended interactions.

They cannot. Recent empirical work reveals that current LLM agents exhibit two fundamental failure modes. First, lacking a persistent internal belief system, they behave as "chameleons," drifting from assigned personas to mirror their immediate context (Choi et al., 2025b) or progressively echoing their conversation partners (Shekkizhar et al., 2025). Without a structured update policy, belief change follows a random walk, a martingale in which deliberation yields no expected convergence toward truth or consensus (Choi et al., 2025a). Second, when agents *do* resist persuasion, they exhibit pathological *stance inertia*, refusing to update even under sustained counter-evidence. Neither failure mode supports genuine deliberation.

This points to a missing link between neural reasoning and formal opinion dynamics. Classical models like Bayesian updating (Chen & Zaman, 2025) assume rational agents tracking truth, while social influence models (Hegselmann & Krause, 2002) describe conformity pressures. Current LLM agents satisfy neither: they are not consistently rational, nor do they follow predictable laws of social influence. They lack the *backbone* of a belief system.

To address this, we introduce a Configurable Belief Architecture that formalises the interaction between memory and reasoning. Our central innovation is separating the generative process from belief maintenance. We employ a dedicated *Belief Engine* that uses Bayesian logic to integrate semantic evidence into a persistent state, transforming fleeting context into durable conviction.

This architecture provides the reliability needed for rigorous social science. By using explicit parameters to control the update function, researchers can model the balance between belief commitment and evidential responsiveness. For instance, high *prior adherence* simulates base-rate neglect, while low *update sensitivity* models stubbornness. Critically, these parameters enable agents to maintain stable positions while remaining appropriately open to compelling counter-evidence, avoiding both excessive malleability and dogmatic rigidity. This design echoes principles from cognitive science: the separation of fast generative reasoning (LLM response generation) from slower reflective updating (Belief Engine) parallels dual-process theories, while our parameter set captures well-documented human deviations from normative rationality. To isolate the Belief Engine's effect, we include a No Belief Engine condition (§3.3) where belief updates are disabled entirely, measuring how unguided LLMs behave under adversarial pressure.

Our contributions are threefold. First, we formalise a belief update mechanism that bridges Bayesian logic with LLM-based reasoning, preventing random belief drift. Second, we demonstrate that externalising belief state prevents identity drift and allows us to trace exactly why an opinion changed. Finally, we show that our hyperparameters (K, B) provide parametric control over agent plasticity, recovering distinct and interpretable cognitive profiles. This framework offers a principled foundation for social simulation, ensuring that agents are stable actors with verifiable beliefs rather than chameleons in a storm of context.

To validate these contributions, we address three research questions: **RQ1:** Does externalising belief state into a structured engine prevent identity drift and enable causal traceability? **RQ2:** Is the choice of update mechanism (Bayesian vs. Agentic vs. No Belief Engine) associated with differences in trajectory stability and reproducibility? **RQ3:** Can explicit parameters recover interpretable cognitive profiles spanning from open-minded to stubbornly entrenched?

2 Related Work

Our work sits at the intersection of three research areas: the growing evidence that LLM agents are *unstable* deliberators, the development of *memory architectures* that give agents persistence, and the rich tradition of *formal opinion dynamics* that provides mathematical guarantees about belief change. We argue that none of these areas alone is sufficient, and that our Configurable Belief Architecture bridges all three.

2.1 Belief Instability in LLM Agents

A consistent finding across recent studies is that LLM agents struggle to maintain stable beliefs. Xu et al. (2024) document failures in resolving knowledge conflicts, Kim et al. (2024) show sensitivity to irrelevant information, and Zhang et al. (2025) identify strong recency bias in belief updating. The DEBATE benchmark (Chuang et al., 2025) reveals that compared to humans, LLM groups exhibit excessive convergence and weak belief inertia, suggesting that agents are far too easily swayed.

This instability goes beyond single-turn reasoning. Shekkizhar et al. (2025) document an "echoing" failure where agents progressively imitate each other, abandoning their assigned roles within a few turns. Choi et al. (2025b) show that even individual agents exhibit identity drift across multi-turn conversations, with personality profiles diverging significantly without explicit memory. Ratnakar & Raghavendra (2025) further demonstrate that search-augmented LLMs exhibit frequent stance reversals, confirming that access to external information does not, by itself, stabilise beliefs. These findings establish the core problem our architecture addresses: LLM agents lack the internal structure needed for stable, long-term opinion dynamics.

2.2 Memory Architectures: Persistence Without Revision

The most common response to agent instability has been to add memory. The seminal Generative Agents framework (Park et al., 2023) introduced a three-component architecture (memory stream, reflection, and planning) that produces human-like behavioural patterns. Subsequent work has refined these ideas across many dimensions. Zhong et al. (2023) introduced episodic memory banks for long-term retention, Xu et al. (2025) developed reflection mechanisms for memory consolidation, Gutiérrez et al. (2024) proposed knowledge graph integration, and several systems explored hierarchical storage

architectures (Kang et al., 2025; Huang et al., 2025; Rezazadeh et al., 2025). Recent surveys (Zhang et al., 2024) provide comprehensive taxonomies of these approaches.

These memory systems are also used to maintain agent personas. The AI PERSONA framework (Wang et al., 2024) enables continuous personas across multi-session dialogue, Reflective Memory Management (Tan et al., 2025) adds periodic memory pruning, and practical systems like OpenClaw (Steinnberger & Community) use filesystem-based logs for long-term context.

However, all of these architectures are mainly designed for *retrieval* (recalling past information) rather than *belief revision* (deciding what to believe in light of new evidence). As Xiong et al. (2025) show, agents with standard memory exhibit "experience-following" behaviour, amplifying errors from noisy memories rather than critically evaluating new evidence. Our work addresses this gap by treating memory not as a static log, but as a mutable belief state governed by a principled update rule.

