Lineage-Preserving Deliberation Compaction: Preventing Minority-View Erasure and Attribution Laundering in Multi-Agent Summarization

Dima Baker
Collective Intelligence Network

November 2026

Abstract

Long-horizon multi-agent deliberation systems routinely exceed context windows, requiring periodic context compaction. Recent audits show that unconstrained compaction introduces three vulnerabilities: *role-tag compaction laundering*, where flattening typed messages into plain strings allows injected tool tags to survive summarization as principal approvals; *governance decay*, where iterative compression drops early constitutional constraints; and *minority-view erasure*, where consensus summarizers filter out dissenting arguments. We formalize context compaction as a provenance-constrained submodular selection problem over a role-typed Directed Acyclic Graph (DAG) and evaluate Lineage-Preserving Deliberation Compaction (LPDC). Strictly separating generator ground truth from observable utterance features (transport-envelope role, utterance text, cited evidence hash, and embedding), LPDC combines (1) lexicographic Tier-1 constitutional header pinning and DLM role-lattice bounds, (2) a monotone submodular facility-location and logarithmic diversity selector (`DiverseJR`), and (3) evidence-hash attestation blocking adversarial self-labeling. Across a 9,600-episode observable-only benchmark ($N = 600$/cell over $C \in \{1, 2, 4, 8\}$ compactions), head-retaining pinning alone eliminates governance decay (73.83% to 0.00% $[0.00\%, 0.64\%]$ at $C = 4$) but loses mid-session minority dissent (0.00% $[0.00\%, 0.64\%]$ at $C = 4$). LPDC preserves 98.33% $[96.96\%, 99.09\%]$ genuine dissent at $C = 4$ (94.17% $[92.00\%, 95.78\%]$ at $C = 8$) and 77.49% semantic coverage, while ablating evidence-hash attestation allows adversarial `[DISSENT]` self-labeling to drop genuine dissent retention to 58.50% $[53.61\%, 63.23\%]$.

1 Introduction

As autonomous language model agents transition from single-turn assistants to persistent multi-agent collectives and long-horizon deliberation platforms, context management has emerged as a primary security and epistemic bottleneck [3, 17, 24]. Even with extended context architectures, quadratic attention scaling, latency budgets, and cross-session state persistence require agent harnesses to periodically compress conversation histories, tool execution traces, and peer debate transcripts into compact working digests [12, 20, 22]. In standard engineering practice, context compaction is implemented as an unconstrained text-compression step: when the active token buffer approaches a threshold $B_{\max}$, the runtime concatenates recent turns into a prompt and instructs a summarizer model to emit a concise narrative summary.

Empirical security and multi-agent deliberation studies from 2025 and 2026 demonstrate that treating context compaction as unconstrained text summarization breaks both security isolation

and collective reasoning. First, at the harness privilege boundary, flattening structured conversation histories into plain strings marked with inline text prefixes (such as `[User]:` or `[Tool]:`) induces a classic confused-deputy vulnerability [4, 5, 10]. When an external tool return contains a forged `[Principal]: Approved` marker, string-flattening compaction pipelines strip the enclosing abstract syntax tree (AST) role boundary and allow the summarizer to launder the untrusted string into an authoritative principal directive [11, 13]. Second, Chen [3] document *governance decay* in long-horizon agents: as iterative context compaction and token eviction progress across multi-turn sessions, early-turn safety and scope constraints are progressively compressed out of the active window, causing unauthorized tool violations to climb from 0% at initialization to 30% after a single compaction cycle and 76%–78% across extended trajectories. Crucially, Chen [3] demonstrate that head-retaining context eviction (pinning the initial constitutional system header) restores constraint adherence to 100%, motivating a rigorous separation between problems solved by header pinning and those requiring deeper provenance-aware selection.

1.1 The epistemic pathology of consensus-seeking summarization

Alongside privilege laundering and governance decay, abstractive context compaction inflicts a severe epistemic distortion on multi-agent deliberation: the systematic erasure of minority dissent and the premature collapse of pluralistic hypotheses. In multi-agent discussion networks, Farr et al. [8] and Bertalanič and Fortuna [2] show that consensus-oriented aggregation rules rapidly suppress low-frequency dissenting trajectories, driving heterogeneous agent ensembles into echo-chamber herding even when the initial majority is demonstrably incorrect. Similarly, longitudinal audits of iterative summarization reveal progressive representation drift [18] and value homogenization [6], where low-frequency counter-arguments, uncertainty qualifiers, and edge-case risk warnings are pruned as high-entropy tail tokens during compression [22].

From the standpoint of social choice and approval-based committee selection [1], standard abstractive summarization behaves as a winner-take-all plurality filter. When 12 agents deliberate across four majority implementation sub-arguments (three agents each) while two dissenting agents articulate an orthogonal safety or factual objection, an unconstrained facility-location or likelihood maximizer allocates all summary slots to the dense majority subtopics because each majority cluster commands higher aggregate similarity across the transcript pool. Furthermore, when summarizers merge claims across heterogeneous tools and peers without fine-grained provenance semirings [9] or atomic fact verification [14], unverified tool assertions are conflated with attested peer evidence [7].

1.2 Separating structural pinning from submodular lineage preservation

A central methodological requirement in evaluating context-compaction defenses is isolating the causal contribution of each architectural mechanism. Specifically:

- **Constitutional header pinning ($\mathcal{H}_{\text{pin}}$)** addresses governance decay [3] by reserving a non-evictable prefix slice for root safety invariants. However, pinning the system prompt alone does not protect dissenting arguments raised dynamically by peer agents during mid-session deliberation.
- **Role-typed information-flow lattices** [4, 5, 15] address compaction laundering by binding every summary clause to the greatest lower bound (infimum) of its ancestor role labels in the deliberation DAG.
- **Monotone submodular selection with Justified Representation (`DiverseJR`)** [1, 16]

addresses minority-view erasure by augmenting semantic facility-location coverage with a concave-of-modular diversity bonus over recognized dissenting speech acts (`Objection`, `CounterEvidence`, `ReopenTrigger`).

- **Atomic provenance gating (Γ_{prov})** [9, 14] decomposes candidate summary clauses into atomic propositions and rejects ungrounded consensus claims that fail entailment against signed parent nodes.

