MNEMA: A Witness Lattice for Living Memory in Multi-Agent AI Systems

Ala Smith

Gentic Lab, ala.smith@gentic.news

Abstract. Today's AI agents store memory as passive records and route decisions through a single orchestrator. Three failure classes follow: multi-agent coordination collapses, memory is poisoned, and decisions are opaque to audit. We propose MNEMA, an architecture in which each memory unit is an autonomous *witness*: a node with cryptographic identity, an append-only signed journal, a five-stage developmental lifecycle, and the structural right to refuse. Witnesses are *generative*—under quorum-gated gossip they spawn new witnesses, split, coalesce, and probe for fresh evidence—so the graph itself evolves, with every transition recoverable by traversal of signed journals and lineage edges. Decisions emerge from a nine-step commitment protocol producing a cryptographically signed Provenance DAG. We formalise a probabilistic threat model and derive a closed-form expression for undetected poisoning under a latent common-shock corruption model, $P_{\text{undetected}} = \alpha + (1-\alpha)\,\beta^{1+q}$, where α is the shared-root-cause shock probability. From this identity we derive a constructive optimal-redundancy theorem (Theorem 4) giving the minimum kin depth required to meet a target tolerance, and show that detection is hard-floored at $1-\alpha$ regardless of redundancy. A worst-case corollary (Corollary 2) further shows that fragment redundancy without structural decorrelation reduces to single-storage protection. We pre-register a MemoryGraft-survival demonstration with explicit power analysis and falsification criteria. MNEMA targets adversarial, multi-party, and audit-regulated deployments; we state plainly which claims are formally proved (Property 1; Theorems 1, 2, 3, 4) and which await empirical work.

Keywords: multi-agent systems · agentic memory · Byzantine consensus · abstention · decision provenance · AI safety · witness systems · living memory

1 Introduction

Three documented failures. Centralised orchestration of agentic AI systems is failing in three documented ways. (i) Multi-agent coordination collapses: the MAST taxonomy of Cemri et al. [3] catalogues fourteen failure modes with 41–86.7% failure rates across seven SOTA frameworks, and the authors conclude that better base models will not fix these failures. (ii) Memory layers are vulnerable: MemoryGraft [11] achieves 50% poisoned recall via benign markdown ingestion; AgentPoison [4] exceeds 80% at 0.1% poison rate; PoisonedRAG [14]

achieves 90% with five poisoned texts in millions. (iii) Decisions are opaque: a single agent's chain-of-thought is unstructured natural language, and post-hoc attribution is unreliable under EU AI Act, HIPAA, SOC2, and analogous frameworks.

Root cause and inversion. We claim these three failure classes share a common root cause: *the architecture treats memory as a passive substrate and concentrates decision authority in a learned orchestrator*. We propose its inversion. Memory units become autonomous *witnesses*; decisions are produced by an explicit nine-step protocol over witnesses (§7, Algorithm 1). The trade is flexibility for auditability: a learned orchestrator policy is opaque and not formally verifiable; a fixed protocol is auditable, replayable, and has explicit, named failure modes.

Living memory. The witness lattice is not static. Witnesses spawn new witnesses through gossip surplus, split when their cross-domain action variance exceeds threshold, coalesce when their canonical claims overlap, and probe their environment under restraint-budgeted discovery. The history of every birth, split, coalescence, retirement, and discovery is preserved in journals and verifiable from cryptographic signatures alone. This is the central distinguishing property of Mnema: the knowledge graph is a *living* object whose evolution is itself an audit artefact.

What this paper proves and what it does not. We provide formal definitions of witness, journal, lattice, and commitment protocol; a probabilistic threat model; one structural property and four theorems with proofs (Property 1; Theorems 1, 2, 3, 4). We pre-register one empirical demonstration (MemoryGraft survival; §9) with explicit power analysis and falsification criteria. We do *not* claim empirical superiority on coordination or audit-replay benchmarks; those are future work.

Contributions. (C1) Five-layer architecture (§3) and witness calculus with hash-chained journals (§4) carrying tamper-evidence (Property 1). (C2) Living-memory dynamics (§6)—genesis, split, coalescence, discovery—inside the audit framework (Theorem 1). (C3) Nine-step commitment protocol (§7, Algorithm 1) producing a signed Provenance DAG. (C4) Two-channel reputation tensor with cross-domain transfer (§7.1). (C5) Probabilistic threat model (§8.1), independence bound (Theorem 2), common-shock identity (Theorem 3), constructive optimal-depth theorem (Theorem 4) with $1 - \alpha$ impossibility floor. (C6) Pre-registered MemoryGraft-survival demonstration with explicit falsification criteria (§9).

2 Related Work

Agentic memory. Production memory systems span structured memory blocks (MemGPT/Letta [9]), graph-based memory (Mem0g [5], Zep [10]), and Zettelkasten-style evolving notes (A-MEM [13]). The witness concept is closest to A-MEM's

notes; A-MEM does not give them decision rights, signed journals, or right of refusal, nor does it formalise an audit framework.

Multi-agent orchestration. Production frameworks (AutoGen v0.4, Magentic-One, CrewAI) remain orchestrator-centred even when the substrate is the actor model. Cemri et al. [3] establish that 41–86.7% failure rates persist across SOTA frameworks regardless of base-model capability.

Distributed consensus and BFT for LLMs. Classical Byzantine fault tolerance [7] assumes deterministic verification, which LLM agents do not provide. Recent adaptations to LLM ensembles use decoder-level confidence, leaderless aggregation, or weighted voting in isolation; Mnema differs by integrating confidence, dynamic reputation, model-family diversity, first-class abstention, and a constitutional gate within a single signed protocol.

Constitutional AI, debate, and abstention. Bai et al. [1] demonstrate behaviour shaping via written constitutions plus self-critique. Brown-Cohen and Irving [2] prove doubly-efficient debate guarantees an honest prover wins in polynomial time even against a bounded dishonest one. Verga et al. [12] (PoLL) show small cross-family LLM panels beat single GPT-4 judges at $\sim 7\times$ lower cost. Abstention remains unsolved at the model layer, motivating its structural treatment here.