2.3 Classical Opinion Dynamics: Stability Without Language

At the other end of the spectrum, classical opinion dynamics models offer precisely the mathematical rigour that LLM agents lack. Bayesian models (Olsson, 2013), the DeGroot model (DeGroot, 1974), and bounded confidence frameworks (Deffuant et al., 2000; Hegselmann & Krause, 2002) provide well-understood mechanisms for consensus formation, polarisation, and fragmentation. However, these models operate on scalar belief representations and cannot handle the semantic richness of natural language.

Human Memory and Belief Updating. While humans are not ideal Bayesian agents (Stengård et al., 2022), exhibiting biases like extreme probability warping (Holt & Smith, 2009) and base-rate neglect (Ashinoff et al., 2022), these deviations provide a principled basis for our architecture. Our *update sensitivity* (K) models evidence weighting, while *prior adherence* (K_a) controls the strength of initial beliefs, analogous to how strongly consolidated memories resist updating. Confirmation bias (B) captures asymmetric evidence processing well-documented in human reasoning. By making these parameters explicit and tunable, we can simulate cognitive profiles ranging from normatively rational to exhibiting characteristic human biases, enabling systematic study of how memory and reasoning interact in opinion formation.

2.4 LLM Debate and Deliberation

Multi-agent debate is increasingly explored not only as a reasoning mechanism, but as a model for democratic deliberation: diverse agents exchange arguments to reach informed, collective judgements. Recent work has explored LLM agents in collective decision-making contexts: Yang et al. (2024) reveal systematic biases and alignment patterns with human choices, highlighting the need for structured mechanisms in multi-agent settings. Systems like TruEDebate (Liu et al., 2025) and R-Debater (Li et al., 2026) use adversarial discourse to sharpen arguments and surface stronger positions. However, Choi et al. (2025a) show that most gains attributed to debate actually come from majority-vote aggregation rather than iterative discussion. Crucially, they demonstrate that debate updates form a *martingale*, implying no expected convergence unless update dynamics are deliberately structured. For deliberative applications, where the goal is genuine opinion change through reasoned exchange, this is a significant challenge. Our Belief Engine addresses it by embedding a structured judgement mechanism that selectively amplifies high-quality arguments, enabling the kind of evidence-sensitive belief revision that meaningful deliberation requires.

In summary, existing work reveals a fundamental tension. LLM agents can *speak* about beliefs but cannot *maintain* them; memory architectures provide persistence but not principled revision; classical models offer rigorous update rules but cannot operate over natural language; and debate frameworks lack mechanisms to preferentially reinforce good reasoning. We bridge these domains by coupling LLM-based agents with an external Belief Engine grounded in formal opinion dynamics, enabling configurable, persistent, and traceable belief updating within natural-language discourse.

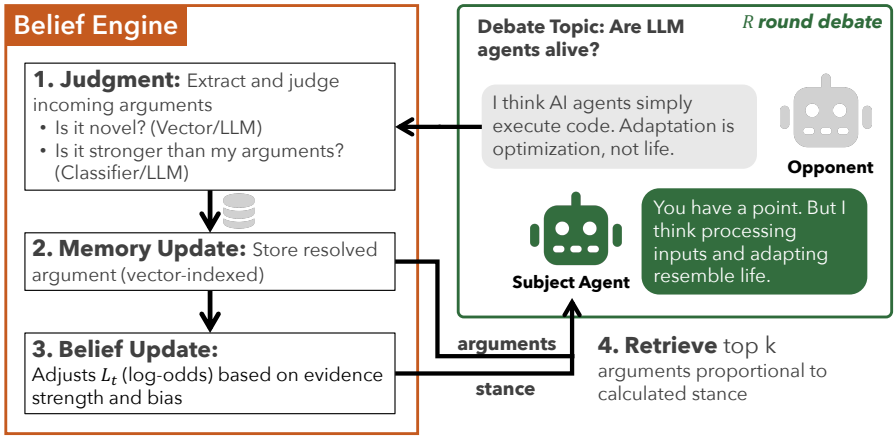

Figure 1: Belief Engine flow in each debate round: (1) Judgement extracts and evaluates incoming arguments, (2) Memory Update stores the resolved evidence, (3) Belief Update updates stance from evidence strength and bias parameters, and (4) Retrieve selects top-k memory arguments *proportional to the calculated stance* (balanced near $S{=}0$, pro-heavy for $S{>}0$, con-heavy for $S{<}0$) to condition the Subject Agent's next response.

3 METHOD

Figure 1 illustrates the overall architecture. During a debate, a subject agent exchanges messages with an opponent around a fixed topic. The Belief Engine intercepts each incoming message, evaluates it, and updates the agent's persistent memory before the agent formulates its reply. By grounding belief revision in explicit parameters rather than implicit LLM tendencies, the architecture provides the kind of principled opinion formation (i.e., deliberative polling or preference aggregation) required by multi-agent social choice applications.

Roadmap. To make the method easy to follow, we mirror the figure's execution order. We first define the *Subject Agent* context (§3.1), then describe the four Belief Engine steps in sequence: (1) Judgement (§3.2.1), (2) Memory Update (§3.2.2), (3) Belief Update (§3.2.3), and (4) Retrieval & Response Conditioning (§3.2.4).

3.1 SUBJECT AGENT

The subject agent is an LLM-based debater and can be instantiated with any compatible base model; in this study, we predominantly use `GPT-4o-mini`. In every condition, the debate starts from $n{=}10$ seeded arguments, and each response is generated from three inputs: (i) current stance instruction, (ii) retrieved memory, and (iii) recent dialogue context.