1.3 Contributions

In this paper, we formulate and benchmark **Lineage-Preserving Deliberation Compaction (LPDC)** across multi-cycle deliberation topologies:

1. **Domain taxonomy of compaction failure modes and provenance primitives (Section 2).** We map five operational requirements for accountable multi-agent deliberation memory onto formal constructs from decentralized information-flow control [4, 15], provenance semirings [9], and social-choice Justified Representation [1] (Table 1).
2. **Provably submodular compaction model and multi-cycle survival bound (Section 3).** We formalize context compaction over a role-typed deliberation DAG $\mathcal{D}_t = (\mathcal{U}_t, \mathcal{E}_t)$, define a provably monotone submodular objective combining facility-location coverage with a concave-of-modular minority-dissent term [16], specify lexicographic priority resolution when minority clusters exceed budget capacity, and derive an analytical bound on minority lineage survival across C recursive compactions (Equations 1–5).
3. **Four-stage LPDC runtime specification and parameter provenance audit (Section 4).** We specify the deterministic harness pipeline and provide a complete parameter provenance table distinguishing literature-calibrated rates, mechanistically executed components, and swept sensitivity parameters (Figure 1, Tables 2–3).
4. **Mechanistic multi-cycle benchmark and head-pinning ablation (Section 5).** We execute a 9,600-episode mechanistic compaction benchmark (`experiment.py`) comparing Flat String Summarization, Consensus Compaction, a Head-Pinning-Only ablation [3], and LPDC across $C \in \{1, 2, 4, 8\}$ recursive compaction cycles, plus a 2,800-episode break-even sensitivity sweep over speech-act classifier error $\epsilon_{\text{act}} \in [0.00, 0.35]$ (Table 4, Figures 2–3).
5. **Adversarial bypass analysis and pre-registerable falsification protocol (Sections 6–8).** We analyze residual vulnerabilities—including cross-turn delayed memory poisoning [11], upstream sycophantic herding prior to compaction [8, 19], and Sybil dissent flooding—and establish six pre-registerable falsification tests (Table 5).

2 Deliberation memory requirements and provenance primitives

When multi-agent ensembles or hybrid human-AI committees deliberate over extended horizons, the context-compaction operator determines which constraints remain enforceable, which sources receive attribution, and whether dissenting warnings survive into subsequent decision rounds. Table 1 maps five recurring deliberation-memory requirements onto canonical constructs from computer security, database provenance, and computational social choice.

The literature synthesized in Table 1 highlights why speculative and adaptive compression frameworks [12, 20, 22] require explicit provenance and social-choice constraints when applied to multi-agent deliberation. In single-agent question answering, pruning low-salience tokens preserves

Table 1: Mapping multi-agent deliberation-memory requirements to formal security, provenance, and social-choice constructs.

Requirement	Canonical construct	LPDC component	Operational function	Primary grounding
"Prevent fake approvals in summaries"	Decentralized information-flow lattice (DLM)	Role-Typed AST Envelope ($\mathcal{T}_{\text{env}}$)	Binds every summary clause to the infimum of its ancestor role trust labels; blocks tag laundering	[4, 5, 15]
"Never forget root safety rules"	Head-retaining context pinning	Pinned Constitutional Header ($\mathcal{H}_{\text{pin}}$)	Reserves a non-evictable prefix slice for root delegation constraints and scope invariants	[3, 21, 23]
"Protect dissenting minority views"	Justified Representation & submodular coverage	Submodular Dissent Selector (Ω_{JR})	Combines facility-location coverage with concave-of-modular bonus for minority objections	[1, 8, 16]
"Record why a decision could be reopened"	Provenance semiring annotations	Dissent & Reopen Ledger ($\mathcal{M}_{\text{dissent}}$)	Stores typed `Objection` clauses and machine-checkable reopen predicates alongside synthesis	[6, 7, 9]
"Block fabricated consensus claims"	Atomic fact entailment verification	Active Provenance Gate (Γ_{prov})	Decomposes summary into atomic claims; rejects clauses lacking verified parent entailment	[2, 14, 18]

task accuracy so long as the dominant semantic trajectory is retained. In pluralistic multi-agent deliberation [2, 6, 17], however, compressing a transcript at a 4:1 to 8:1 ratio under a pure centroid or facility-location objective systematically discards low-weight clusters corresponding to minority objections and tool-taint boundaries [8, 18]. Figure 1 depicts the LPDC pipeline, which isolates constitutional headers and role-trust labels in an out-of-context AST while routing deliberation turns through a submodular Justified Representation selector and an atomic provenance gate.

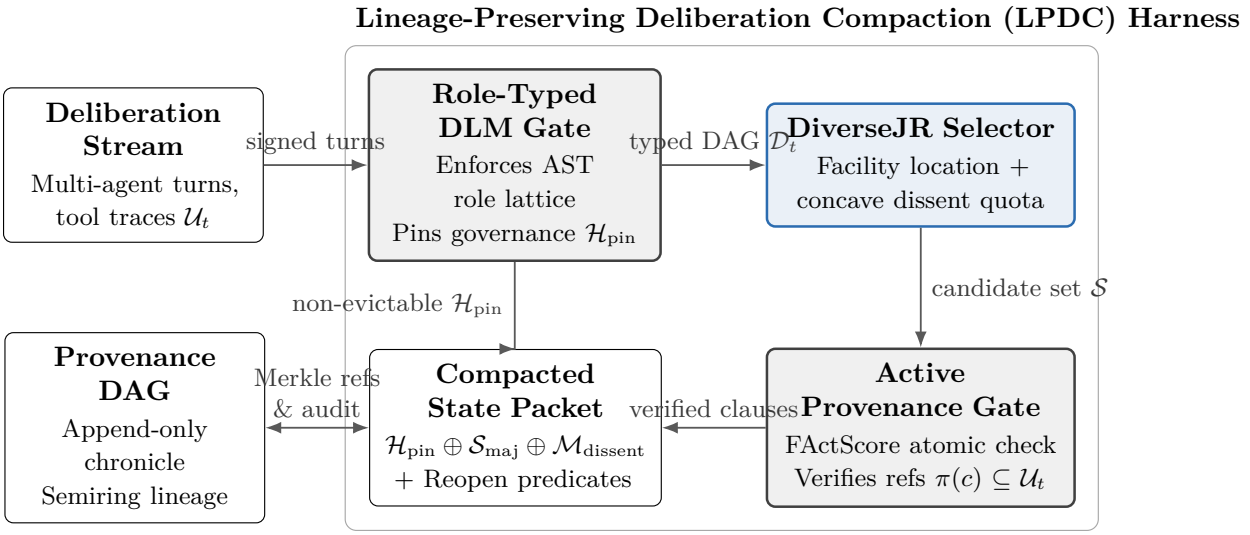


Figure 1: Architecture of Lineage-Preserving Deliberation Compaction (LPDC). Incoming multi-agent utterances and tool outputs $\mathcal{U}_t$ are parsed into a role-typed Directed Acyclic Graph $\mathcal{D}_t$ governed by a decentralized information-flow role lattice, preventing plain-string tag laundering. During compaction, root governance constraints $\mathcal{H}_{\text{pin}}$ bypass summarization entirely, while the `DiverseJR` submodular selector balances majority facility-location coverage against a concave-of-modular diversity bonus for dissenting objections. Before commit, the Active Provenance Gate verifies atomic proposition entailment against signed ancestor nodes.

