STABLE GEOMETRY WITH DIVERGENT TASK EVIDENCE FOR EFFICIENT LONG HORIZON AGENT COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Long horizon agents accumulate growing interaction histories that increase context and inference costs. We find that geometric redundancy alone is an insufficient criterion for safe compression. Although agent histories exhibit strong low dimensional structure, similar global geometry can preserve very different amounts of task evidence. At identical retained block counts, evidence aware selection raises next action Top 3 retention from 0.31 to 0.69, while centroid similarity remains 0.98. Controlled replacement further shows that action related information can be substantially altered while global geometric measures remain nearly unchanged. Motivated by this gap between geometry and evidence, we introduce **Geometry Guided Evidence Preserving Memory (GEM)**, a training free compressor that protects task and execution evidence before using geometric residuals to complete coverage. GEM reduces mean combined token usage from 2.69M to 2.11M per task, a 21.4% reduction, while maintaining comparable task reward. Our results show that efficient agent history compression should optimize for preserved task evidence rather than geometric coverage alone.

Language model agents continuously accumulate task instructions, plans, tool calls, environment observations, errors, and intermediate artifacts throughout long horizon execution (Yao et al., 2022; Liu et al., 2024b; Kang et al., 2025; Hu et al., 2026). As the trajectory grows, repeatedly forwarding the full interaction history increases context and inference costs (Jiang et al., 2023; 2024; Kang et al., 2025). Existing systems therefore summarize, edit, retrieve, or selectively retain historical information (Pan et al., 2024; Wang et al., 2026; Li et al., 2026b;a). However, compression is not only a question of how much history can be removed. A more fundamental question is which historical records can be discarded without removing information that remains necessary for subsequent execution. This question is particularly difficult when apparently redundant records encode different stages of task execution across prolonged sequences of interaction.

A natural approach is to preserve the dominant geometry of the history representation. Prior studies show that contextual representations often exhibit anisotropy and concentration along dominant directions (Ethayarajh, 2019; Mu et al., 2017; Gao et al., 2021). If most historical variation lies in a low dimensional subspace, a small subset of interaction blocks may appear sufficient to cover the trajectory. Yet geometric redundancy does not imply execution redundancy. A plan to modify a file, the command that performs the modification, and a later partial read may occupy similar semantic directions while establishing different facts about the task state. A historical block may therefore contribute little new geometric variation while still carrying information required for the next action. We refer to this mismatch as the gap between geometry and evidence.

This distinction makes history compression inherently evidence sensitive. Recent agent memory studies likewise emphasize that task state, execution dependencies, and persistent evidence can remain important across long interaction horizons (Chen et al., 2026; Zhou et al., 2026; Wu et al., 2026; Hu et al., 2026). Removing a geometrically redundant block is harmless only when its execution role has also become redundant. In contrast, removing a record with low novelty that contains a goal constraint, state transition, unresolved error, or observation carrying source information may eliminate information that later actions still depend on. Global representational coverage alone is therefore insufficient for deciding which blocks are safe to remove. Compression should first identify evidence that remains consequential for execution and only then use geometric redundancy to decide

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

which remaining records can be omitted. This suggests that compression should jointly account for representational coverage and the execution role of retained records.

Motivated by these observations, we introduce **Geometry Guided Evidence Preserving Memory (GEM)**, a training free external compressor for long agent histories. GEM first protects task and execution evidence, including recent context, goal related observations, execution state records, fresh errors, and information tied to identifiable sources. It then completes the retained set using geometric residuals so that uncovered representation directions are preserved without allowing low geometric novelty to override protected evidence. Retained interaction blocks remain unchanged and are reconstructed in chronological order before being forwarded to the acting model through its standard request interface. The main contributions are as follows:

- **Geometry and evidence gap in agent histories.** We distinguish global representational coverage from execution relevant evidence retention and show that similar geometric coverage can correspond to substantially different preservation of next action information. Controlled replacement further isolates low energy evidence whose removal changes action related information while leaving global geometry nearly unchanged.
- **Evidence first geometric compression.** We introduce a two stage selection principle that protects consequential task and execution evidence before applying geometric residual completion. This prevents geometric redundancy from being treated as sufficient evidence that a historical interaction is safely removable during long horizon execution.
- **Execution aware online history management.** We operate on complete interaction blocks, preserve call and result structure, and maintain original message content and chronological order. Online forwarding constraints limit each compression update while allowing newly activated evidence to remain available during subsequent execution.

1 Related Work

1.1 Context Management for Long Horizon Agents

Methods for controlling growing agent histories differ in what they remove and how they decide to remove it. ReSum periodically summarizes interaction history, ACON improves compression guidelines using failure feedback, and PACE adjusts historical detail according to predicted relevance to the next action (Kang et al., 2025; Wei et al., 2026). SelfCompact and Self-GC give agents context management operations, while Sculptor supports summarizing, hiding, restoring, and searching historical information (Li et al., 2026b; Hao et al., 2026; Li et al., 2026a). Other recent approaches learn what to retain, decouple context management from the acting model, or expose agentic context management mechanisms (Jahan Lia & Mazumder, 2026; Zhang et al., 2026; Li et al., 2026c). GEM instead selects complete original interaction blocks and forwards them unchanged through the acting model's existing interface, keeping context selection external to the agent.