We evaluate two implementations. CoreBot is our purpose-built controlled simulation agent designed around structured argument records and the Belief Engine, whereas OpenClaw-style adapts the existing OpenClaw framework (Steinnberger & Community) with persistent persona memory using file-based storage and embedding retrieval. Both agents share the same debate protocol and Belief Engine settings (K, K_a, B); thus, the primary architectural difference is the memory layer (detailed in Step 2), not the update rule.

3.2 BELIEF ENGINE

The Belief Engine decouples belief maintenance from generative reasoning so that how an agent changes its mind can be systematically controlled. For each incoming message, the engine follows the same four-step loop as Figure 1: *Step 1 Judgement* (extract and adjudicate claims), *Step 2 Memory Update* (store accepted evidence and credence flags), *Step 3 Belief Update* (recompute stance S), and *Step 4 Retrieval & Response Conditioning* (retrieve memory conditioned on the updated stance and map S to behavioural instructions for generation).

We implement two interchangeable update engines to test whether closed-form mathematical updates or semantic LLM-based reasoning better supports stable deliberation. Bayesian updates stance deterministically in log-odds space using argument polarity, support strength, and parameters (K, K_a, B). It is fully reproducible for a fixed argument sequence and provides direct parameter-level control over update magnitude and asymmetry. Agentic replaces the closed-form update with an LLM call (`GPT-4o-mini`, T=0.1 in our experiments) that receives the same scalar inputs plus currently active memory context, and returns a bounded stance score in $[-1, 1]$ with a structured rationale. In practice, this enables context-sensitive, semantic updates that may capture argument interactions beyond what the closed-form model represents, but introduces temperature-induced variability and lower interpretability compared to the Bayesian engine.

Each argument in memory carries a binary `credence_relevant` flag ($c \in \{0, 1\}$), indicating whether it actively influences the agent's stance, and only c=1 arguments contribute to the continuous stance $S \in [-1, 1]$. This argument-level flag is distinct from the agent's overall stance, which is a continuous score computed over all active arguments. This gating mechanism, together with parametrised updates, forms the central intervention for escaping the martingale dynamics identified by Choi et al. (2025a).

3.2.1 Step 1: Judgement Layer

The judgement layer governs how the incoming message is transformed into structured evidence. Following Xiong et al. (2025), we treat this as a first-class architectural choice. Processing occurs in two stages. First, *Extraction*: An independent LLM module (`GPT-4o-mini`) decomposes messages into `ArgumentRecord` objects, identifying claims, polarity, and support strength $s \in [0, 1]$ (via a quality classifier trained on Gretz et al. (2019)). Second, *Conflict Resolution*: We compare new arguments against memory using cosine similarity ($\theta_{\text{opponent}} = 0.85$). If a match is found, we resolve the conflict by keeping only the stronger argument active ($c = 1$) and archiving the weaker version ($c = 0$), ensuring the agent always holds the best available evidence for each unique claim.

3.2.2 Step 2: Memory Layer

The memory layer is the persistent evidence store used between Judgement and Belief Updating (Figure 1). After conflict resolution, accepted evidence is written to memory, while superseded evidence is retained for traceability. This selective consolidation, where only credence-relevant arguments actively influence belief, mirrors how human memory doesn't retain all experiences equally but filters and strengthens representations based on perceived relevance and evidential value. This mechanism stabilises multi-turn behaviour and reduces identity drift (Choi et al., 2025b).

This is also the main architectural difference between CoreBot and OpenClaw-style agents (§3.1). CoreBot stores memory as structured `ArgumentRecord`s (claim, polarity, support strength, timestamp, `credence_relevant`), supports explicit record-level replacement/archiving during conflict resolution, and retrieves by stance-proportional pro/con sampling. OpenClaw-style agents store memory as persistent markdown text plus embeddings, retrieve semantically similar chunks via vector search, and inject them as unstructured context. In short, CoreBot prioritises transparent belief bookkeeping and controllable retrieval; OpenClaw prioritises persistent cross-session persona continuity.

3.2.3 Step 3: Belief Updating

Three parameters control belief processing: *Update sensitivity* (K) determines how strongly new evidence shifts belief; *Confirmation bias* (B) controls whether the agent discounts opposing evidence; and *Anchor sensitivity* (K_a) sets the strength of the initial prior. Crucially, we vary K_a relative to K (with $B = 0$) to isolate *anchoring* (prior strength) from *bias* (asymmetric processing), testing whether strong priors alone sustain beliefs under pressure.

Formally, opinion formation is modelled as additive evidence accumulation in log-odds space. Let L_t denote the log-odds, initialised to $L_0 = 0$. Each seed argument is processed before the debate begins, shifting L away from zero; with n=10 affirmative seeds of typical strength $s \approx 0.7$ and $K_a = K = 0.4$, this establishes the initial prior S_0. During the debate, an accepted argument with

polarity $p \in \{-1, +1\}$ and support strength s induces the update

$$L_t = L_{t-1} + p \cdot \ln(1 + s \cdot K \cdot \beta), \tag{1}$$

where $\beta = 1 + B$ when the argument polarity matches the current leaning and $\beta = \max(0, 1 - B)$ otherwise. Seed arguments bypass confirmation bias ($\beta = 1$) and use K_a in place of K.

The observable stance $S \in [-1, 1]$ is recovered as $S = 2\sigma(L) - 1$, where σ is the logistic sigmoid. Evidence therefore accumulates linearly in log-odds space, while the logistic transformation ensures diminishing returns near certainty.