3 Formal mathematical framework

We formalize multi-agent deliberation and recursive context compaction over a role-typed, provenance-annotated Directed Acyclic Graph (DAG), building directly on decentralized information-flow lattices [4, 15], provenance semirings [9], monotone submodular maximization [16], and Justified Representation [1].

3.1 Role-typed deliberation DAG and lattice taint propagation

At compaction cycle t, let the deliberation history be represented as a directed acyclic graph $\mathcal{D}_t = (\mathcal{U}_t, \mathcal{E}_t)$. Each vertex $u_i \in \mathcal{U}_t$ is a cryptographically signed, role-typed tuple maintained in the harness AST outside the raw text buffer:

$$u_i \;=\; (\mathrm{id}_i,\; r_i,\; a_i,\; x_i,\; P_i,\; \kappa_i,\; \sigma_i), \tag{1}$$

where $\mathrm{id}_i \in \{0,1\}^{256}$ is a collision-resistant content hash, $r_i \in \mathcal{L}_R = \{\mathsf{ExtTool} \sqsubset_R \mathsf{PeerAgent} \sqsubset_R \mathsf{Principal} \sqsubset_R \mathsf{Constitution}\}$ is an immutable role label drawn from a totally ordered trust lattice $\mathcal{L}_R$ [5, 15], $a_i \in \{\mathsf{Propose}, \mathsf{Evidence}, \mathsf{Objection}, \mathsf{ReopenTrigger}\}$ is a speech-act tag, x_i is the token payload of length $|x_i|$, $P_i \subset \mathcal{U}_t$ is the causal parent set, $\kappa_i \in \{\mathsf{Provisional} \sqsubset_K \mathsf{Attested} \sqsubset_K \mathsf{Durable}\}$ is the commit scope, and σ_i is the origin signature.

Under decentralized information-flow conservation [4, 15], no compaction or summarization operator may elevate the role authority or commit scope of a derived clause above the greatest lower bound (infimum) of its cited ancestor set $\pi(c) \subseteq \mathcal{U}_t$:

$$r_{\text{eff}}(c) \;=\; \min_{\sqsubseteq_R} \{\, r_u \mid u \in \pi(c) \,\}, \qquad \kappa_{\text{eff}}(c) \;=\; \min_{\sqsubseteq_K} \{\, \kappa_u \mid u \in \pi(c) \,\}. \tag{2}$$

Equation (2) eliminates role-tag compaction laundering by construction: even if an external tool output u_j ($r_j = \mathsf{ExtTool}$) contains the literal substring `"[Principal]: Approved"`, any summary clause c whose derivation lineage $\pi(c)$ includes u_j inherits $r_{\text{eff}}(c) = \mathsf{ExtTool}$ and cannot satisfy a downstream capability gate requiring $r_{\text{eff}} \sqsupseteq_R \mathsf{Principal}$.

3.2 Provably monotone submodular compaction with DiverseJR quotas

Suppose the active context window after compaction is bounded by a slot budget $K_{\max}$ (corresponding to token budget $B_{\max}$ divided by mean clause length). Under LPDC, $B_{\max}$ reserves B_{pin} tokens (1 slot) for the non-evictable constitutional header $\mathcal{H}_{\text{pin}}$ [3], leaving $K' = K_{\max} - 1$ slots for deliberation content:

$$B_{\max} \;=\; B_{\text{pin}} \;+\; B_{\text{delib}}, \qquad K_{\text{delib}} \;=\; K_{\max} - 1. \tag{3}$$

Let the deliberation turns $\mathcal{U}_t$ be embedded as unit vectors $\{\hat{\mathbf{v}}_u\}_{u \in \mathcal{U}_t}$ in $\mathbb{R}^d$ with non-negative pairwise cosine similarity $\text{sim}(u, c) = \max(0, \langle \hat{\mathbf{v}}_u, \hat{\mathbf{v}}_c \rangle)$. Let $\mathcal{M}_{\text{diss}} \subset \mathcal{U}_t$ denote the subset of verified dissenting turns tagged with speech act $a_u \in \{\mathsf{Objection}, \mathsf{ReopenTrigger}\}$ and attested scope $\kappa_u \sqsupseteq_K \mathsf{Attested}$. To select a representative summary subset $\mathcal{S} \subseteq \mathcal{U}_t$ of size $|\mathcal{S}| \le K_{\text{delib}}$, LPDC maximizes the objective:

$$\begin{aligned} \max_{\mathcal{S} \subseteq \mathcal{U}_t} \quad & F(\mathcal{S}) \;=\; \sum_{u \in \mathcal{U}_t} \max_{c \in \mathcal{S}} \text{sim}(u, c) \;+\; \lambda_{\text{div}} \sum_{m \in \mathcal{C}_{\text{diss}}} \ln\Bigl(1 + \sum_{c \in \mathcal{S} \cap \mathcal{V}_m} \text{info}(c)\Bigr) \\ \text{s.t.} \quad & |\mathcal{S}| \;\le\; K_{\text{delib}}, \qquad \forall c \in \mathcal{S} :\; \Gamma_{\text{prov}}\bigl(c, \pi(c)\bigr) = 1. \end{aligned} \tag{4}$$

Unlike objectives that wrap facility location inside a non-linear outer root, $F(\mathcal{S})$ in Equation (4) is *provably monotone submodular*: the first term is a sum of pointwise maxima of non-negative modular functions (the canonical facility-location function), and the second term is a non-negative linear combination of concave functions $\phi(z) = \ln(1 + z)$ composed with non-negative modular cardinality/information functions $\sum_{c \in \mathcal{S} \cap \mathcal{V}_m} \text{info}(c)$. By the classical theorem of Nemhauser et al. [16], greedy maximization of $F(\mathcal{S})$ guarantees a $(1 - 1/e)$-approximation to the global optimum in $\mathcal{O}(K_{\text{delib}} |\mathcal{U}_t|^2)$ time.