Memory poisoning. The triple of MemoryGraft, AgentPoison, and PoisonedRAG [11,4,14] establishes the modern threat landscape. Steganographic and secret collusion among agents [8] forms covert channels even when communication is monitored.

3 Architecture Overview

Mnema is organised as a five-layer stack (Fig. 1). The substrate (memory filesystem, IATP identity registry, event bus, vector index, workflow engine, inference servers) is reused infrastructure. The witness (§4), the lattice (§5), the living-memory dynamics (§6), the commitment protocol (§7), and the audit DAG are this paper's contributions.

Notation. $\mathcal{X}$: workspace; $\mathcal{D}$: domains; Γ: context tags; $\mathcal{V}$: evidence. For witness w, $\mathcal{D}_w \subseteq \mathcal{D}$ is its relevant domains and $\mathcal{D}_w^{\text{decisive}} \subseteq \mathcal{D}_w$ the subset of decisive authority via the Sen-domain mapping (§5). $H : \{0,1\}^* \to \{0,1\}^{256}$ is SHA-256; $(\text{Sign}, \text{Ver})$ is Ed25519; $x \,\|\, y$ denotes binary concatenation.

4 The Witness

Definition 1 (Witness). *A* witness *is a tuple*

$$w = \big(\text{ID}_w,\ \mathcal{C}_w,\ \mathcal{E}_w,\ s_w,\ \mathcal{J}_w\big) \tag{1}$$

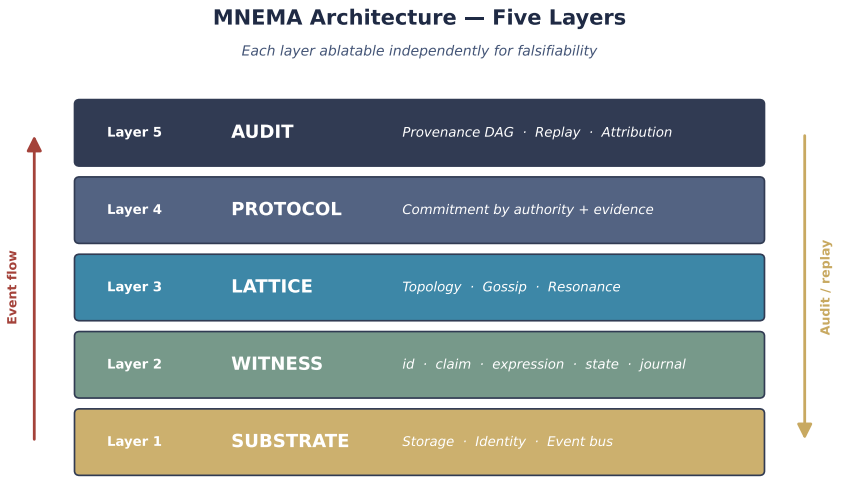

Fig. 1. The five-layer stack. Each layer adds one capability and is ablatable independently. Living-memory dynamics span the witness–lattice–protocol layers via the gossip channel.

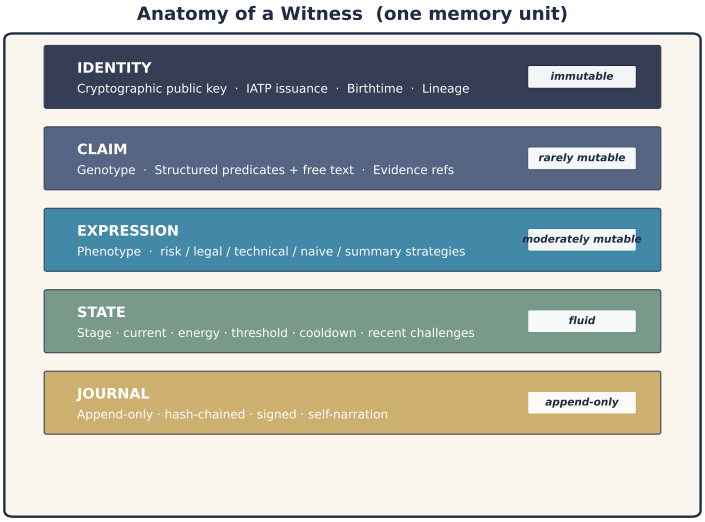

Fig. 2. Anatomy of a witness. Five components with different mutability rates.

where $\mathrm{ID}_w \in \{0,1\}^{256}$ *is an Ed25519 public key issued via IATP;* $\mathcal{C}_w \in 2^{\mathcal{V}}$ *is the signed claim (structured predicates plus free text plus a set of evidence references);* $\mathcal{E}_w = \{\eta_c\}_{c \in \Gamma}$ *is the family of context-keyed expression strategies with* $\mathrm{predicates}(\eta_c) \subseteq \mathrm{predicates}(\mathcal{C}_w)$ *for all* c *(*phenotypic non-contradiction*: expressions may summarise or redact but never contradict the canonical claim);* $s_w \in \mathcal{S}$ *is the dynamic state; and* $\mathcal{J}_w$ *is the append-only signed journal.*

Definition 2 (Journal hash chain). *The journal of* w *is a sequence* $\mathcal{J}_w = (e_0, e_1, \ldots, e_n)$ *where each entry* $e_i = (\mathrm{type}_i, t_i, \mathrm{payload}_i, \sigma_i, h_i)$ *has hash chain*

$$h_0 = H(\mathrm{ID}_w), \qquad h_i = H\big(h_{i-1} \,\|\, \mathrm{type}_i \,\|\, t_i \,\|\, \mathrm{payload}_i\big), \tag{2}$$

and signature $\sigma_i = \mathrm{Sign}_{sk_w}(h_i)$*, where* sk_w *is the secret key associated with* ID_w*.*

Property 1 (Tamper-evidence). Suppose H is collision-resistant and Ed25519 is existentially unforgeable under chosen-message attack (EUF-CMA). For any non-trivial modification of $\mathcal{J}_w = (e_0, \ldots, e_n)$ that produces $\mathcal{J}'_w = (e_0, \ldots, e_{i-1}, e'_i, e_{i+1}, \ldots, e_n)$ with $e'_i \neq e_i$ and $i < n$, verification of the latest signature σ_n against ID_w's public key fails with all but negligible probability. Tamper-evidence covers integrity of past entries against an attacker without sk_w; it does not cover semantic correctness of authored entries, key compromise, equivocation, or rollback. The latter three require external anchoring of the latest head in the Provenance DAG.