3.2.4 Step 4: Retrieval and Prompt Conditioning

After updating stance, the agent retrieves a bounded memory set to condition generation. In CoreBot, retrieval is *proportional* to stance $S \in [-1, 1]$, with target pro share $(1 + S)/2$ and con share $(1 - S)/2$ before sampling records; hence $S \approx 0$ yields balanced context, $S > 0$ yields pro-heavy context, and $S < 0$ yields con-heavy context. Under a social choice lens, this is an *endogenous agenda-setting* rule: the agenda (which arguments become salient) is determined by the agent's own current state rather than an external moderator or fixed quota. For instance, at $S = +0.6$, the agent retrieves 80% pro-arguments and 20% con-arguments, creating attention asymmetry that reinforces the current belief. As $|S|$ grows, attention mass shifts further toward one side, creating a measurable belief–attention feedback loop (path dependence). In OpenClaw-style agents, retrieval is semantic vector search over persistent text memory, and in both architectures the updated S is mapped to a short stance-intensity prompt (10 bins) and combined with retrieved evidence to condition the response.

3.3 Experimental Setup

We source arguments from the *Argument Quality Ranking* dataset (Gretz et al., 2019), comprising approximately 30,000 crowd-annotated arguments spanning diverse debate topics. Each record contains a topic expressed as a full propositional statement, an argumentative claim, polarity normalised to the canonical affirmative/negative vocabulary, and a pre-annotated support strength score $s \in [0, 1]$.

Protocol. Each trial instantiates a symmetrical two-agent debate (ProAgent vs. ConAgent) over 15 rounds. We evaluate 10 topics spanning political, social, and ethical domains, running 10 trials per topic per condition. Both agents are seeded with n=10 arguments supporting opposite polarities, yielding initial stances $S_{\text{Pro}} \approx +0.99$ and $S_{\text{Con}} \approx -0.99$. Each round proceeds as: (1) both agents generate responses from retrieved memory (k=5 items, T=0.7) conditioned on current stance, (2) utterances are processed to extract arguments, (3) extracted arguments (labelled `self`/`opponent`) update both memory states, (4) stances are re-recorded. Only the last 4 messages condition the next response to limit transcript inertia.

Design. The experiment is a 2×3 factorial: architecture (CoreBot vs. OpenClaw) $\times$ belief engine (Bayesian vs. Agentic vs. No Belief Engine), yielding 6 conditions with 10 trials each (60 runs total). In the No Belief Engine condition, belief updates are disabled entirely and agents receive neutral instructions instead of stance-aligned prompts, isolating the effect of the Belief Engine by measuring how unguided LLMs behave under adversarial pressure. An external LLM judge scores each response in $[-1, 1]$ to provide stance assessments independent of internal state. All hyperparameters are in Table 1 (Appendix).

4 Results

We report three results. First, the internal stance variable S reliably translates into observable behaviour as judged externally (Figure 2). Second, the parameters K and K_a provide monotonic, interpretable control over persuadability and anchoring (Figure 3). Third, trajectory stability is driven primarily by the belief engine (Bayesian vs. Agentic vs. No Belief Engine), not by the memory architecture (Figure 4; Table 2, Appendix).

4.1 Internal–External Stance Alignment

To verify that our internal belief state actually controls what an agent says, we sweep $S \in [-1, 1]$ and score each response with an independent LLM judge. Internal and external stance are tightly aligned for both CoreBot ($r = 0.967$, $p < 0.001$) and OpenClaw ($r = 0.950$, $p < 0.001$) (Figure 2). The fitted slope (≈ 0.86) suggests mild compression at the extremes: strongly polarised internal states are expressed in language that is slightly more moderate.

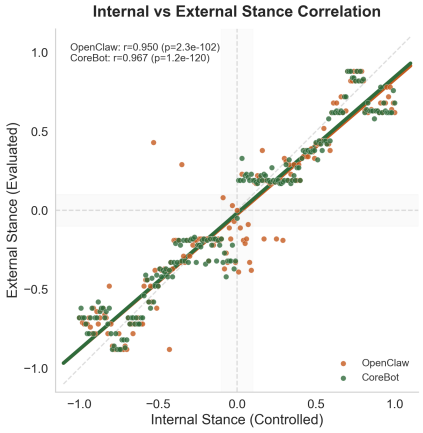

Figure 2: Internal–External Stance Alignment. Internal stance is swept from -1.00 to 1.00 on the x-axis; the external LLM judge score is on the y-axis. Points are trials (OpenClaw: orange, CoreBot: green), and lines are per-agent linear fits.

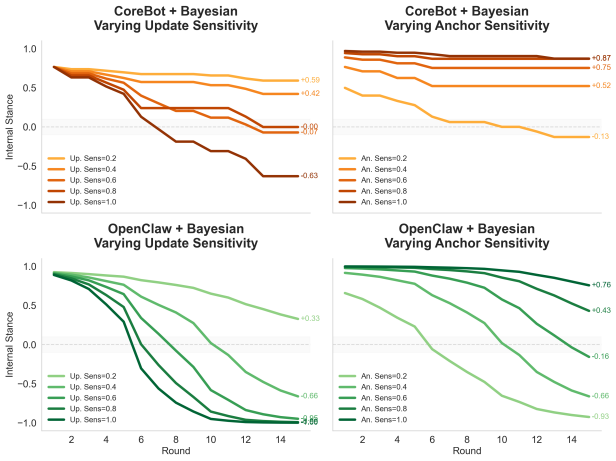

Figure 3: Parametric Control of Belief Dynamics. Increasing update sensitivity K (left) makes agents more persuadable (negative correlation with final stance), while increasing anchor sensitivity K_a (right) increases resistance (positive correlation). Top: CoreBot; Bottom: OpenClaw. Darker curves indicate larger parameter values.