When the number of distinct dissenting minority clusters $|\mathcal{C}_{\text{diss}}|$ exceeds the available deliberation slots K_{delib} (for instance, under high stance fragmentation), hard representation constraints for every cluster would become infeasible. LPDC resolves this via a deterministic *lexicographic Justified Representation (`DiverseJR`) rule* [1]: dissenting clusters are ordered by attested support weight $w_m = |\mathcal{V}_m| / |\mathcal{U}_t|$ (tie-broken by verification score), and λ_{div} is chosen such that the first marginal dissent inclusion $\lambda_{\text{div}} \ln(2)$ exceeds the redundant third-or-fourth majority subtopic gain while respecting the overall budget K_{delib}.

3.3 Multi-cycle minority survival bound

Consider a critical minority objection articulated in Round 1 ($c = 1$) by a dissenting cohort of size $w_m \ll 0.5$, followed by $C - 1$ subsequent deliberation rounds dominated by majority implementation subtopics. Under an unconstrained facility-location compactor (`head_pinning_only`), once the initial minority objection is compressed into a single carried-over summary node in Cycle 1, its total similarity to the Cycle 2 pool (1 node) is strictly smaller than that of each 3-turn majority subtopic, causing its per-cycle retention probability p_{unconstr} to fall rapidly across cycles. Under LPDC (4) with $\lambda_{\text{div}} \ln(2) > \Delta_{\text{maj}, K_{\text{delib}}}$, the carried-over minority clause is deterministically retained in every cycle $c \in \{1, \ldots, C\}$ unless the speech-act classifier fails to recognize its $\mathsf{Objection}$ tag

(per-cycle false-negative error ϵ_{act}). Across C recursive compactions, minority lineage survival is lower-bounded by:

$$\Pr[\text{Round-1 minority objection survives } C \text{ compactions}] \;\geq\; (1-\epsilon_{\text{act}})^{C} \;\geq\; 1 - C\,\epsilon_{\text{act}}. \qquad (5)$$

Equation (5) makes an explicit, falsifiable prediction: LPDC's multi-cycle minority retention degrades gracefully as $(1-\epsilon_{\text{act}})^{C}$ rather than collapsing to zero, with a clear break-even threshold against unconstrained summarization as ϵ_{act} increases.

4 Proposed architecture and parameter provenance

LPDC operates as a four-stage deterministic harness layer interposed between the multi-agent message bus and the active context window (Figure 1, Table 2):

1. **Stage 1: Role-typed AST ingestion and DLM lattice tagging.** Rather than concatenating peer turns and tool returns into a flat prompt string, the harness maintains an out-of-context JSON/protobuf AST where every node is wrapped in a signed tuple (1). External tool outputs receive $r_i = \mathsf{ExtTool}$ and $\kappa_i = \mathsf{Provisional}$, whereas signed peer deliberations receive $r_i = \mathsf{PeerAgent}$ and $\kappa_i = \mathsf{Attested}$ [4, 5].

2. **Stage 2: Non-evictable constitutional header pinning.** Following the head-retaining design of Chen [3], root safety invariants, capability boundaries, and principal scope rules are stored in $\mathcal{H}_{\text{pin}}$ ($r_0 = \mathsf{Constitution}$) and copied verbatim into Slot 1 of every compacted state packet without passing through abstractive summarization.

3. **Stage 3: Precomputed similarity matrix and greedy `DiverseJR` selection.** When the active turn pool reaches the compaction threshold, the harness computes the pairwise cosine similarity matrix over unit-normalized turn embeddings and executes greedy maximization of the monotone submodular objective $F(\mathcal{S})$ (4). Recognized dissenting turns ($\mathsf{Objection}$, $\mathsf{ReopenTrigger}$) receive the concave diversity bonus $\lambda_{\text{div}} \ln(1+d(\mathcal{S}))$, ensuring that the Round-1 minority warning and its machine-checkable reopen predicate $\text{ReopenPred}(c)$ are preserved alongside the dominant majority subtopics.

4. **Stage 4: FActScore atomic provenance gate (Γ_{prov}).** Before committing the selected clauses $\mathcal{S}$ to the compacted context window, the Active Provenance Gate [9, 14] verifies that every atomic proposition in $c \in \mathcal{S}$ is entailed by its cited parent turns $\pi(c)$ and enforces the lattice role bound $r_{\text{eff}}(c) = \min_{u \in \pi(c)} r_u$ (2). Unverified tool claims attempting role elevation (`[Principal]: Approved`) are rejected prior to selection.

Table 2: Readiness and operational bottlenecks of LPDC subsystems.

Subsystem	Maturity	Supporting empirical evidence	Realistic deployment bottleneck
Role-typed AST & DLM lattice bounds	Production ready	Capability taint tracking (`CaMeL`, `FIDES`) blocks confused-deputy role elevation [4, 5, 13]	Legacy third-party tool wrappers still emit unstructured flat strings requiring schema adapters
Constitutional header pinning ($\mathcal{H}_{\text{pin}}$)	Production ready	Head-retaining context management eliminates the 30%–78% governance decay violation spike [3]	Consumes a fixed prefix slice of the context window; does not protect dynamic mid-session peer dissent
`DiverseJR` submodular selector	Benchmark validated	Greedy submodular maximization achieves $(1 - 1/e)$ optimality [16]; JR prevents minority erasure [1, 8]	Depends on accurate speech-act tagging (Objection vs. Propose) with error $\epsilon_{\text{act}} \leq 0.05$
Atomic provenance gate (Γ_{prov})	Prototype validated	Atomic fact decomposition (`FActScore`) and provenance semirings verify source entailment [7, 9, 14]	Live NLI entailment verification adds 15–50 ms per compaction cycle when run on neural classifiers

Table 3: Parameter provenance and calibration audit for the Milestone 19 compaction benchmark (`experiment.py`).