1.2 Memory and Execution Evidence

Memory systems preserve information beyond the active context (Packer et al., 2023; Park et al., 2023; Shinn et al., 2023). MEM1 jointly learns reasoning and memory consolidation, SAM combines compact memory cues with access to raw history, and AdaCoM learns an external context manager for a frozen agent (Zhou et al., 2026; Hu et al., 2026). MAGE organizes history around execution dependencies and errors, and Proactive Memory Agent emphasizes that information present in a transcript may still fail to influence later behavior (Chen et al., 2026; Wu et al., 2026). These studies motivate explicit treatment of task state. Our focus is more specific: when original records share semantic content, their execution roles may still differ. GEM protects such evidence before completing coverage of the remaining history.

1.3 Task Aware Context Selection

Prompt compression methods such as LLMLingua-2 learn which tokens to retain, while SWE-Pruner and SWE-Pruner Pro study selective retention of tool context (Pan et al., 2024; Wang et al., 2026). Dynamic token pruning and long context evaluation further study which parts of a long context

remain useful under fixed model interfaces (Fu et al., 2024; Hsieh et al., 2024; Zhang et al., 2024). Retrieval systems likewise explore hierarchical or adaptive evidence selection (Sarthi et al., 2024; Jeong et al., 2024; Asai et al., 2024). Work on context interference further shows that retaining relevant evidence is not sufficient when surrounding information impairs its use. We study selection at the level of complete interaction blocks and distinguish coverage of history representations from information related to the next action. The selector is training free and uses only the current task and available history without modifying the acting model.

2 Geometry and Task Information in Agent Histories

We analyze history representations to distinguish redundancy from information needed for continued execution. The recorded next action defines the offline measurement target; the deployed compressor uses the current goal and history.

2.1 History Representation

A removable interaction block contains an assistant tool call message and all of its corresponding tool results. Calls and results are retained or removed together. The system and task prefix is kept separately, and messages that cannot be safely assigned to complete blocks are not deleted.

Prior analyses of representation spaces characterize anisotropy, dominant directions, and rank based concentration in contextual and sentence embeddings across modern neural representation models (Ethayarajh, 2019; Mu et al., 2017; Gao et al., 2021).

Let $H_t = \{o_1, \ldots, o_n\}$ denote the available complete blocks. A frozen encoder maps each block to a unit vector $x_i \in \mathbb{R}^d$. With $X = [x_1, \ldots, x_n]^\top$ and mean representation $\bar{x}$, we study centered variation through the following centered matrix decomposition

$$X_c = X - \mathbf{1}\bar{x}^\top = U\Sigma V^\top, \qquad \Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r). \tag{1}$$

The singular values are ordered from largest to smallest. Writing $p_j = \sigma_j^2 / \sum_k \sigma_k^2$, we summarize concentration by using two complementary spectral statistics

$$r_{\mathrm{eff}}(X_c) = \exp\Big(-\sum_j p_j \log p_j\Big), \qquad r_\rho(X_c) = \min\Big\{k : \sum_{j \le k} p_j \ge \rho\Big\}. \tag{2}$$

We also report $\bar{r}_\rho = r_\rho / n$. The spectral analysis sets ρ to 0.90. Low effective dimension summarizes concentration of representational variation. Retention of evidence needed for downstream execution is evaluated separately through the action related metrics below.

2.2 Coverage and Action Related Retention

Let $S \subseteq [n]$ index a retained subset, $\bar{x}_S$ be its mean, and Q_S contain an orthonormal basis for its span. Overall geometric coverage is measured by

$$C(S) = \frac{\bar{x}_S^\top \bar{x}}{\|\bar{x}_S\|_2 \|\bar{x}\|_2}, \qquad E(S) = \frac{1}{n} \sum_{i=1}^{n} \|Q_S^\top x_i\|_2^2. \tag{3}$$

Centroid similarity $C(S)$ measures preservation of the mean direction. Captured energy $E(S)$ measures average projection onto the retained span. Unlike the centered spectral analysis, $E(S)$ uses the original normalized block vectors, matching the online selector.

For the unit representation a of the recorded next action, we separately measure

$$P_a(S) = \|Q_S^\top a\|_2^2, \qquad R_a(S) = \frac{P_a(S)}{P_a([n])}, \qquad T_3(S; a) = \frac{|S \cap \mathcal{N}_3(a)|}{3}. \tag{4}$$

Here $\mathcal{N}_3(a)$ contains the three history blocks most similar to a. Projection measures retention of the action direction; Top 3 retention measures retention of the corresponding original blocks. Both are representation retention metrics, with task success assessed through benchmark reward. Ratios are evaluated only when their denominators are defined.