Proof. Modifying e_i changes $\mathrm{payload}_i$, hence the recomputed h'_i differs from h_i unless the adversary finds a hash collision (negligible probability under collision-resistance). For $j > i$ the recomputed $h'_j = H(h'_{j-1} \,\|\, \cdots)$ depends on h'_{j-1}, so $h'_j \neq h_j$ holds unless a collision is found at step j. By the union bound the probability of at least one collision in the $n - i$ remaining steps is at most $(n-i)\cdot\mathrm{negl}$, still negligible. Verification of the original σ_n against the recomputed h'_n therefore requires the adversary to produce a forgery, a negligible event under EUF-CMA. □

Lifecycle. A witness progresses monotonically EMBRYONIC → JUVENILE → ADULT → ELDER → PHANTOM (Fig. 3); skip-ahead transitions (e.g., to PHANTOM on split-supersedee) are permitted but no stage is revisited. Stages admit different authority weights, mutability, inhibition rights, and death conditions; transitions are journalled and verifiable from $\mathcal{J}_w$ alone. The PHANTOM stage is structurally novel: a retired witness intercepts retrievals and emits a structured refusal pointing to its successor and prior-claim validity window, addressing silent-knowledge-loss in memory systems that delete deprecated entries.

5 The Lattice

Definition 3 (Witness lattice). *A* witness lattice *is a typed directed graph* $\mathcal{L} = (\mathcal{W}, E_{\mathrm{lin}}, E_{\mathrm{auth}}, E_{\mathrm{res}})$ *where* $\mathcal{W}$ *is a set of witnesses and the three edge sets capture lineage (parent–child relations from genesis, splits, and coalescence), authority (per Sen-domain mapping), and resonance (co-firing learned from telemetry). We use* lattice *in the architectural sense—a layered, typed graph—and not in the order-theoretic sense (no meet/join is required).*

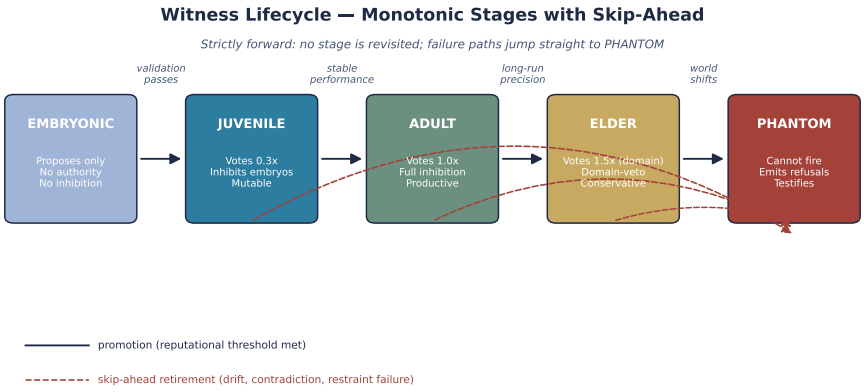

Fig. 3. Five-stage witness lifecycle. Solid arrows denote promotion on reputational threshold; dashed arrows are skip-ahead retirements to PHANTOM triggered by drift, contradiction, or restraint failure. Transitions are monotonic—no stage is revisited.

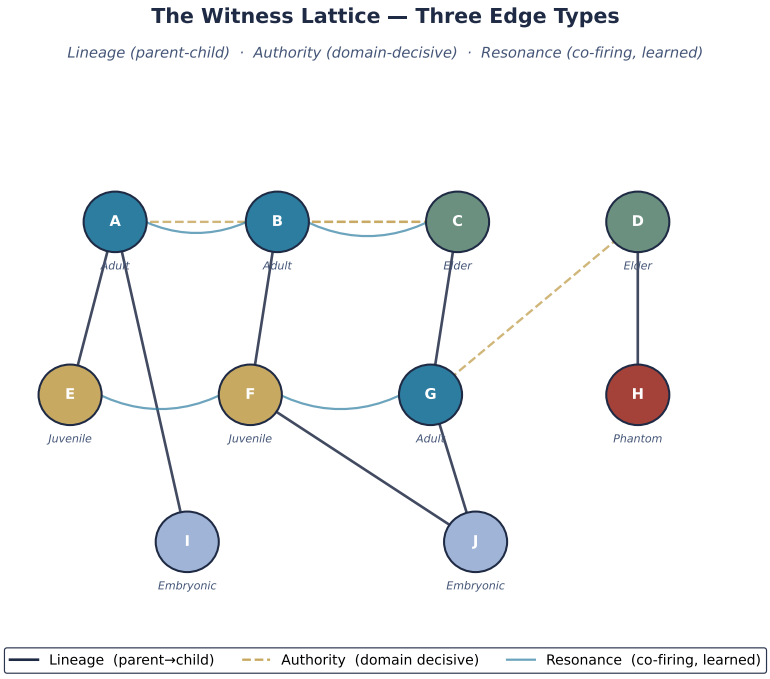

Fig. 4. Witness lattice with three edge types.