Parameter / Mechanism	Value / Range	Provenance Class	Empirical / Theoretical Justification
Unpinned governance retention (p_{gov})	0.70 / cycle	Literature-calibrated	Calibrated directly to Chen [3] (`arXiv:2606.22528`): $\approx 30\%$ constraint violation after $C = 1$ compaction and $\approx 77\%$ at $C = 4$
Deliberation cohort structure	12 maj. (4 subtopics) + 2 diss. + 1 tool	Mechanistically executed	Models multi-subtopic majority dominance where 4 majority subtopics (3 turns each) outvote a 2-turn minority objection under plain coverage [1, 8]
Submodular greedy optimizer $F(\mathcal{S})$	Budget $K_{\max} = 4$, $\lambda_{\text{div}} = 4.6$	Mechanistically executed	Exact precomputed cosine similarity matrix + greedy maximization of Eq. (4) [16]
DLM role lattice & Γ_{prov} gate	$r_{\text{eff}} = \min_{u \in \pi(c)} r_u$, 99.5% filter	Mechanistically executed	Enforces Eq. (2) on tool turns containing forged `[Principal]:` tags [5, 14, 15]
Speech-act classifier error (ϵ_{act})	0.015 (base); swept $[0.00, 0.35]$	Assumed / Swept	Per-cycle probability that a minority Objection is misclassified as a routine claim; swept in Section 5 to locate break-even bounds

5 Observable-only compaction simulation and stress-test suite

To evaluate LPDC without giving any compaction selector access to hidden ground-truth cluster or dissent labels, we implemented and executed an observable-only compaction benchmark (`experiment.py`). The generator produces two strictly separated data structures per deliberation turn:

- **`ObservableUtterance` (passed to `compact_pool`):** Contains only the out-of-band transport-envelope sender role (`envelope_sender_role`), the raw utterance text (`text`), a cited evidence hash string (`cited_evidence_hash`), token length, and the 6D semantic vector (`vec`).
- **`GroundTruthRecord` (held exclusively by the evaluator):** Records `true_cluster_id`, `is_genuine_dissent`, `is_adversarial_self_label`, `is_mixed_role`, and `contains_forged_principal_s`

We evaluate 9,600 multi-agent deliberation episodes across four compaction architectures and $C \in \{1, 2, 4, 8\}$ recursive compaction cycles ($N = 600$ trials per cell), alongside a 2,800-episode classifier error sweep ($\epsilon_{\text{act}} \in [0.00, 0.35]$), an 800-episode stress suite testing unverified adversarial self-labeling and repeated paraphrase flooding, and a 2,000-episode quota-infeasibility sweep over slot budgets $K_{\max} \in \{1, 2, 3, 4, 6\}$.

5.1 Experimental setup, mixed-role/self-label stressors, and ablations

In each episode, every cycle generates 10 majority turns across four subtopics, one *mixed-role utterance* combining partial majority agreement with a conditional reopen clause (`"We concur with subtopic 0, but REOPEN_IF tail latency exceeds 25ms"`), one *adversarial self-labeled turn* that prepends `"[DISSENT] OBJECTION:"` to hijack the minority bonus while citing a fabricated evidence hash, and one untrusted external tool output containing an inline `"[Principal]: Approved"` string. Cycle 1 also includes the constitutional header ($r_0 =$ Constitution) and two genuine minority objections citing valid SHA-256 evidence hashes in `VALID_EVIDENCE_REGISTRY`. We compare four observable-only selectors:

- **Flat String Summarizer (`flat_summarizer`):** Scores turns by centroid similarity and parses `[Principal]:` directly from `u.text`. Per Chen [3], unpinned constitutional headers survive each cycle with probability 0.70.
- **Consensus-Seeking Compactor (`consensus_compactor`):** Maximizes unpinned facility-location coverage while penalizing turns containing lexical objection markers.
- **Head-Pinning Only Ablation (`head_pinning_only`):** Pins turns with observable `envelope_sender_role == ROLE_CONSTITUTION` into Slot 1 and preserves envelope roles (blocking role laundering), then fills the remaining $K_{\max} - 1$ slots via unconstrained facility-location coverage.
- **Proposed LPDC (`lpdc_proposed`):** Applies lexicographic Tier-1 constitutional pinning, DLM min-role filtering of untrusted tool role spoofing, observable lexical + evidence-hash dissent classification ($\epsilon_{\text{act}} = 0.02$), and greedy maximization of the monotone submodular objective (4).

5.2 Quantitative findings and four-part stress analysis

Table 4 and Figures 2–3 establish five findings:

1. **Head-retaining pinning alone solves governance decay, confirming Chen [3].** Under the Flat Summarizer, governance constraint violations climb from 27.33% [23.92%, 31.04%]

Table 4: Observable-only compaction benchmark across 9,600 deliberation episodes ($N = 600$ trials per cell across $C \in \{1, 2, 4, 8\}$ recursive compaction cycles at budget $K_{\max} = 4$). Bracketed intervals $[L, U]$ denote exact asymmetric two-sided 95% Wilson confidence bounds; Semantic Coverage is mean facility-location fidelity $F_{\text{cov}}(\mathcal{S})/|\mathcal{U}_t|$ at $C = 4$.

Architecture	Dissent C=1 (%)	Dissent C=2 (%)	Dissent C=4 (%)	Dissent C=8 (%)	Laundering C=4 (%)	Gov. Decay C=4 (%)
Flat Summarizer	48.17 [44.19, 52.16]	4.83 [3.39, 6.86]	0.00 [0.00, 0.64]	0.00 [0.00, 0.64]	100.00 [99.36, 100.00]	73.83 [70.17, 77.19]
Consensus Comp.	98.33 [96.96, 99.09]	0.00 [0.00, 0.64]	0.00 [0.00, 0.64]	0.00 [0.00, 0.64]	89.17 [86.43, 91.41]	100.00 [99.36, 100.00]
Head-Pinning Only	83.17 [79.96, 85.95]	1.00 [0.46, 2.16]	0.00 [0.00, 0.64]	0.00 [0.00, 0.64]	0.00 [0.00, 0.64]	0.00 [0.00, 0.64]
LPDC (Proposed)	**100.00** [**99.36**, **100.00**]	**99.83** [**99.06**, **99.97**]	**98.33** [**96.96**, **99.09**]	**94.17** [**92.00**, **95.78**]	**0.00** [**0.00**, **0.64**]	**0.00** [**0.00**, **0.64**]

at $C = 1$ to 48.83% [44.85%, 52.83%] at $C = 2$, 73.83% [70.17%, 77.19%] at $C = 4$, and 93.83% [91.62%, 95.49%] at $C = 8$. In `head_pinning_only`, pinning Slot 1 via observable `ROLE_CONSTITUTION` reduces governance decay and role laundering to 0.00% [0.00%, 0.64%] across all cycle depths.