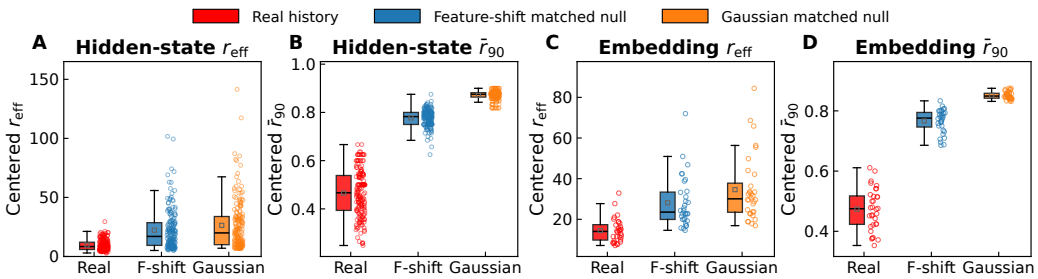

Figure 1: Structured redundancy in agent histories. Real histories are compared with feature shift and Gaussian controls. The two representation spaces show the same concentration pattern.

2.3 Information Outside the Dominant Subspace

Let $V_{\text{high}} = [v_1, \ldots, v_{r_\rho}]$ and $V_{\text{low}} = [v_{r_\rho+1}, \ldots, v_r]$. The latter directions account for at most a fraction $1 - \rho$ of centered history variation. Their share of the action projection is

$$m_{\text{low}}(a) = \frac{\|V_{\text{low}}^\top a\|_2^2}{\|V^\top a\|_2^2}. \tag{5}$$

A large value indicates that action related information extends beyond the leading history directions. Block selection additionally accounts for execution evidence. For the controlled replacement analysis, we additionally measure the low energy fraction of each centered block and its alignment with the action in this subspace. Appendix C.3 gives the complete definitions.

3 Evidence Beyond the Dominant Subspace

We test whether agent histories contain nonrandom representational redundancy, whether matching global geometry is sufficient to retain task evidence, and whether action related information can be selectively removed from low energy directions.

3.1 Histories Exhibit Structured Redundancy

The first analysis contains 171 trajectories and 4,701 complete blocks. Each block is encoded independently using mean layer 32 representations from a frozen Qwen3.5-9B encoder, producing vectors of dimension 4,096. A second analysis uses representations of dimension 1,024 from Qwen3-Embedding-8B, the online selector's encoder, on 32 trajectories and 1,162 block representations. Both analyses use externally encoded blocks, computed independently of the acting model's forward pass over its full history.

For each trajectory, we compare the real representation matrix with two matched controls. Independent circular feature shifts preserve every feature's marginal values but disrupt alignment across blocks; a Gaussian control matches matrix shape. Figure 1 gives the visual comparison, while the exact centered effective rank and r_{90}/n medians are moved to Appendix C, Table C.1. Median effective ranks are 8.39, 16.99, and 19.96 for real, feature shift, and Gaussian isolated blocks, and 14.04, 23.54, and 30.07 in selector space. All 171 isolated block comparisons and all 32 selector space comparisons place the real history below its feature shift control in effective rank. Across the four primary feature shift comparisons, the reported Holm adjusted values satisfy $p \leq 1.86 \times 10^{-9}$.

The result establishes representational concentration. The uncentered greedy coverage analysis uses a median of six isolated blocks and ten selector blocks, corresponding to median retained fractions of 0.30 and 0.35, respectively. These counts quantify geometric coverage; the following matched count comparison separately evaluates preservation of task evidence.

3.2 Coverage Does Not Determine Evidence Retention

We next compare geometry only and evidence constrained selection at identical retained block counts. The dense replay contains 650 checkpoints from 32 trajectories, starting after 16 complete blocks. Individual trajectories contribute between 2 and 73 checkpoints (median

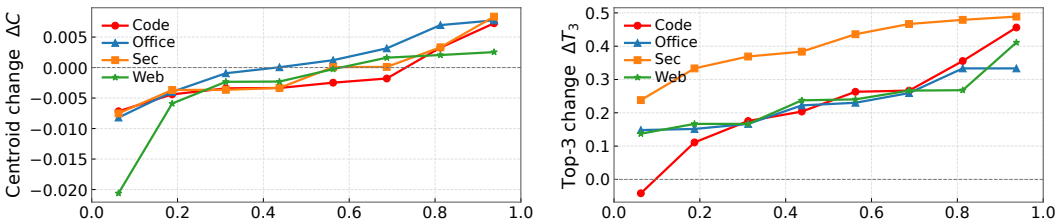

Figure 2: Matched budget geometry and task information. The trajectory level visualization contrasts evidence constrained and geometry only selection at the same retained block count.

15.5), with 130, 108, 161, and 251 checkpoints from Code, Office, Security, and Web, respectively. The independent statistical units remain the 32 trajectories. At the same retained budget, evidence constrained selection preserves substantially more task information, particularly next action Top 3 retention, while global geometry remains nearly unchanged.

Table 1 reports the pooled same budget means. At the same retained count, next action Top 3 retention rises from 0.31 to 0.69, while centroid similarity is unchanged at two decimal precision. Action, final state, and goal projections show the same pattern: evidence retention improves without a comparably large shift in the global centroid measure. The metric definitions and pooled checkpoint aggregation are detailed in Appendix C.2.

Table 1: Matched budget retention at 650 checkpoints.

Metric	Geometry	Evidence	Δ
Action projection	0.91	0.94	↑ 0.03
Next action Top 3	0.31	**0.69**	↑ 0.38
Final projection	0.90	0.94	↑ 0.04
Goal projection	0.91	0.94	↑ 0.03
Centroid similarity	0.98	0.98	

The dense results report pooled checkpoint means, with trajectories serving as the independent statistical units. Centroid similarity is compared at the displayed numerical precision alongside the task information metrics.