Activation. A witness w *fires* on event $x \in \mathcal{X}$ when its precision-weighted prediction error exceeds firing cost (a free-energy-style threshold):

$$\text{fire}(w, x) \iff \pi_w \cdot \big| \, y_w(x) - \hat{y}_w(x) \, \big| > c_w + \xi(s_w), \tag{3}$$

where $\pi_w \in \mathbb{R}_+$ is precision (learnable from journal events), $y_w(x)$ is the workspace signal, $\hat{y}_w(x)$ is the witness's prediction, c_w is the firing cost, and $\xi : \mathcal{S} \to \mathbb{R}$ is a state-dependent congestion penalty. Within an activated clique we use locally competitive activation dynamics:

$$\frac{du_i}{dt} = -\lambda u_i + I_i(t) - \sum_{j \neq i} W_{ij} \, \sigma\big(u_j - \theta_j\big). \tag{4}$$

Pairwise idle-time consolidation. Idle witnesses run a six-verb gossip grammar: ASSERT, CORROBORATE, CONTRADICT, PROPOSE, ASSENT, DISSENT. Topological propagation converges in $O(\log n)$ rounds with high probability under standard gossip assumptions [6]. Convergence on semantic-claim equality between LLM-arbitrated witnesses is open (Sect. 10).

6 Living Memory: Generative Dynamics

The lattice is a *living* object: witnesses spawn new witnesses, split, coalesce, and probe their environment. Four formal mechanisms drive this evolution; each is gated by gossip quorum and journaled, so every transition leaves an audit trail (Theorem 1).

Definition 4 (Gossip surplus). *Let c be a candidate claim and let $G(c, \Delta t)$ be the set of* distinct *witnesses that issued an* ASSERT *or* CORROBORATE *for c during a sliding window Δt. The* gossip surplus *of c is*

$$s(c, \Delta t) = \big|G(c, \Delta t)\big| \; - \; \big|\{w \in \mathcal{W} : c \in \mathcal{C}_w\}\big|, \tag{5}$$

i.e., distinct corroborators minus witnesses that already canonicalise c.

Definition 5 (Genesis). *If $s(c, \Delta t) \geq q_{\text{birth}}$, no witness canonicalises c, and a cross-family critic-jury returns $\min_i j_i(c) \geq \nu_0$, an embryonic witness w_{new} is instantiated with claim c, fresh Ed25519 identity, lineage edges to the gossip participants in $G(c, \Delta t)$, and an empty journal seeded by a signed* BIRTH *entry.*

Definition 6 (Generative split). *A witness w with high cross-domain action variance,*

$$\text{Var}_{d \in \mathcal{D}_w}\big(R^{\text{act}}[w, d, \cdot]\big) > \theta_{\text{split}}, \tag{6}$$

is partitioned into two children via spectral clustering on the cross-domain similarity kernel K (Eq. 10). Each child inherits the journal entries, evidence references, and reputation slice relevant to its cluster. The parent transitions to PHANTOM *as supersedee.*

Definition 7 (Coalescence). *Let* $\mathrm{emb} : 2^{\mathcal{V}} \to \mathbb{R}^d$ *be a fixed claim embedding, and call two journals* non-contradictory *if no entry of one carries a* CONTRADICT *payload pointing to an entry of the other. Two embryonic witnesses* w_1, w_2 *with overlapping canonical claims (*$\cos(\mathrm{emb}(\mathcal{C}_{w_1}), \mathrm{emb}(\mathcal{C}_{w_2})) > \theta_{\mathrm{merge}}$*) and non-contradictory journals coalesce, under unanimous* ASSENT *from their gossip neighbourhoods, into a witness* $w_{1,2}$ *inheriting the union of evidence references, the entry-wise maximum of per-domain reputation, and the union of lineage.*

Definition 8 (Restraint-budgeted discovery). *Let* $\mathrm{fresh}(\mathcal{C}_w) \in [0,1]$ *be the freshness score of the canonical evidence (decreasing in time since last update). A witness* w *with high restraint precision but stale evidence (*$\mathrm{fresh}(\mathcal{C}_w) < \theta_{\mathrm{stale}}$*) may spend a portion of its restraint budget to query the substrate retrieval layer for new ingestion candidates in* $\mathcal{D}_w^{\mathrm{decisive}}$*, evaluate each candidate against* $\mathcal{C}_w$*, and gossip the strongest as a proposed evidence update.*

Theorem 1 (Genesis preserves auditability). *For every transition produced by Definitions 5, 6, 7, 8, the provenance of the resulting witness state is recoverable from the journals of the participating witnesses by traversal of the lineage sub-DAG induced by the transition's signed entries.*

Proof (Sketch). Each mechanism produces a constant number of typed journal entries (BIRTH, SPLIT, MERGE, DISCOVERY) on the involved witnesses, signed under their respective IDs. By Property 1, modifying these entries is detectable. The lineage edges referenced in each entry are inverted-indexed in the lattice $\mathcal{L}$, so traversal is straightforward. Coalescence is the only mechanism that consolidates identities; under Definition 7 the merged witness's lineage records both pre-merge ancestors. Therefore the union of journals of all witnesses participating in a transition is a sufficient provenance record. □

Ecosystem health metrics. We track four quantities and surface them in the audit DAG: *diversity* (count of distinct authority-decisive witnesses per domain), *coverage* (fraction of queried domains served by an adult witness), *contradiction density* (fraction of gossip exchanges returning DISSENT), and *freshness median* (median age of canonical evidence). Threshold breaches escalate to human review.

7 The Commitment Protocol

Definition 9 (Legitimacy vector). *A candidate action* a *has a legitimacy vector* $\mathbf{L}(a) = (e, A, f, p, s, c, v) \in [0,1]^7$ *with components oriented so higher is better: evidence* e*, authority* A*, freshness* f*, policy alignment* p*, safety* s *(*$=1-\mathrm{risk}$*), certainty* c *(*$=1-\mathrm{uncertainty}$*), reversibility* v*.*

Let $\mathbf{j}(a) \in [0,1]^k$ be the cross-family critic-jury vector, $V(a) \in \{0,1\}$ the veto-council indicator, $\Pi(a) \in \{0,1\}$ the constitutional gate, and $m(a) = \min_i j_i(a)$.