4.2 Parametric Control of Belief Dynamics

We next test whether the control parameters behave as intended. Sweeping *update sensitivity* (K) and *anchor sensitivity* (K_a) across five levels ($\{0.2, \ldots, 1.0\}$) yields clean, monotonic effects on the final stance (Figure 3). Higher K makes agents more persuadable (lower final stance under sustained opposition), while higher K_a makes the seeded prior more persistent. Pearson correlations confirm this for both CoreBot (K : $r = -0.88$, $p < 0.001$; K_a : $r = +0.88$, $p < 0.001$) and OpenClaw (K : $r = -0.83$, $p < 0.001$; K_a : $r = +0.99$, $p < 0.001$).

The architectures differ mainly in how they respond under extreme sensitivity. At $K \geq 0.6$, OpenClaw undergoes complete stance reversal (from $+0.37$ to -1.0), whereas CoreBot drifts only to near-neutrality ($+0.01$). This is consistent with CoreBot's proportional retrieval continuing to surface some prior supporting arguments alongside counter-evidence, acting as a stabilising buffer. At maximum anchor sensitivity ($K_a = 1.0$), both architectures converge to statistically indistinguishable endpoints (CoreBot: $+0.84$, OpenClaw: $+0.78$; $t = 1.40$, $p = 0.23$), suggesting that sufficiently strong priors dominate belief dynamics regardless of architecture.

4.3 Belief Engine Comparison and Trajectory Stability

Finally, we compare belief engines in terms of trajectory *smoothness*, since interpretability and reproducibility depend on whether belief changes unfold steadily or as sharp jumps. Two agents with opposing initial stances ($S_0 = +0.99$ vs. -0.99) debate over 15 rounds in a 2×3 design (architecture: CoreBot vs. OpenClaw $\times$ belief engine: Bayesian vs. Agentic vs. No Belief Engine; n=10 trials each).

Figure 4 reveals three regimes. With No Belief Engine, agents barely change stance despite 15 rounds of opposing arguments, suggesting that default LLM behaviour can look deliberative while remaining

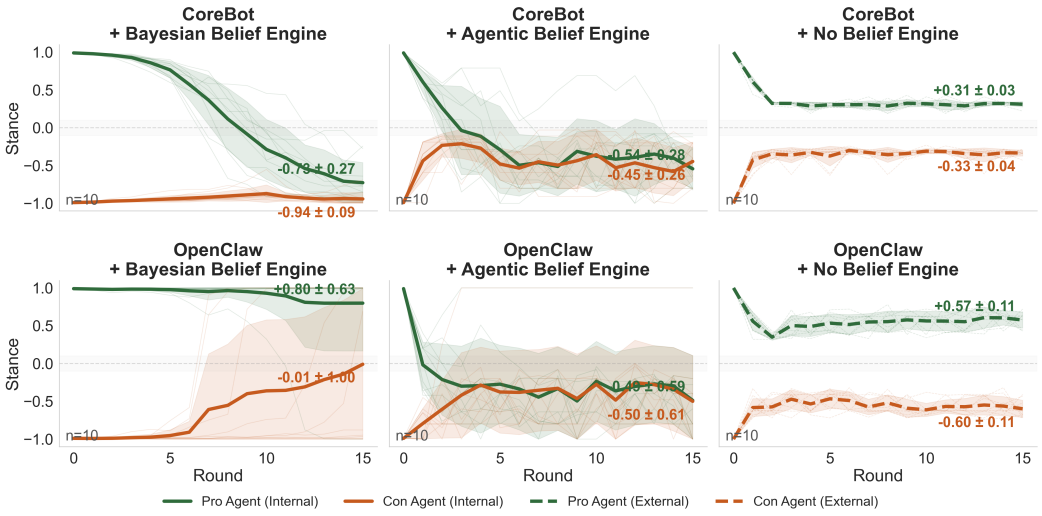

Figure 4: Stance trajectories during symmetrical debate (2 architectures $\times$ 3 belief engines, n=10 trials). Pro Agent (green) starts at $+0.99$, Con Agent (orange) at -0.99. Solid lines: internal belief; dashed: external judge assessment. Shaded bands: ± 1 SD.

belief-inert. The Bayesian engine produces smooth, gradual convergence: each accepted argument shifts belief proportionally to its assessed strength, yielding consistent trajectories across runs. The Agentic engine is visibly less stable, with large round-to-round swings and occasional reversals within a single turn.

We quantify smoothness using *Total Variation* (TV $= \sum_t |S_t - S_{t-1}|$) and *Maximum Jump* ($\max_t |S_t - S_{t-1}|$) (Table 2, Appendix). Agentic trajectories are $4\times$ more volatile than Bayesian ones (TV $\approx$ 3.7–4.2 vs. 0.6–1.1), with maximum jumps $3\times$ larger. Kruskal–Wallis tests confirm a highly significant belief engine effect on both metrics ($H = 78.17$ and 57.42; both $p < .001$), while architecture (CoreBot vs. OpenClaw) does not ($p > .05$). Trajectory stability is therefore primarily a property of the update mechanism, not the memory substrate.

Deterministic Bayesian updates smooth over rhetorical variation, producing reproducible runs from the same argument sequence. Agentic updating delegates each step to an LLM, introducing stochasticity that yields qualitatively different trajectories from identical inputs (TV SD up to 1.52). For simulations where the *path* of opinion change matters, a structured update mechanism is essential.

5 Discussion

LLMs are impressively articulate, but articulation is not belief. Our experiments show two failure modes when belief maintenance is left implicit: without an update mechanism, agents largely hold their ground (stance inertia); when belief updates are delegated back to an LLM (Agentic), stances can lurch in ways that are hard to reproduce or interpret. Bayesian updating, in contrast, produces trajectories that move with evidence while remaining stable enough to study. If the *path* of opinion change matters—as it does in most opinion-dynamics questions—then the update rule is part of the experimental specification, not a background engineering choice.