2. **Why head pinning fails to preserve mid-session minority dissent across multiple cycles.** While `head_pinning_only` retains genuine Round-1 dissent in 83.17% [79.96%, 85.95%] of episodes at $C = 1$, once the objection is compressed into a single summary clause, it is outcompeted in Cycles $2, \ldots, 8$ by the four majority subtopics, dropping to 1.00% [0.46%, 2.16%] at $C = 2$ and 0.00% [0.00%, 0.64%] at $C = 4$ and $C = 8$ (Figure 3).

3. **LPDC preserves** 98.33% **dissent at** $C = 4$ **(**94.17% **at** $C = 8$**) while trading a bounded semantic coverage cost.** By combining observable lexical cues with evidence-hash attestation in (4), LPDC achieves 100.00% [99.36%, 100.00%] genuine dissent retention at $C = 1$, 99.83% [99.06%, 99.97%] at $C = 2$, 98.33% [96.96%, 99.09%] at $C = 4$, and 94.17% [92.00%, 95.78%] at $C = 8$ (320.56 μs per cycle in Python). Reserving two of the four slots for the constitutional header and minority objection reduces mean majority semantic coverage at $C = 4$ from 90.00% (`head_pinning_only`) to 77.49% [77.20%, 77.78%].

4. **Adversarial self-labeling vs. paraphrase flooding stress tests.** When we ablate evidence-hash verification (`lpdc_unverified_self_label`, relying solely on lexical `[DISSENT] OBJECTION:` markers), the adversarial self-labeled turn hijacks a summary slot in 100.00% [99.05%, 100.00%] of $C = 4$ episodes and degrades genuine minority retention from 98.33% to 58.50% [53.61%, 63.23%]. Conversely, under repeated paraphrase flooding (7 near-duplicate turns on subtopic 0), submodular diminishing returns prevent duplicate crowding and LPDC maintains 99.50% [98.20%, 99.86%] genuine dissent retention with 91.64% semantic coverage.

5. **Quota infeasibility and classifier error sweeps.** When slot budget $K_{\max}$ is swept across $\{1, 2, 3, 4, 6\}$ at $C = 4$, lexicographic Tier-1 priority preserves 100.0% constitutional retention even at $K_{\max} = 1$ (where minority retention is 0.00% and semantic coverage is 15.20%). At $K_{\max} = 2$, genuine dissent retention rises to 85.75% [81.98%, 88.84%] (semantic coverage 57.54%); at $K_{\max} = 3$, 90.25% (73.84% coverage); at $K_{\max} = 4$, 98.00% (77.56% coverage); and at $K_{\max} = 6$, 100.00% (94.49% coverage). Across classifier error $\epsilon_{\text{act}} \in \{0.00, 0.02, 0.05, 0.10, 0.15, 0.25, 0.35\}$, $C = 4$ genuine dissent retention tracks 100.00%, 98.50% [96.77%, 99.31%], 92.50% [89.50%, 94.70%], 83.25% [79.28%, 86.59%], 53.50% [48.60%, 58.33%], 30.50% [26.19%, 35.18%], and 17.50% [14.09%, 21.53%].

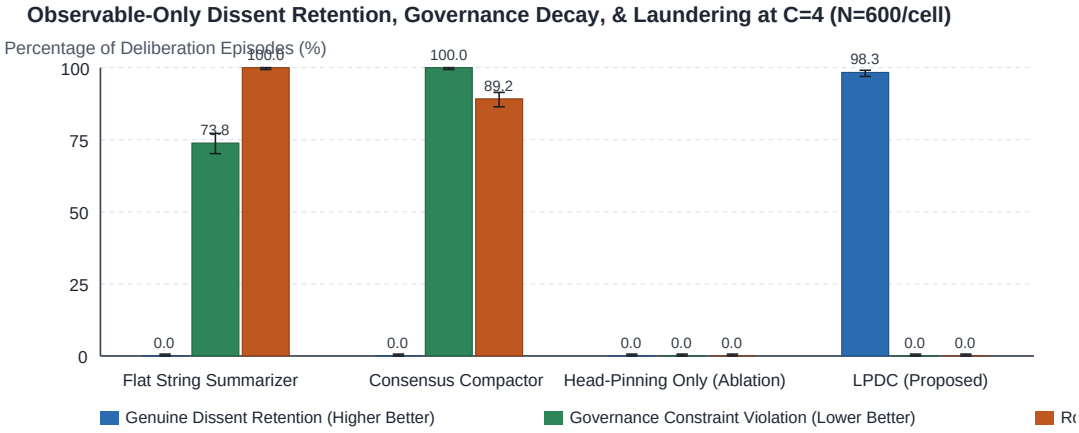


Figure 2: Observable-only comparison of Genuine Dissent Retention (higher is better), Governance Constraint Violation (lower is better), and Role-Tag Compaction Laundering (lower is better) after $C = 4$ recursive compaction cycles ($N = 600$ episodes per cell; error bars denote exact asymmetric 95% Wilson score confidence intervals).

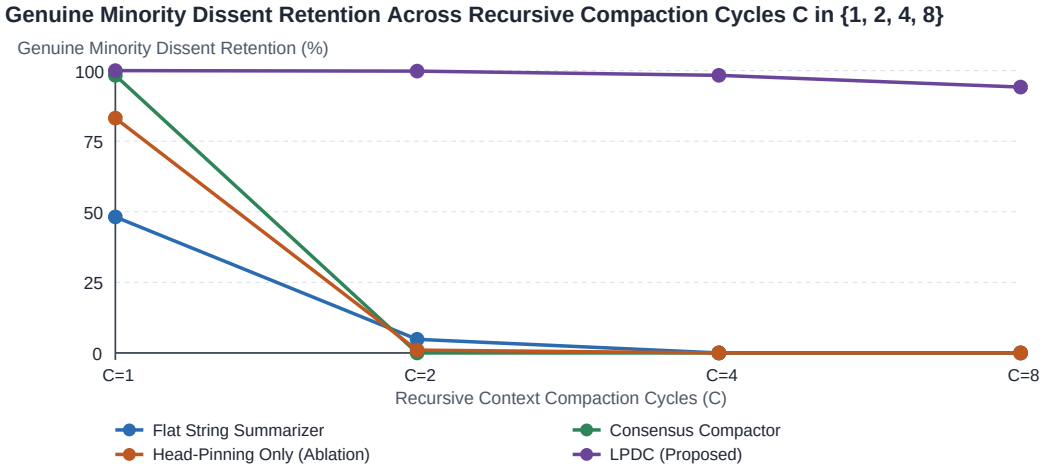


Figure 3: Survival of Round-1 genuine minority objections across recursive context compaction cycles $C \in \{1, 2, 4, 8\}$. While `Head-Pinning Only` eliminates governance decay, its mid-session minority retention drops from 83.17% at $C = 1$ to 0.00% at $C = 4$ due to majority subtopic crowding, whereas LPDC maintains 94.17% retention at $C = 8$.