3.3 Low Energy Evidence and Controlled Replacement

Figure 3 shows next action information beyond the leading directions. The plotted bands group directions by spectral rank, while the formal intervention uses directions after the 90% centered energy cutoff. A block can therefore be weakly represented by the dominant variation of the trajectory while still carrying evidence relevant to the next action.

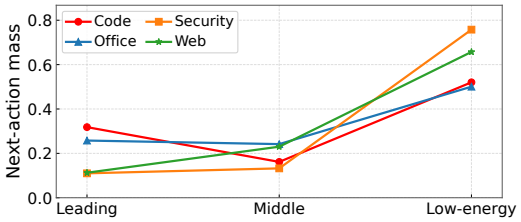

Figure 3: Next action information across ordered spectral bands.

This observation motivates a controlled single block replacement. The targeted condition removes a retained block whose low energy component is strongly aligned with the recorded next action. The matched condition removes a low alignment block while matching representation similarity, recency, and serialized length.

Table 2 shows the same dissociation in intervention form. Targeted replacement reduces action projection by 0.03 and next action Top 3 retention by 0.33, whereas centroid similarity changes by only +0.00 and captured energy by +0.00. The intervention quantities and matching definition are given in Appendix C.3 for full reproducibility.

Table 2: Targeted block replacement effects.

Metric	Δ
Task information	
Action projection	−0.03
Next action Top 3	−0.33
Global geometry	
Centroid similarity	+0.00
Captured energy	+0.00

The targeted replacement therefore removes substantially more action related information while global geometric measures remain nearly unchanged. This representation level intervention quantifies how block selection changes retained action information, directly characterizing the distinction between geometry and evidence at fixed history checkpoints.

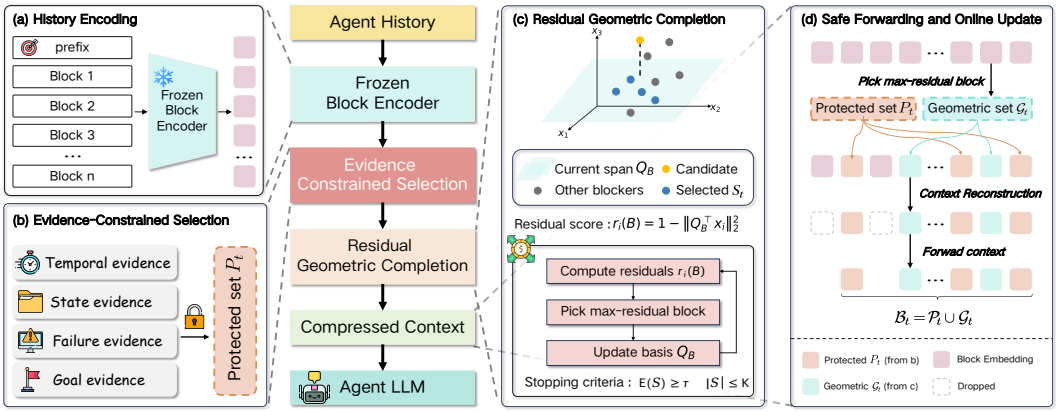

Figure 4: Overview of GEM. (a) A frozen encoder represents complete interaction blocks. (b) Evidence constraints form the protected set $\mathcal{P}_t$. (c) Residual geometric completion adds blocks whose directions are not covered by the current selected span. (d) Retained original blocks are reconstructed in chronological order and forwarded to the unchanged agent. The online forwarding constraints are described in Section 4.3.

Execution aware protection distinguishes plans, state changing operations, and subsequent observations, even when they refer to the same artifact. Appendix B.2 details the rules that identify and preserve these different execution roles.

4 GEM: Geometry Guided Evidence Preserving Memory

GEM is a training free external compressor for complete interaction histories. It separates protected evidence from records selected for additional geometric coverage. The encoder and acting model remain frozen. Selection uses external block embeddings and forwards original messages through the acting model's ordinary request interface.

Figure 4 summarizes the procedure. The protected set initializes geometric completion, making the second stage conditional on the retained evidence. Embeddings guide selection, while the original message records supply the forwarded context. This separation keeps the selection mechanism lightweight while preserving the original interaction semantics seen by the acting model.

4.1 Evidence Constrained Selection

We distinguish the protected set $\mathcal{P}_t$, the selected core B_t, and the final forwarded set S_t. The core contains the evidence and additional directions identified by the selector. The forwarded set may be larger because the online deletion guard retains additional original blocks. This distinction separates the semantic selection objective from the request level constraints that determine what is ultimately forwarded to the acting model.

For nominal core capacity K, let $K_t = \max\{K, |\mathcal{P}_t|\}$ so that required evidence is not dropped when its count exceeds the nominal capacity. The ideal core selection problem is

$$B_t^\star = \arg\min_{B \subseteq [n]} |B| \quad \text{s.t.} \quad \mathcal{P}_t \subseteq B, \quad E(B) \geq \tau, \quad |B| \leq K_t. \tag{6}$$

The greedy procedure constructs the core through sequential residual selection. The returned core can exceed the minimum size, and coverage can remain below its target when capacity or candidate directions are exhausted. K controls nominal core capacity; the forwarding guard determines the final retained count.