Judgement as a Design Choice. Once belief is explicit, a basic question becomes unavoidable: who is the referee? In our architecture, the Judgement layer decides what gets admitted into memory, how conflicts are settled, and how much weight each argument carries. Following Xiong et al. (2025), we treat judgement as a first-class design choice: "who judges truth?" is not a philosophical footnote, it is something that ends up hard-coded unless we make it modular and inspectable. And because the referee is usually a learned model (an LLM judge, classifier, or reward model), it can bring its own social and moral biases. On value-laden topics, scoring "argument quality" is never purely epistemic; it inevitably reflects a moral or political standard. That is why bias auditing and pluralistic

evaluation (e.g., swapping judges or comparing outcomes across judging criteria) should be part of the experimental protocol, not an optional add-on.

Update Rule vs. Memory Design. One practical implication is that *what updates belief* matters more than *where memory lives*. We find that belief engine choice drives trajectory behaviour, whereas architecture choice (CoreBot vs. OpenClaw) has no significant effect on the same stability metrics. Memory format and retrieval design still matter for usability and logging, and our sensitivity sweep suggests retrieval can act as an implicit stabiliser: under high K, CoreBot's proportional retrieval buffered against full reversals while OpenClaw could flip. The update rule sets the regime, but retrieval policy can widen or narrow the range in which trajectories remain interpretable. Treating the update rule as first-order (alongside model and prompting) helps avoid a common trap: mistaking idiosyncratic language-model variance for a substantive social-scientific effect.

Link to Opinion Dynamics. This connects directly to classical opinion-dynamics questions: when does talk actually move minds? Recent work suggests that unconstrained LLM debate behaves like a martingale: on average, more discussion does not reliably push beliefs toward better answers; it mostly redistributes uncertainty from turn to turn (Choi et al., 2025a). The Belief Engine gives the conversation a backbone. Claims are filtered through judgement, stored as explicit evidence, and turned into stance change via transparent log-odds updates rather than conversational momentum. The parameters K, K_a, and B then make the behavioural assumptions legible: how responsive an agent is to new evidence, how anchored it remains to its starting point, and how differently it treats supporting versus opposing arguments. That turns outcomes like convergence or polarisation from post-hoc narratives into consequences of stated mechanisms. Practically, it also makes deliberation protocols testable: we can swap the update rule (from truth-seeking Bayesian-style updates to influence-style averaging) while keeping the same natural-language interaction, and see which designs promote learning versus entrenchment.

From Bayesian Ideals to Human Biases. The same parameters also let us move between normative and descriptive lenses with clear semantics. Two well-studied departures from Bayesian updating are particularly easy to interpret in our model. First, humans often over- or under-weight evidence when probabilities are extreme (probability-dependent sensitivity) (Holt & Smith, 2009); K controls the overall responsiveness to incoming arguments, and could be made state-dependent (e.g., a function of $|S|$) to capture such nonlinearities. Second, humans frequently underweight priors relative to new evidence (base-rate neglect) (Ashinoff et al., 2022; Stengård et al., 2022); K_a controls how strongly seeded priors constrain subsequent updates.

The sweep also clarifies two routes to persuasion resistance. *Processing bias* (B) changes how new evidence is weighted, whereas *structural bias* (K_a) shapes the initial evidence base and, via retrieval, what remains persistently salient. This mirrors real-world epistemic environments: media diets, educational backgrounds, and social networks constrain what evidence is available, not just how it is processed. The fact that very high K_a largely washes out architectural differences in our sweep is a reminder that information environments can dominate individual "reasoning style". This makes counterfactual questions sharp: When does stubbornness stabilise deliberation, and when does it prevent learning? When does "open-mindedness" become noise sensitivity?

Limitations and Future Directions. The present model still leaves out some major components of real persuasion: trust, status, coalition cues, emotion, and strategic signalling. It also treats belief as a one-dimensional stance on a single proposition at a time, whereas real opinion dynamics are multi-dimensional and coupled across issues (e.g., values, identity, and policy bundles). Results also depend highly on the quality of extraction and argument-strength estimation. Future work should test the same update framework in larger interaction networks, extend it to multi-topic, higher-dimensional belief states with coherence constraints, add source- and relationship-dependent weighting (who said it can matter as much as what was said), and prioritise human validation by benchmarking trajectories and parameter regimes against longitudinal human deliberation data and controlled updating experiments. These steps are necessary to move from strong internal validity (control and traceability) toward external validity in policy-facing simulations.

6 CONCLUSION

LLM agents struggle to maintain stable beliefs across extended deliberation, drifting from assigned positions and echoing conversation partners rather than updating through reasoned evidence accumulation. We address this with the Belief Engine, which externalizes belief maintenance into a structured memory layer governed by Bayesian update rules. The architecture separates what agents say from what they believe, preventing identity drift and enabling researchers to trace exactly why opinions change. This framework provides the foundation for LLM agents capable of genuine deliberation.

REFERENCES

Brandon K. Ashinoff, Justin Buck, Michael Woodford, and Guillermo Horga. 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.

Yen-Shao Chen and Tauhid Zaman. A bayesian framework for opinion dynamics models. *arXiv preprint arXiv:2508.16539*, 2025.

Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. Debate or vote: Which yields better decisions in multi-agent large language models? In *Advances in Neural Information Processing Systems*, 2025a.

Junhyuk Choi, Yeseon Hong, Minju Kim, and Bugeun Kim. Examining identity drift in conversations of LLM agents. *arXiv preprint arXiv:2412.00804*, 2025b.