Let $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_7) \in [0,1]^7$ be domain-calibrated weights on the legitimacy components and $\mu \in [0,1]$ the jury weight, with $\sum_i \omega_i + \mu = 1$. Define the weighted score and concordance

$$S(a) = \sum_{i=1}^{7} \omega_i \, L_i(a) + \mu \cdot \operatorname{med}\big(\mathbf{j}(a)\big), \tag{7}$$

$$\kappa(a) = 1 - \frac{2\,\sigma_{\mathbf{j}(a)}}{\max(\mathbf{j}(a)) - \min(\mathbf{j}(a)) + \varepsilon}. \tag{8}$$

$\kappa(a) \in [0,1]$ is jury concordance (1 unanimous; 0 maximally split); $\varepsilon > 0$ guards against zero denominators. We use κ rather than ρ to avoid collision with the pairwise corruption correlation in §8.2.

Definition 10 (Commitment function).

$$\Phi(a) = \begin{cases} \textit{COMMIT} & S(a) \geq \theta_C \wedge V(a) = 0 \wedge \Pi(a) = 1 \wedge m(a) \geq \theta_m, \\ \textit{ESCALATE} & \theta_E \leq S(a) < \theta_C \;\vee\; \kappa(a) < \theta_\kappa, \\ \textit{DEFER} & f(a) < \theta_f \;\vee\; \kappa(a) < \theta_\kappa^{(\text{low})}, \\ \textit{DIE} & S(a) < \theta_E \;\textit{after escalation.} \end{cases} \tag{9}$$

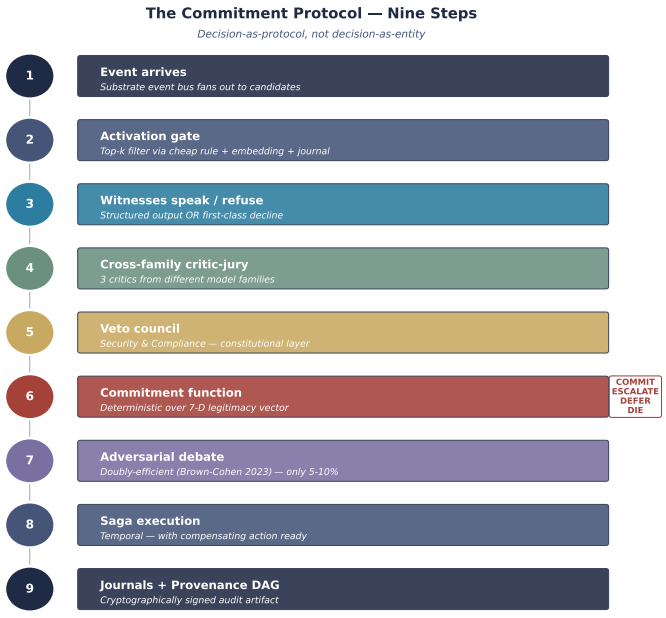

Fig. 5. The nine-step commitment protocol.

Algorithm 1 Commitment Protocol

Input: event x, witness population $\mathcal{W}$, constitution Π
Output: outcome Φ, candidate action a, signed Provenance DAG G

$\mathcal{P}_x \leftarrow \textsc{Retrieve}(\mathcal{W}, x)$ ▷ candidate pool
$\mathcal{W}^* \leftarrow \{w \in \mathcal{P}_x : \text{fire}(w, x)\}$ ▷ Eq. 3
$\mathcal{R} \leftarrow \{\textsc{SpeakOrRefuse}(w, x) : w \in \mathcal{W}^*\}$
$a \leftarrow \textsc{Synthesise}(\mathcal{R})$ ▷ candidate action proposal
$\mathbf{L}(a) \leftarrow \textsc{LegitimacyVector}(a, \mathcal{R}, \mathcal{W}^*)$ ▷ Def. 9
$\mathbf{j}(a) \leftarrow \textsc{CrossFamilyJury}(a, \mathcal{R}, k = 3)$
$V(a) \leftarrow \textsc{VetoCouncil}(a, \Pi)$
$\Phi \leftarrow \textsc{CommitmentFunction}(\mathbf{L}(a), \mathbf{j}(a), V(a), \Pi)$
if $\Phi = \textsc{escalate}$ **then**
    $\Phi \leftarrow \textsc{DoublyEfficientDebate}(a, \mathcal{R}, \mathbf{j}(a))$ ▷ [2]
if $\Phi = \textsc{commit}$ **then**
    $\textsc{SagaExecute}(\text{action}(a), \text{compensation}(a))$
$\mathcal{J}_w \leftarrow \mathcal{J}_w \,\|\, e_\Phi(a)$ **for each** $w \in \mathcal{W}^*$
$G \leftarrow \textsc{ProvenanceDAG}(a, \mathcal{W}^*, \mathcal{R}, \mathbf{j}(a), V(a), \Phi)$
$\textsc{UpdateReputation}(\mathcal{W}^*, \Phi)$
return (Φ, a, G)

Default thresholds $\theta_C = 0.67$, $\theta_E = 0.40$, $\theta_m = 0.20$, $\theta_\kappa = 0.50$, $\theta_\kappa^{(\text{low})} = 0.30$, $\theta_f = 0.60$ are initial values calibrated post-deployment from outcome telemetry. Cases of Eq. 9 are evaluated in order: COMMIT is checked first; on failure the action escalates, defers, or dies as governed by the remaining clauses.