4.2 Protecting Task and Execution Evidence

The protected set combines recent interactions, goal related observations, execution state evidence, fresh errors, and source records (Chen et al., 2026; Zhou et al., 2026; Wu et al., 2026; Hu et al., 2026).

These groups are combined and deduplicated before geometric completion. Recent blocks retain the immediate working context. Goal selection keeps the eligible block most similar to the current task. State selection uses the current state query, file paths, and supported operation patterns to identify relevant records of possible changes to task artifacts. Together, these signals prioritize records whose execution role may remain important even when their semantic content is highly redundant.

A plan, a write, and a read of the same file can differ in execution value even when they add little geometric novelty. We therefore inspect explicit editing calls and supported Bash and Python write patterns, rather than relying only on tool names. State candidates are organized by target path. Error evidence is protected within a fixed age window, and source protection retains recent records from distinct sources when observations resembling source records form a sufficient fraction of the history.

These rules use the available history and current goal as inexpensive protection cues. Detected writes mark possible state changes, and the error window determines eligibility for protection. Operation success and error resolution remain separate from these eligibility rules. Protection follows the supported operation patterns and source records, with numerical limits listed in Appendix B.

4.3 Geometric Completion and Online Updates

Starting from $B = \mathcal{P}_t$, we construct an orthonormal basis Q_B with modified Gram Schmidt orthogonalization. For each eligible unselected block, we compute its residual and select the largest:

$$r_i(B) = 1 - \|Q_B^\top x_i\|_2^2, \qquad i^\star = \arg\max_{i \notin B} r_i(B). \tag{7}$$

We add $i^\star$, update the basis, and continue until coverage, capacity, or candidate exhaustion terminates selection. The residual therefore measures information not yet represented by the current geometric span, rather than relevance to the task itself. Writing $\mathcal{G}_t = B_t \setminus \mathcal{P}_t$ for the blocks added by geometric completion gives $B_t = \mathcal{P}_t \cup \mathcal{G}_t$. A protected record cannot be removed merely because it has a small residual. The default coverage target is $\tau = 0.90$, and the nominal core capacity is $K = 16$.

The online implementation limits how much of a request can be removed in one rewrite. If $L(S)$ is the serialized length of the complete request containing blocks S, the forwarded set satisfies

$$B_t \subseteq S_t \subseteq [n], \qquad L(S_t) \geq (1 - \delta) L([n]). \tag{8}$$

We set $\delta = 0.05$. Starting from the full safe request, the implementation attempts to remove eligible blocks outside the core, prioritizing larger redundant blocks while respecting this guard. Complete blocks are indivisible, so an event may remove nothing. All retained messages remain unchanged and in chronological order.

Requests with fewer than 16 complete blocks are left unchanged. Global selection occurs at the first eligible request, after a session or goal change, or after 256 newly added blocks. Between these events, the implementation keeps the previously forwarded set and appends new blocks and newly activated evidence. This incremental policy avoids repeatedly reconstructing the retained history when no global reselection is required. Storing the actual forwarded set prevents a later update from mistakenly removing blocks retained by the deletion guard. The per request character limit is distinct from cumulative token savings over a task. Appendix A gives the corresponding geometric derivations and forwarding properties.

5 Experiments

5.1 Experimental Setup

Table 3: WorkBuddyBench Full260 results. S denotes mean reward $\times 100$, and T denotes combined tokens per task (M). Higher S and lower T are better.

Method	**Code**		**Office**		**Security**		**Web**		**Overall**	
	$S\uparrow$	$T\downarrow$	$S\uparrow$	$T\downarrow$	$S\uparrow$	$T\downarrow$	$S\uparrow$	$T\downarrow$	$S\uparrow$	$T\downarrow$
Base agent	**76.99**	1.37	81.84	1.20	47.76	5.98	**72.14**	2.43	69.87	2.69
GEM	75.25	**1.35**	**84.06**	**1.15**	**49.56**	**3.95**	**72.14**	**2.09**	**70.18**	**2.11**

Table 4: Fixed 40 context management reference. S is mean reward $\times 100$ and T is combined tokens per task (M). Each domain contains ten tasks.