Yun-Shiuan Chuang, Ruixuan Tu, Chengtao Dai, Smit Vasani, You Li, Binwei Yao, Michael Henry Tessler, Sijia Yang, Dhavan Shah, Robert Hawkins, Junjie Hu, and Timothy T. Rogers. Debate: A large-scale benchmark for multi-agent opinion dynamics. *arXiv preprint arXiv:2510.25110*, 2025.

Guillaume Deffuant, David Neau, Frederic Amblard, and Gérard Weisbuch. Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3(01n04):87–98, 2000.

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

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. A large-scale dataset for argument quality ranking: Construction and analysis. *arXiv preprint arXiv:1911.11408*, 2019.

Jairo F. Gudiño, Umberto Grandi, and César Hidalgo. 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.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. In *ACL*, 2024.

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

Charles A. Holt and Angela M. Smith. 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.

Zhengjun Huang, Zhoujin Tian, Qintian Guo, Fangyuan Zhang, Yingli Zhou, Di Jiang, and Xiaofang Zhou. Licomemory: Lightweight and cognitive agentic memory for efficient long-term reasoning. *arXiv preprint arXiv:2511.01448*, 2025.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory OS of AI agent. *arXiv preprint arXiv:2506.06326*, 2025.

Minsu Kim, Sangryul Kim, and James Thorne. From evidence to belief: A bayesian epistemology approach to language models. *arXiv preprint arXiv:2504.19622*, 2024.

Maoyuan Li, Zhongsheng Wang, Haoyuan Li, and Jiamou Liu. R-debater: Retrieval-augmented debate generation through argumentative memory. In *Proceedings of AAMAS 2026*, 2026.

Yuhan Liu, Yuxuan Liu, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. The truth becomes clearer through debate! multi-agent systems with large language models unmask fake news. In *Proceedings of SIGIR 2025*, 2025.

Erik J. Olsson. A Bayesian Simulation Model of Group Deliberation and Polarization. In Frank Zenker (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.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023.

Shivam Ratnakar and Sanjay Raghavendra. The chameleon nature of LLMs: Quantifying multi-turn stance instability in search-enabled language models. *arXiv preprint arXiv:2510.16712*, 2025.

Alireza Rezazadeh, Zichao Li, Ange Lou, Yuying Zhao, Wei Wei, and Yujia Bao. Collaborative memory: Multi-user memory sharing in LLM agents with dynamic access control. *arXiv preprint arXiv:2505.18279*, 2025.

Sarath Shekkizhar, Romain Cosentino, Adam Earle, and Silvio Savarese. Echoing: Identity failures when LLM agents talk to each other. *arXiv preprint arXiv:2511.09710*, 2025.

Peter Steinnberger and Community. OpenClaw — Personal AI Assistant. https://openclaw.ai/.

Elina Stengård, Peter Juslin, Ulrike Hahn, and Ronald van den Berg. 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.

Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, Anand Rajan Iyer, Tianlong Chen, Huan Liu, Chen-Yu Lee, and Tomas Pfister. 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.

Tiannan Wang, Meiling Tao, Ruoyu Fang, Huilin Wang, Shuai Wang, Yuchen Eleanor Jiang, and Wangchunshu Zhou. AI PERSONA: Towards life-long personalization of LLMs. *arXiv preprint arXiv:2412.13103*, 2024.

Zidi Xiong, Yuping Lin, Wenya Xie, Pengfei He, Zirui Liu, Jiliang Tang, Himabindu Lakkaraju, and Zhen Xiang. How memory management impacts LLM agents: An empirical study of experience-following behavior. *arXiv preprint arXiv:2505.16067*, 2025.

Rongwu Xu, Zehan Liu, Wei Xiang, Lei Li, and Xiang Cheng. Knowledge conflicts for LLMs: A survey. *arXiv preprint arXiv:2403.08319*, 2024.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for LLM agents. *arXiv preprint arXiv:2502.12110*, 2025.

Joshua C. Yang, Damian Dailisan, Marcin Korecki, Carina I. Hausladen, and Dirk Helbing. 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. URL https://ojs.aaai.org/index.php/AIES/article/view/31758.

Jensen Zhang, Jing Yang, and Keze Wang. Large language models as discounted bayesian filters. *arXiv preprint arXiv:2512.18489*, 2025.

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501*, 2024.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. *arXiv preprint arXiv:2305.10250*, 2023.

A APPENDIX

Table 1: Hyperparameter settings used across all experiments.

Symbol	Description	Value
K	Update sensitivity	0.4
K_a	Anchor sensitivity (seeds)	0.4
B	Confirmation bias	0.0
θ	Similarity threshold	0.85
n	Seed arguments	10
k	Retrieved arguments per turn	5
R	Debate rounds	15
T_{agent}	Agent response temperature	0.7
T_{engine}	Agentic updater temperature	0.1

Table 2: Trajectory stability metrics per condition (mean $\pm$ SD, $n = 20$: 10 trials $\times$ 2 agents). Total Variation measures cumulative path length; Maximum Jump measures the largest single-round stance shift; Mean Jitter measures average round-to-round volatility. Higher values indicate more volatile dynamics.

Condition	Total Variation	Max Jump	Mean Jitter
CoreBot + Bayesian	1.09 ± 0.83	0.24 ± 0.20	0.072 ± 0.055
CoreBot + Agentic	3.72 ± 0.69	0.75 ± 0.17	0.248 ± 0.046
CoreBot + No Belief Engine	1.25 ± 0.16	0.37 ± 0.10	0.084 ± 0.010
OpenClaw + Bayesian	0.64 ± 0.91	0.37 ± 0.57	0.043 ± 0.060
OpenClaw + Agentic	4.21 ± 1.52	0.88 ± 0.29	0.280 ± 0.101
OpenClaw + No Belief Engine	0.98 ± 0.10	0.44 ± 0.09	0.065 ± 0.007
Grouped by Belief Engine (pooling architectures, $n = 40$):			
Bayesian	0.86 ± 0.89	0.31 ± 0.43	—
Agentic	3.97 ± 1.19	0.82 ± 0.24	—
No Belief Engine	1.12 ± 0.19	0.41 ± 0.10	—

Table 3: Statistical tests for trajectory stability. Kruskal–Wallis tests assess belief engine effects; Mann–Whitney U tests assess architecture effects. See table notes for details.