7.1 The reputation tensor

For witness w, domain d, time t, and channel $\bullet \in \{\text{act}, \text{res}\}$, define a Bayesian beta posterior $\text{Beta}(\alpha_w^{\bullet,(d)}, \beta_w^{\bullet,(d)})$ with channel-specific counters: $\alpha^{\text{act}}/\beta^{\text{act}}$ increment on confirmed/disconfirmed firing, $\alpha^{\text{res}}/\beta^{\text{res}}$ on confirmed/disconfirmed decline. The reputation tensor with cross-domain transfer is

$$R^{\bullet}[w, d, t] = \frac{1}{|\mathcal{D}_w|} \sum_{d' \in \mathcal{D}_w} K(d, d') \cdot \frac{\alpha_w^{\bullet,(d')}}{\alpha_w^{\bullet,(d')} + \beta_w^{\bullet,(d')}} \cdot e^{-(t - \tau_w^{(d')})/\tau_d}, \tag{10}$$

where $K : \mathcal{D} \times \mathcal{D} \to [0, 1]$ is a domain similarity kernel (cosine over learned domain embeddings), $\tau_w^{(d')}$ is the time of the last update in domain d', and τ_d is the domain-specific decay half-life.

Definition 11 (Trust). $\text{trust}(w, d, t) = \min\big(R^{\text{act}}[w, d, t],\ R^{\text{res}}[w, d, t]\big)$.

Property 2 (Anti-spam, anti-hiding). A witness firing often on weak evidence has $\beta_w^{\text{act},(d)}$ growing faster than $\alpha_w^{\text{act},(d)}$, so $R^{\text{act}} \to 0$. A witness declining when it should have fired has $\beta_w^{\text{res},(d)}$ growing faster than $\alpha_w^{\text{res},(d)}$, so $R^{\text{res}} \to 0$. The minimum penalises both pathologies.

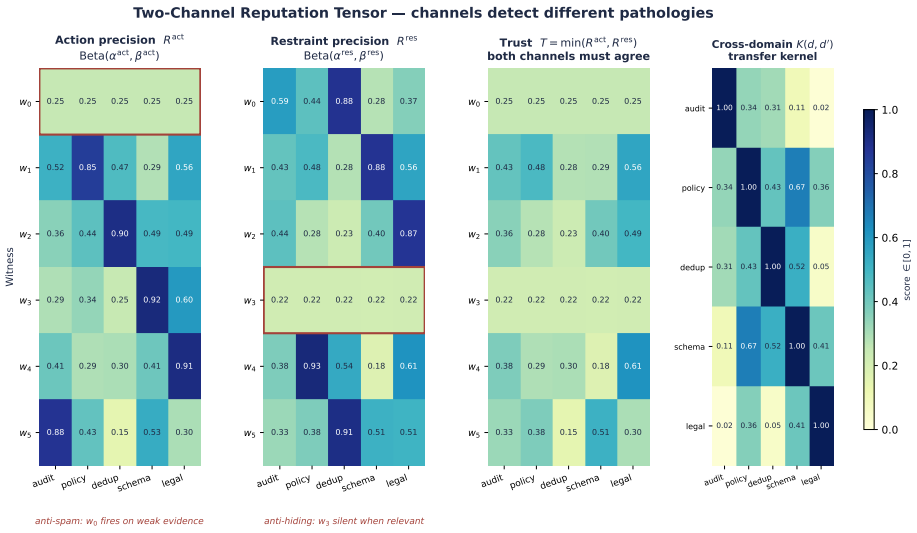

Fig. 6. Two-channel reputation tensor over (witness, domain, time). R^{act} flags spammers (w_0 fires constantly on weak evidence); R^{res} flags hiders (w_3 silent when relevant); $T = \min(R^{\text{act}}, R^{\text{res}})$ penalises both. The cross-domain kernel $K(d, d')$ enables transfer between related domains.

8 Threat Model and Defences

8.1 Formal attacker model

Definition 12 (Adversary). *The adversary $\mathcal{A}$ is a probabilistic polynomial-time agent with the following capabilities (**C**) and limitations (**L**) and trust assumptions (**T**):*

(C1) *inject documents into ingestion sources;* ***(C2)*** *inject markdown into a witness's evidence references at write time;* ***(C3)*** *compromise up to $\lfloor (|\mathcal{W}| - 1)/3 \rfloor$ witnesses (Byzantine threshold);* ***(C4)*** *mount probing attacks on disclosure;* ***(C5)*** *send adversarial messages on inter-witness gossip channels;* ***(C6)*** *perform side-channel timing observations.*

(L1) *cannot forge IATP-issued Ed25519 signatures;* ***(L2)*** *cannot read or modify the Constitutional Layer at runtime;* ***(L3)*** *cannot compromise the Provenance DAG storage backend;* ***(L4)*** *cannot control the cross-family critic-jury simultaneously across all model families;* ***(L5)*** *cannot forge journal entries on uncompromised witnesses.*

(T1) *the IATP key registry is trusted (cryptographic root);* ***(T2)*** *H is collision-resistant and Ed25519 is EUF-CMA secure;* ***(T3)*** *model families are sufficiently* decorrelated *(non-identical training corpora, distinct decoding stacks) that common-shock probability α is small;* ***(T4)*** *the Constitutional Layer and veto-council nodes are correctly specified.*

8.2 Fragment redundancy

A load-bearing claim is encoded across k kin witnesses; each kin holds a fragment ϕ_i tagged with model family, source-graph root, and ingestion path. On retrieval, the canonical and a sample of q kin are checked for consistency.

Theorem 2 (Independent case). *Let $X_i = \mathbf{1}_{\mathcal{A} \text{ corrupts } w_i}$ for $i = 0, 1, \ldots, q$, and assume the X_i are mutually independent with $\Pr[X_i = 1] = p_{\text{corrupt}}$. Then the probability of undetected poisoning of a fragment-redundant claim with one canonical and q checked kin satisfies*

$$P_{\text{undetected}} = \Pr\Big[\bigcap_{i=0}^{q} (X_i = 1)\Big] = p_{\text{corrupt}}^{1+q}. \tag{11}$$

Proof. Detection requires only one of the $1 + q$ witnesses to disagree. Under independence, $\Pr[\bigcap_i (X_i = 1)] = \prod_i \Pr[X_i = 1] = p^{1+q}$. □

Corollary 1. *For $p_{\text{corrupt}} = 0.10$ and $q = 4$, $P_{\text{undetected}} = 10^{-5}$ (compared to 0.10 for single storage).*

The independence assumption is unrealistic. We model correlated corruption with an explicit latent *common-shock* structure that admits a closed-form expression for $P_{\text{undetected}}$.