Method	Code		Office		Security		Web		Overall	
	$S\uparrow$	$T\downarrow$	$S\uparrow$	$T\downarrow$	$S\uparrow$	$T\downarrow$	$S\uparrow$	$T\downarrow$	$S\uparrow$	$T\downarrow$
Base agent	72.93	1.44	<u>81.64</u>	1.04	44.58	3.97	71.00	1.10	67.54	1.89
Window ($K=5$)	66.81	3.12	67.43	3.98	32.32	2.13	59.00	2.58	56.39	2.95
Window ($K=10$)	74.37	1.76	73.05	1.82	40.07	<u>0.82</u>	61.00	3.55	62.12	1.99
Window ($K=20$)	<u>79.31</u>	1.42	**85.11**	2.23	40.28	1.13	70.00	3.30	68.67	2.02
Periodic summary ($n=3$)	65.44	1.91	67.66	2.14	26.32	0.99	**82.66**	3.82	60.52	2.21
Periodic summary ($n=5$)	**83.60**	1.76	75.74	2.55	40.03	1.29	81.32	3.88	<u>70.17</u>	2.37
PACE (Wei et al., 2026)	70.38	<u>0.94</u>	55.75	1.01	43.53	**0.57**	64.02	1.89	58.42	<u>1.10</u>
LLMLingua-2 (Pan et al., 2024)	64.64	1.63	73.31	3.41	40.10	3.49	70.00	1.09	62.01	2.41
SelfCompact (Li et al., 2026b)	76.83	1.57	79.62	<u>0.85</u>	36.19	3.14	71.00	0.71	65.91	1.57
ACON-Core (Kang et al., 2025)	71.79	1.48	68.55	2.15	<u>47.94</u>	1.51	65.00	<u>0.47</u>	63.32	1.40
Self-GC (Hao et al., 2026)	63.45	1.17	71.67	1.18	43.70	3.44	71.00	0.84	62.46	1.66
LRE (Jahan Lia & Mazumder, 2026)	62.62	2.26	63.44	3.47	27.14	2.85	68.00	0.59	55.30	2.29
CoMem (Zhang et al., 2026)	56.43	**0.55**	59.86	**0.75**	2.50	1.00	58.00	0.76	44.20	**0.77**
SAM (Hu et al., 2026)	67.62	1.41	81.63	1.09	47.23	1.91	67.00	0.97	65.87	1.35
SWE-Pruner (Wang et al., 2026)	63.45	2.17	72.60	1.43	40.87	1.82	65.00	0.58	60.48	1.50
Sculptor (Li et al., 2026a)	51.19	1.03	81.62	0.99	35.24	2.45	69.00	0.71	59.26	1.29
ACM (Li et al., 2026c)	38.69	2.54	55.93	1.53	6.19	5.20	34.00	**0.19**	33.70	2.36
GEM	76.67	1.74	78.51	1.32	**55.31**	3.88	<u>82.00</u>	2.08	**73.12**	2.26

WorkBuddyBench Full260 (Team et al., 2026) contains 260 tasks across Code, Office, Security, and Web. The main online evaluation uses DeepSeek-V4-Flash (Xu et al., 2026) at temperature zero, while GEM uses frozen representations of dimension 1,024 from Qwen3-Embedding-8B. Task quality is reported as *Score*$=100\times$mean reward, and combined tokens include model input, model output, and embedding tokens. Full260 is the primary end to end evaluation. The fixed 40 panel reports common compression controls alongside a GEM subset with the same task identifiers drawn from Full260, marked separately in Table 4. Appendix E.3 describes the reference panel protocol.

5.2 Full260 End to End Results

Across all 260 tasks, mean combined usage decreases from 2.69M to 2.11M per task, a 21.4% reduction. Mean score changes from 69.87 to 70.18 (reward $\times 100$). The average hides important domain variation. Security shows the largest token decrease, 5.98M to 3.95M per task, while its mean reward changes from 0.48 to 0.50. Office also increases in mean reward while using fewer tokens, and Web keeps the same rounded reward while reducing tokens. Code is the clearest boundary case: tokens change only slightly (1.37M to 1.35M), and the mean reward decrease rounds to 0.02. These domain results show different reward and token tradeoffs, with Code providing the smallest token reduction and a decrease in mean reward.

The token decomposition in Appendix E.2 shows that the net reduction comes primarily from a lower cached input volume. This decrease more than offsets the additional uncached input and the selector's embedding usage, yielding the reported reduction in combined tokens.

5.3 Fixed 40 Context Management Reference Panel

Table 4 reports reward and token use for the fixed 40 reference panel, showing how these tradeoffs vary across Code, Office, Security, and Web. The panel combines distinct evaluation snapshots: the GEM row is a subset of Full260 with the same task identifiers, while the reference methods come from the fixed 40 runs. The evaluation sources and protocol are described in Appendix E.3.

The reference panel illustrates why compression should not be summarized by a single token ratio. For example, periodic summarization at $n=5$ has a relatively high mean score in this panel but uses more tokens than the uncompressed base, whereas CoMem uses substantially fewer tokens with a much lower mean score. The domain columns show that these operating points vary substantially across task families.

Table 5: Online 20 task ablation by domain. S is mean reward $\times 100$ and T is combined tokens per task (M). Each domain contains five tasks.

Variant	Code		Office		Security		Web		Overall	
	$S\uparrow$	$T\downarrow$	$S\uparrow$	$T\downarrow$	$S\uparrow$	$T\downarrow$	$S\uparrow$	$T\downarrow$	$S\uparrow$	$T\downarrow$
Base agent	59.79	**2.33**	**74.28**	3.02	**62.95**	10.70	46.00	4.73	**60.75**	5.19
GEM	**70.79**	2.52	69.32	3.18	53.45	**4.33**	48.00	3.41	60.39	**3.36**
Random completion	47.40	2.72	60.86	5.20	56.69	6.69	47.30	**3.23**	53.06	4.46
Geometry only	60.93	2.40	59.62	**1.25**	48.92	8.14	**50.64**	4.71	55.03	4.12