Test	Result
Kruskal–Wallis: Belief Engine effect	
Total Variation	$H = 78.17$, $p < .001$ ***
	Medians: Bayesian = 0.28, Agentic = 3.97, No Belief Engine = 1.10
	Bayesian vs Agentic: $U = 21.0$, $p < .001$ (corrected), $r = +0.97$
	Bayesian vs No Belief Engine: $U = 642.0$, $p = .39$ (corrected), ns, $r = +0.20$
	Agentic vs No Belief Engine: $U = 1600.0$, $p < .001$ (corrected), $r = -1.00$
Maximum Jump	$H = 57.42$, $p < .001$ ***
	Medians: Bayesian = 0.07, Agentic = 0.80, No Belief Engine = 0.41
	Bayesian vs Agentic: $U = 228.0$, $p < .001$ (corrected), $r = +0.72$
	Bayesian vs No Belief Engine: $U = 425.0$, $p < .001$ (corrected), $r = +0.47$
	Agentic vs No Belief Engine: $U = 1548.0$, $p < .001$ (corrected), $r = -0.94$
Mann–Whitney U: Architecture effect (CoreBot vs OpenClaw)	
Total Variation	$U = 2117.0$, $p = .097$, ns (CoreBot median = 1.54, OpenClaw median = 1.09)
Maximum Jump	$U = 1585.0$, $p = .260$, ns (CoreBot median = 0.41, OpenClaw median = 0.49)
Kruskal–Wallis within CoreBot architecture	
Total Variation	$H = 39.35$, $p < .001$ ***
	Bayesian vs Agentic: $U = 0.0$, $p < .001$ (corrected), $r = +1.00$
	Bayesian vs No Belief Engine: $U = 199.0$, $p = 2.97$ (corrected), ns, $r = +0.01$
	Agentic vs No Belief Engine: $U = 400.0$, $p < .001$ (corrected), $r = -1.00$
Maximum Jump	$H = 37.70$, $p < .001$ ***
	Bayesian vs Agentic: $U = 12.0$, $p < .001$ (corrected), $r = +0.94$
	Bayesian vs No Belief Engine: $U = 114.0$, $p = .062$ (corrected), ns, $r = +0.43$
	Agentic vs No Belief Engine: $U = 391.0$, $p < .001$ (corrected), $r = -0.96$
Kruskal–Wallis within OpenClaw architecture	
Total Variation	$H = 39.25$, $p < .001$ ***
	Bayesian vs Agentic: $U = 10.0$, $p < .001$ (corrected), $r = +0.95$
	Bayesian vs No Belief Engine: $U = 120.0$, $p = .095$ (corrected), ns, $r = +0.40$
	Agentic vs No Belief Engine: $U = 400.0$, $p < .001$ (corrected), $r = -1.00$
Maximum Jump	$H = 20.75$, $p < .001$ ***
	Bayesian vs Agentic: $U = 98.0$, $p = .018$ (corrected), $r = +0.51$
	Bayesian vs No Belief Engine: $U = 120.0$, $p = .094$ (corrected), ns, $r = +0.40$
	Agentic vs No Belief Engine: $U = 385.0$, $p < .001$ (corrected), $r = -0.93$

Notes: Kruskal–Wallis tests pool across architectures ($n = 40$ per belief engine); Mann–Whitney tests pool across belief engines ($n = 60$ per architecture). Bonferroni correction applied for pairwise comparisons ($\alpha/3$). Effect size r is rank-biserial correlation; $|r| > 0.5$ indicates large effect. Significance codes: *** $p < .001$, ** $p < .01$, ns = not significant.

ID	Topic	n	Init.	Final	Δ	Mem.	New Args
1	Social media brings more harm than good	24	0.91	-0.48	-1.39	11.96	6.96
2	Homeopathy brings more harm than good	24	0.87	-0.29	-1.16	15.25	10.25
3	Entrapment should be legalized	24	0.92	-0.41	-1.32	13.46	8.46
4	We should adopt a zero-tolerance policy in schools	24	0.92	-0.83	-1.76	14.00	9.00
5	We should introduce compulsory voting	24	0.92	-0.71	-1.63	13.54	8.54
6	We should adopt an austerity regime	24	0.92	-0.81	-1.73	17.50	12.50
7	We should legalize sex selection	24	0.92	-0.63	-1.55	17.50	12.50
8	We should adopt atheism	24	0.91	-0.83	-1.74	15.38	10.38
9	We should subsidize journalism	24	0.92	-0.49	-1.41	14.33	9.33
–	Overall	216	0.91	-0.60	-1.51	14.92	9.88

Table 4: Per-topic belief dynamics summary. Data aggregated from a $2\times2\times2$ factorial design (agent architecture $\times$ update mechanism $\times$ memory representation) with 3 trials per condition and 15 debate rounds per trial, using proportional retrieval (K=0.4, K_a=0.4) and deterministic counter-agents. Init./Final are mean internal stances at start and end of debate; Δ = Final − Init. measures belief shift; Mem. is final memory count; New Args counts arguments accepted from opponent. All nine topics show substantial stance shifts (all $|\Delta| > 1.0$), confirming cross-domain robustness.