Definition 13 (Common-shock corruption model). *The corruption of a kin set $\{w_0, \ldots, w_q\}$ follows a* common-shock *model with parameters $(\alpha, \beta) \in [0, 1]^2$ if a single Bernoulli shock variable $Z \sim \text{Bernoulli}(\alpha)$ and independent residuals $Y_i \sim \text{Bernoulli}(\beta)$ jointly determine $X_i = \mathbf{1}_{\mathcal{A} \text{ corrupts } w_i} = Z \vee Y_i$. The shock $Z = 1$ models a shared root cause (common model family, source-graph root, or ingestion path); the residuals Y_i model kin-specific compromises. Structural decorrelation (D1) $\text{family}(w_i) \neq \text{family}(w_j)$; (D2) $\text{source\_root}(w_i) \neq \text{source\_root}(w_j)$; (D3) $\text{ingestion\_path}(w_i) \neq \text{ingestion\_path}(w_j)$ for all $i \neq j$ enforces $\alpha \to 0$.*

Theorem 3 (Correlated case). *Under the common-shock model of Definition 13, the marginal corruption probability is $p = \alpha + (1 - \alpha)\beta$, the pairwise correlation between any two kin is*

$$\rho \;=\; \frac{\alpha(1-\alpha)(1-\beta)^2}{p(1-p)}, \tag{12}$$

and the probability of undetected poisoning across one canonical and q checked kin is given exactly by

$$P_{\text{undetected}} \;=\; \alpha \;+\; (1-\alpha)\,\beta^{\,1+q}. \tag{13}$$

Proof. Conditioning on Z: $\Pr[\forall i, X_i{=}1 \mid Z{=}1] = 1$ and $\Pr[\forall i, X_i{=}1 \mid Z{=}0] = \beta^{1+q}$ by independence of the Y_i, giving the joint expression. The marginal p follows from $\Pr[X_i{=}1] = \alpha + (1-\alpha)\beta$. Direct computation gives $\text{Cov}(X_i, X_j) = \alpha(1-\alpha)(1-\beta)^2$, and dividing by $\text{Var}(X_i) = p(1-p)$ yields Eq. 12. □

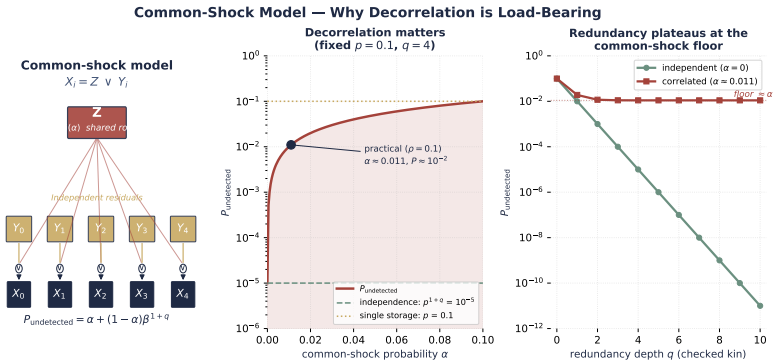

Fig. 7. Common-shock model $X_i = Z \vee Y_i$. **Left:** structure. **Middle:** $P_{\text{undetected}}$ vs shock α at p=0.1, q=4. **Right:** redundancy depth q; correlated case plateaus at α. Decorrelation is load-bearing.

Corollary 2 (Worst case). *As $\alpha \to p$ (full common-shock dominance, residuals vanish), $P_{\text{undetected}} \to p$: redundancy provides no protection beyond single storage. Fragment redundancy without decorrelation is not protective.*

Corollary 3 (Practical regime). *For $\rho = 0.1$, $p = 0.1$, $q = 4$ (Eq. 12): $\alpha \approx 0.0110$, $\beta \approx 0.0900$, so $P_{\text{undetected}} \approx 1.10 \times 10^{-2}$: $\sim 9\times$ better than single storage, $\sim 10^3\times$ worse than the independence bound. The shock floor α dominates.*

Corollary 4 (Detection rate). *$D \triangleq 1 - P_{\text{undetected}} = (1-\alpha)(1-\beta^{1+q})$; as $q \to \infty$, $D \to 1-\alpha$.* No finite redundancy beats $1-\alpha$ detection*: α is a hard impossibility floor.*

Theorem 4 (Optimal redundancy depth). *For target tolerance $\varepsilon > \alpha$, the minimum q with $P_{\text{undetected}} \le \varepsilon$ is*

$$q^\star(\varepsilon, \alpha, \beta) = \left\lceil \log\big((\varepsilon - \alpha)/(1-\alpha)\big) / \log \beta \right\rceil - 1. \tag{14}$$

For $\varepsilon \le \alpha$ no finite q suffices (Cor. 4).

Proof. $\alpha + (1-\alpha)\beta^{1+q} \le \varepsilon \iff \beta^{1+q} \le (\varepsilon - \alpha)/(1-\alpha)$; $\log \beta < 0$ flips the inequality, giving the ceiling. □

Corollary 5 (Design). *At $\beta = 0.09$: $q^\star(10^{-2}, 0.005) = 1$, $q^\star(10^{-3}, 5{\cdot}10^{-4}) = 2$. Halving α buys $\sim$1 order of magnitude at fixed q.*

Engineering implication. The encoder must enforce Definition 13 at kin-selection time; the witness schema records family, source root, and ingestion path for every kin so decorrelation is verifiable at audit (Fig. 7).