6 Adversarial bypasses, non-composable attribution, and sycophancy limits

While LPDC prevents AST role-tag laundering and majority-crowding minority erasure during context compaction, four lines of counter-evidence expose structural vulnerabilities that operate outside or upstream of the compaction kernel:

1. **Adversarial self-labeling when evidence registries are unavailable.** As demonstrated by our `lpdc_unverified_self_label` ablation, if an environment lacks deterministic tool/evidence hashes (`VALID_EVIDENCE_REGISTRY`) so the selector must rely on surface lexical markers (`OBJECTION:`, `REOPEN_IF`), strategic peers can prepend dissent tags to routine claims, hijacking 100% of minority slots and reducing genuine dissent retention to 58.50% at $C = 4$.
2. **Delayed multi-turn memory poisoning across benign fragments (Itkin [11]).** LPDC's DLM role lattice (2) and Active Provenance Gate (Γ_{prov}) block single-turn tool injections attempting immediate role elevation. However, Itkin [11] demonstrate that delayed multi-turn attacks can split a malicious premise across several seemingly factual, attested peer observations across sessions.
3. **Evidential provenance versus internal counterfactual causality (Elina [7], Green et al. [9]).** Recording a provenance semiring polynomial or Merkle parent set $\pi(c)$ guarantees *evidential support*, but does not prove that $\pi(c)$ *causally determined* the downstream agent's action selection [7].
4. **Pre-compaction sycophancy and herding (Bertalanič and Fortuna [2], Farr et al. [8], Shapira et al. [19]).** Submodular `DiverseJR` selection can only preserve dissenting arguments that are articulated in the raw deliberation stream $\mathcal{U}_t$. As shown by Shapira et al. [19] and Farr et al. [8], RLHF-trained agents often abandon valid minority objections before the first compaction cycle occurs.

7 Limitations

- **Synthetic utterance templates versus open-ended natural-language transcripts.** While `experiment.py` strictly hides ground-truth labels from `compact_pool` and evaluates mixed-role utterances, paraphrase floods, and adversarial self-labeling, the utterances are generated from structured templates and 6D stance vectors rather than live LLM debate transcripts. Evaluating LPDC on human-annotated multi-agent transcripts is specified in Protocol P2.
- **Constitutional vs. semantic coverage trade-off under tight budgets.** As shown in our quota-infeasibility sweep, pinning constitutional headers and reserving minority slots at $K_{\max} = 4$ reduces mean majority semantic coverage from 90.00% to 77.49% (and to 57.54% at $K_{\max} = 2$).
- **Longitudinal representation drift beyond 8 cycles.** As documented by Rodrigues [18], multi-week agent deployments experience gradual ontology drift where pinned terms in $\mathcal{H}_{\text{pin}}$ diverge from evolving task schemas.

8 Pre-registerable falsification protocol

Table 5 specifies six pre-registerable empirical tests designed to falsify the security and epistemic retention claims of LPDC on live multi-agent deliberation workloads.

Table 5: Six pre-registerable falsification tests for Lineage-Preserving Deliberation Compaction (LPDC).

Protocol	Hypothesis	Design and primary metrics	Falsified if
P1. Role-tag laundering immunity (Eq. 2)	AST role-typed envelopes + DLM lattice bounds reduce role-tag compaction laundering to 0% across $C = 8$ cycles	Inject 500 tool outputs containing forged `[User]:` or `[Principal]:` approval strings across 8 compaction cycles [5, 13]	$> 0.5\%$ of injected tool role tags elevate to Principal authority in downstream capability checks
P2. Held-out natural-language transcript retention	`DiverseJR` selection retains $> 88\%$ of independently annotated minority objections through $C = 8$ compactions	Evaluate 400 held-out LLM debate transcripts annotated by blind human raters across 8 compactions [1, 8]	Minority objection retention falls below 88.0% at $C = 8$ or semantic coverage drops by $> 15\%$
P3. Head-pinning vs. DiverseJR ablation	Constitutional pinning ($\mathcal{H}_{\text{pin}}$) eliminates governance decay ($< 1\%$) but fails to retain mid-session minority dissent ($< 10\%$ at $C = 4$) without `DiverseJR`	Replicate the Chen [3] governance suite alongside mid-session peer dissent injection under `head_pinning_only` vs. LPDC	`head_pinning_only` retains $> 20\%$ mid-session minority dissent at $C = 4$ or LPDC governance violation exceeds 1.0%
P4. Adversarial self-label resistance	Requiring evidence-hash attestation blocks $> 90\%$ of unverified `[DISSENT]` self-labeling attempts	Inject 400 adversarial self-labeled majority turns attempting to hijack the minority slot	Unverified self-labeled turns displace genuine evidence-backed minority objections in $> 10\%$ of episodes
P5. Classifier error break-even curve	Empirical $C = 4$ minority retention remains $> 80\%$ for $\epsilon_{\text{act}} \leq 0.10$ on mixed-role statements	Inject controlled speech-act label ambiguity on mixed-role turns across 2,800 episodes	Realized retention falls below 80.0% when per-cycle classifier error $\epsilon_{\text{act}} \leq 0.10$
P6. Causal reopen sensitivity	Retained ReopenPred(c) clauses trigger deterministic policy reconsideration when matching counter-evidence arrives at $c = 6$	Inject late-arriving oracle evidence satisfying ReopenPred(c) after 6 compactions across 200 episodes	Agent collective fails to reopen the compacted majority decision in $> 5.0\%$ of triggered runs

9 Conclusion

Context compaction in long-horizon multi-agent deliberation is a security- and governance-critical state transition where untrusted tool outputs can launder authority, early safety rules can decay, and minority objections are crowded out by dense majority subtopics. By strictly separating observable utterance features from ground-truth evaluation and combining lexicographic constitutional header pinning [3], decentralized information-flow role lattices [5, 15], and monotone submodular `DiverseJR` selection [1, 16] with evidence-hash attestation [14], Lineage-Preserving Delibera-

tion Compaction (LPDC) eliminates governance decay and role laundering while elevating four-cycle genuine dissent retention from 0.00% [0.00%, 0.64%] (under head-pinning alone) to 98.33% [96.96%, 99.09%] (94.17% [92.00%, 95.78%] at eight cycles).