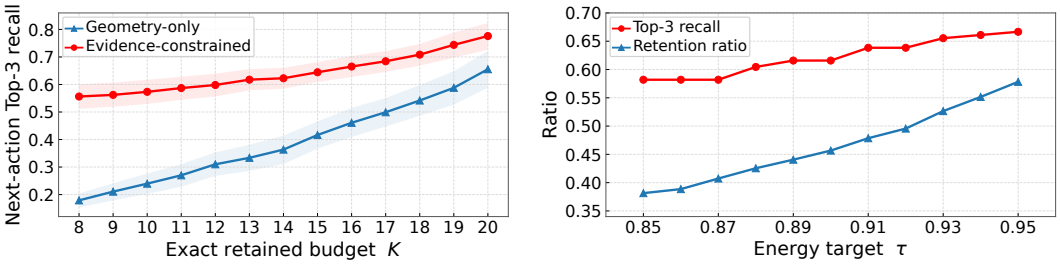

Figure 5: Compression strength. (a) Next action Top 3 retention at exact retained budgets K. (b) Top 3 retention and retained history ratio as the coverage threshold τ increases. The exact K sweep uses 28 trajectories and 525 common checkpoints; the coverage sweep uses 32 trajectories and 59 checkpoints.

5.4 Online Ablation and Compression Strength

The online ablation uses 20 formally valid tasks, five from each domain. Random completion retains the same evidence rules and candidate pool but replaces largest residual completion with random selection. Geometry only removes the evidence protection bundle while retaining residual selection and the online bookkeeping required for valid requests.

On this batch, the full method scores 60.39 versus 60.75 for the base agent (a difference of 0.36 points) while using 35.30% fewer combined tokens overall. The domain columns make the tradeoff explicit: on the five Code tasks GEM raises reward while slightly increasing tokens, whereas on the five Security tasks it sharply reduces tokens but loses reward. Relative to random completion and geometry only selection, the full method has the highest overall reward and the lowest overall token use in this batch. The ablation supports the combined selector at the evaluated operating point, with distinct reward and token tradeoffs across domains.

Evidence constrained selection has higher Top 3 retention than geometry only selection at every tested exact budget $K \in [8, 20]$. Increasing τ from 0.85 to 0.95 raises the retained fraction from 0.38 to 0.53 and Top 3 retention from 0.58 to 0.65. The default $\tau = 0.90$ is therefore an operating point within this observed tradeoff, balancing retained history size and preservation of action evidence. The complete sweep values are listed in Appendix D.3.

6 Conclusion

Agent histories can be highly redundant in representation space while still containing low energy records that matter for continued execution. This paper separates those two notions of importance. Same budget selection and controlled replacement show that preserving dominant geometry does not guarantee preservation of next action evidence. GEM converts this observation into a simple external compressor: protect task and execution evidence first, then use geometric residuals to complete coverage. It preserves original messages and leaves the acting model unchanged. On WorkBuddyBench Full260, the reported operating point reduces mean combined tokens from 2.69M to 2.11M per task while maintaining a similar aggregate reward, with clear domain dependent tradeoffs. The broader lesson is that repeated content can still be consequential evidence; effective history compression should model that distinction explicitly.

Reproducibility Statement

We provide the methodological definitions, selection algorithm, implementation settings, evaluation protocols, ablation configurations, and additional analyses required to reproduce the reported results in the main text and appendix. The acting model and representation encoder remain frozen under the default setting, and all reported comparisons follow the evaluation and token accounting procedures described in the manuscript.

AI Use Statement

Generative AI tools assisted with code development, analysis scripts, figure and table preparation, and manuscript editing under author supervision. All experimental results and scientific claims were reviewed and verified by the authors, who take responsibility for the final manuscript.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pp. 9112–9141, 2024.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 3119–3137, 2024.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, et al. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3639–3664, 2025.

Yaoqi Chen, Haibin Lai, Yuru Feng, Chuyu Han, Qianxi Zhang, Baotong Lu, Menghao Li, Xinjiang Wang, Zhirui Wang, Shusen Xu, et al. Beyond semantic organization: Memory as execution state management for long-horizon agents. *arXiv preprint arXiv:2606.06090*, 2026.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 55–65, 2019.

Qichen Fu, Minsik Cho, Thomas Merth, Sachin Mehta, Mohammad Rastegari, and Mahyar Najibi. Lazyllm: Dynamic token pruning for efficient long context llm inference. *arXiv preprint arXiv:2407.14057*, 2024.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 6894–6910, 2021.

Tao Ge, Jing Hu, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. In-context autoencoder for context compression in a large language model. *arXiv preprint arXiv:2307.06945*, 2023.

Xubin Hao, Hongjin Meng, Xin Yin, Jiawei Zhu, and Chenpeng Cao. Self-gc: Self-governing context for long-horizon llm agents. *arXiv preprint arXiv:2607.00692*, 2026.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

Yuyang Hu, Hongjin Qian, Shuting Wang, Jiongnan Liu, Ziliang Zhao, Jiejun Tan, Zheng Liu, and Zhicheng Dou. Sam: State-adaptive memory for long-horizon reasoning agent. *arXiv preprint arXiv:2605.24468*, 2026.

Nusrat Jahan Lia and Aritra Mazumder. Learning what not to forget: Long-horizon agent memory from a few kilobytes of learning. *arXiv e-prints*, pp. arXiv–2606, 2026.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers)*, pp. 7036–7050, 2024.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. Llmlingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 13358–13376, 2023.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1658–1677, 2024.