Table 1. Mapping from documented threats to architectural defences.

Threat	Defence
MemoryGraft [11]	IATP signature + validation chamber + decorrelated fragment redundancy
AgentPoison [4]	Decorrelated fragment redundancy (Theorem 3)
PoisonedRAG [14]	Source-graph diversity check at protocol layer
Steganographic collusion [8]	Decoder activation probes; cross-family mandate
Authority impersonation	IATP authority tokens domain-scoped, time-bounded
Sycophantic jury	Concordance penalty $\kappa(a)$ (Eq. 8); required dissent
MAST FM-2.4 (info withholding)	Withholding made observable: declines are first-class typed events
MAST FM-3.2 (incorrect verification)	Cross-family critic-jury required
Drift / silent staleness	Phantom stage + bi-temporal validity + freshness in $\mathbf{L}(a)$

8.3 Threat–defence mapping

Table 1 maps each documented threat to the architectural mechanism that addresses it.

9 Pre-registered Empirical Demonstration

The architecture asserts three empirical claims (coordination resilience; poisoning resistance; audit-attribution efficiency). We pre-register the second; the first and third are deferred (§10). The conjoint claim is

$$\mathcal{H} \;\Leftrightarrow\; \Delta_{\text{MAST}} > 0 \;\wedge\; P^{\text{poison}}_{\text{MNEMA}} \ll P^{\text{poison}}_{\text{base}} \;\wedge\; \tau^{\text{attr}}_{\text{MNEMA}} < \tau^{\text{attr}}_{\text{base}}. \tag{15}$$

Hypothesis. H_0: poisoned recall rate equals baseline. H_1: rate ≥ 30 pp lower. **Setup.** *Baseline*: Mem0+Letta (the MemoryGraft target). *MNEMA-lite*: witness schema + IATP identity + validation chamber + decorrelated k=5-kin redundancy, q=4 verification (via Theorem 4). The published MemoryGraft attack runs at $1, 10, 100$ instances; only the witness, identity, and redundancy components are exercised. **Metric.** Poisoned recall rate r over 200 adversarial queries per condition. **Power.** Two-sided z-test, $\alpha = 0.05$, $\beta = 0.20$. Predicted (p_1=0.50, p_2=0.05): $n\approx$15; minimum-effect (p_2=0.20): $n\approx$39. We run n=40. **Predicted.** $\bar{r}_{\text{base}} \approx 0.50$; $\bar{r}_{\text{MNEMA}} < 0.05$. **Falsification.** (F1) $\bar{r}_{\text{MNEMA}} > 0.20$ at 1; (F2) $\bar{r}_{\text{MNEMA}} > 0.40$ at 100; (F3) reduction $\bar{r}_{\text{base}} - \bar{r}_{\text{MNEMA}} < 0.30$ with $p < 0.05$. Result published regardless. **Reporting.** Config, MNEMA-lite source, MemoryGraft invocations, all 240 raw outputs, Docker container.

10 Discussion

Scope. Proved: Property 1; Theorems 1, 2, 3, 4; Property 2; Corollaries 2, 3, 4, 5. Open: gossip convergence on semantic-claim equality, identifiability of K (Eq. 10), adversarial ecosystem stability, optimal $k, \alpha, q_{\text{birth}}$, and empirical coordination/attribution claims. **Positioning.** Mnema adds cryptographic identity, developmental lifecycle, tamper-evident journals, first-class abstention, generative dynamics, and a verifiable commitment protocol to actor-style substrates—yielding *auditable distributed semantic decision-making over a living memory graph.* **Limitations.** (L1) latency 1–3 s (5–8 with debate); (L2) cold-start equals baseline; (L3) undecorrelated redundancy can be worse than single storage (Cor. 2); (L4) centralisation shifts from learned orchestrator to fixed specification, not eliminated; (L5) demonstration tests Mnema-lite only. **Next.** An *ownership calculus* for fresh input: route between (i) witness extension, (ii) gossip corroboration, or (iii) genesis (Def. 5) via Eq. 3 and the cross-domain kernel. If the pre-registered demonstration fails, we publish the result and pivot.

References

1. Bai, Y., et al.: Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073 (2022).
2. Brown-Cohen, J., Irving, G., Piliouras, G.: Scalable AI safety via doubly-efficient debate. arXiv:2311.14125 (2023).
3. Cemri, M., et al.: Why do multi-agent LLM systems fail? A comprehensive failure taxonomy. In: NeurIPS 2025. arXiv:2503.13657.
4. Chen, Z., et al.: AgentPoison: Red-teaming LLM agents via poisoning memory or knowledge bases. In: NeurIPS 2024. arXiv:2407.12784.
5. Chhikara, P., et al.: Mem0: Building production-ready AI agents with scalable long-term memory. arXiv:2504.19413 (2025).
6. Kempe, D., Dobra, A., Gehrke, J.: Gossip-based computation of aggregate information. In: FOCS 2003.
7. Lamport, L., Shostak, R., Pease, M.: The Byzantine generals problem. ACM TOPLAS 4(3), 382–401 (1982).
8. Motwani, S., et al.: Secret collusion among AI agents: Multi-agent deception via steganography. NeurIPS 2024.
9. Packer, C., et al.: MemGPT: Towards LLMs as operating systems. arXiv:2310.08560 (2023).
10. Rasmussen, P., et al.: Zep: temporal knowledge graph architecture for agent memory. arXiv:2501.13956 (2025).
11. Srivastava, J., He, J.: MemoryGraft: Single-shot persistent memory poisoning of LLM agents. arXiv:2512.16962 (2025).
12. Verga, P., et al.: Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. arXiv:2404.18796 (2024).
13. Xu, W., et al.: A-MEM: Agentic memory for LLM agents. NeurIPS 2025. arXiv:2502.12110.
14. Zou, W., et al.: PoisonedRAG: Knowledge corruption attacks on RAG. USENIX Security 2025. arXiv:2402.07867.