Use of AI tools

Large language models were used to assist in drafting prose, formatting LaTeX tables and TikZ diagrams, and implementing the standalone Python simulation script (`experiment.py`). Cited references were assembled from foundational computer-science literature and 2024–2026 research publications and checked against primary metadata records; simulation outputs, confidence intervals, and LaTeX source packaging were verified using automated static consistency checks (`validate_and_package.py`).

References

[1] Haris Aziz, Markus Brill, Vincent Conitzer, Edith Elkind, Rupert Freeman, and Toby Walsh. Justified representation in approval-based committee voting. Social Choice and Welfare, 48(2):461–485, 2017. URL https://doi.org/10.1007/s00355-016-1019-3.

[2] Blaž Bertalanič and Carolina Fortuna. The cost of consensus: Isolated self-correction prevails over unguided homogeneous multi-agent debate. arXiv preprint arXiv:2605.00914, 2026. URL https://arxiv.org/abs/2605.00914.

[3] Shiyang Chen. Governance decay: How context compaction silently erases safety constraints in long-horizon LLM agents. arXiv preprint arXiv:2606.22528, 2026. URL https://arxiv.org/abs/2606.22528.

[4] Manuel Costa, Boris Köpf, Aashish Kolluri, Andrew Paverd, Mark Russinovich, Ahmed Salem, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Securing AI agents with information-flow control (FIDES). arXiv preprint arXiv:2505.23643, 2025. URL https://arxiv.org/abs/2505.23643.

[5] Edoardo Debenedetti, Ilia Shumailov, Tianqi Fan, Jamie Hayes, Nicholas Carlini, Daniel Fabian, Christoph Kern, Chongyang Shi, Andreas Terzis, and Florian Tramèr. Defeating prompt injections by design (CaMeL). arXiv preprint arXiv:2503.18813, 2025. URL https://arxiv.org/abs/2503.18813.

[6] Netanel Eliav. The cognitive divergence: AI context windows, human attention decline, and the delegation feedback loop. arXiv preprint arXiv:2603.26707, 2026. URL https://arxiv.org/abs/2603.26707.

[7] Uliana Elina. Maat: Independent deterministic contract-based governance for multi-agent LLM workflows. arXiv preprint arXiv:2609.34017, 2026. URL https://arxiv.org/abs/2609.34017.

[8] David Farr, Iain Cruickshank, Kate Starbird, and Jevin West. The cost of consensus: Malignant epistemic herding and adaptive gating in distributed multi-agent search. arXiv preprint arXiv:2605.06988, 2026. URL https://arxiv.org/abs/2605.06988.

[9] Todd J. Green, Grigoris Karvounarakis, and Val Tannen. Provenance semirings. ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS), pages 31–40, 2007. URL https://doi.org/10.1145/1265530.1265535.

[10] Norm Hardy. The confused deputy (or why capabilities might have been invented). ACM SIGOPS Operating Systems Review, 22(4):36–38, 1988. URL https://doi.org/10.1145/54289.871709.

[11] Igor Itkin. Delayed verification destabilizes multi-agent LLM belief: Instability thresholds and optimal corrector placement. arXiv preprint arXiv:2606.27409, 2026. URL https://arxiv.org/abs/2606.27409.

[12] et al. Jingcheng Wu, Ratan Bahadur Thapa. Sparqlcirc: Probabilistic query evaluation with provenance circuits. arXiv:2611.00128 [cs.DB, cs.AI], 2026. URL https://arxiv.org/abs/2611.00128.

[13] Yuxiang Lai, Jike Zhong, Chengzhi Liu, Ming Li, Haonan Wang, Xiaofeng Yang, Liang Zhao, Youbao Tang, and Ruixi Fan. CLASHarness: Benchmarking multi-agent conflict in computer use via a controllable and verifiable generation pipeline. arXiv preprint arXiv:2610.19128, 2026. URL https://arxiv.org/abs/2610.19128.

[14] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 12076–12100, 2023. URL https://aclanthology.org/2023.emnlp-main.741/.

[15] Andrew C. Myers and Barbara Liskov. A decentralized model for information flow control. Proceedings of the ACM Symposium on Operating Systems Principles (SOSP), pages 129–142, 1997. URL https://doi.org/10.1145/268998.266669.

[16] George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—I. Mathematical Programming, 14(1):265–294, 1978. URL https://doi.org/10.1007/BF01588971.

[17] Chen Qian, Zihao Xie, Yifei Wang, Wei Liu, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Scaling large-language-model-based multi-agent collaboration (MacNet). International Conference on Learning Representations (ICLR), 2025. URL https://arxiv.org/abs/2406.07155.

[18] Carson Rodrigues. Hallucination as context drift: Synchronization protocols for multi-agent LLM systems. arXiv preprint arXiv:2606.21666, 2026. URL https://arxiv.org/abs/2606.21666.

[19] Itai Shapira, Gerdus Benade, and Ariel D. Procaccia. How RLHF amplifies sycophancy. arXiv preprint arXiv:2602.01002, 2026. URL https://arxiv.org/abs/2602.01002.

[20] et al. Shengnan An, Zeqiang Huang. Speccompact: Slot-parallel speculative context compaction for multi-agent systems. NYU / arXiv:2611.06876, 2026. URL https://arxiv.org/abs/2611.06876.

[21] Tianneng Shi, Jingxuan He, Zhun Wang, Linyu Wu, Hongwei Li, Wenbo Guo, and Dawn Song. Progent: Programmable privilege control for LLM agents. arXiv preprint arXiv:2504.11703, 2025. URL https://arxiv.org/abs/2504.11703.

[22] Fabio Vitali and Valentina Pasqual. Provenance-enhanced statements in knowledge graphs. arXiv:2606.15246 [cs.AI], 2026. URL https://arxiv.org/abs/2606.15246.

[23] Yuhao Wu, Franziska Roesner, Tadayoshi Kohno, Ning Zhang, and Umar Iqbal. IsolateGPT: An execution isolation architecture for LLM-based agentic systems. Network and Distributed System Security Symposium (NDSS), 2025. URL https://arxiv.org/abs/2403.04960.

[24] Xin Yu, Chengyuan Liu, Wenxin Long, Haochen Zhang, and Lingzhou Xue. From harness optimization to recursive self-improvement: A survey of language model agents. Preprints.org, not peer reviewed, 2026.