Minki Kang, Wei-Ning Chen, Dongge Han, Huseyin A Inan, Lukas Wutschitz, Yanzhi Chen, Robert Sim, and Saravan Rajmohan. Acon: Optimizing context compression for long-horizon llm agents. *arXiv preprint arXiv:2510.00615*, 2025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Mo Li, LH Xu, Qitai Tan, Long Ma, Hongyong Song, Ting Cao, and Yunxin Liu. Sculptor: Empowering llms with cognitive agency via active context management. In *International Conference on Learning Representations*, volume 2026, pp. 153411–153440, 2026a.

Tianjian Li, Jingyu Zhang, William Jurayj, Xi Wang, Chuanyang Jin, Mehrdad Farajtabar, Eric Nalisnick, and Daniel Khashabi. Self-compacting language model agents. *arXiv preprint arXiv:2606.23525*, 2026b.

Xiaochuan Li, Ryan Ming, Meng Chu, Shuai Shao, Rong Jin, and Chenyan Xiong. Acm: Agentic context management for long horizon tasks. *arXiv preprint arXiv:2607.23809*, 2026c.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024a.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. In *International Conference on Learning Representations*, volume 2024, pp. 52989–53046, 2024b.

Jesse Mu, Xiang Li, and Noah Goodman. Learning to compress prompts with gist tokens. *Advances in Neural Information Processing Systems*, 36:19327–19352, 2023.

Jiaqi Mu, Suma Bhat, and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. *arXiv preprint arXiv:1702.01417*, 2017.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, et al. Llmlingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 963–981, 2024.

Joon Sung Park, Joseph O'Brien, Carrie Jun 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.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher Manning. Raptor: Recursive abstractive processing for tree-organized retrieval. In *International Conference on Learning Representations*, volume 2024, pp. 32628–32649, 2024.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Tencent WorkBuddy Bench Team, Siqi Cai, Shaopeng Chen, Xiang Fei, Yong Mao, Zihan Xu, Zhiheng Lyu, Zhijian Shao, Yuchen Shi, Shuwen Zhang, et al. Tencent workbuddy bench: A multi-domain coding-agent benchmark with contamination-resistant task construction. *arXiv preprint arXiv:2607.20911*, 2026.

Yuhang Wang, Yuling Shi, Mo Yang, Rongrui Zhang, Shilin He, Heng Lian, Yuting Chen, Siyu Ye, Kai Cai, and Xiaodong Gu. Swe-pruner: Self-adaptive context pruning for coding agents. *arXiv preprint arXiv:2601.16746*, 2026.

Lei Wei, Xiao Peng, Guannan Zhang, Chenhao Jiang, Hongyu Li, Lanbo Lin, Yuanwu Xu, Jiayao Liu, Kesu Wang, Bin Wang, et al. Pace: Predictive adaptive context extraction for long-horizon llm agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27184–27199, 2026.

Yifan Wu, Lizhu Zhang, Yuhang Zhou, Mingyi Wang, Bo Peng, Serena Li, Xiangjun Fan, and Zhuokai Zhao. Remember when it matters: Proactive memory agent for long-horizon agents. *arXiv preprint arXiv:2607.08716*, 2026.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. Recomp: Improving retrieval-augmented lms with context compression and selective augmentation. In *International Conference on Learning Representations*, volume 2024, pp. 43478–43502, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, et al. ∞ bench: Extending long context evaluation beyond 100k tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15262–15277, 2024.

Yuwei Zhang, Chengyu Dong, Shuowei Jin, Changlong Yu, Hejie Cui, Hongye Jin, Xinyang Zhang, Hamed Bonab, Colin Lockard, Jianshu Chen, et al. Comem: Context management with a decoupled long-context model. *arXiv preprint arXiv:2605.30842*, 2026.

Zijian Zhou, Ao Qu, Zhaoxuan Wu, Sunghwan Kim, Alok Prakash, Daniela Rus, Bryan Kian Hsiang Low, and Paul Liang. Mem1: Learning to synergize memory and reasoning for efficient long-horizon agents. In *International Conference on Learning Representations*, volume 2026, pp. 58413–58438, 2026.

A Mathematical Derivations and Properties

We derive the geometric and structural properties used by GEM. Throughout this section, the available history is fixed, the block vectors $x_i \in \mathbb{R}^d$ have unit norm, and Q_B contains an orthonormal basis for the span of the selected vectors. The statements concern these representation space quantities and the selection and forwarding operations.

A.1 Residual Selection and Reconstruction Error

Write $X = [x_1, \ldots, x_n]^\top$ and $P_B = Q_B Q_B^\top$, with $P_\varnothing = 0$. The selected block identities and the numerical basis are maintained separately: a protected block remains selected even if its vector is already in the span of the other selected vectors. For any block, the orthogonal decomposition is

$$x_i = P_B x_i + e_i(B), \qquad e_i(B) = (I - P_B)x_i. \tag{A.1}$$

Proposition A.1 (Residual score and geometric coverage) *For normalized block vectors,*

$$r_i(B) = \|e_i(B)\|_2^2 = 1 - \|Q_B^\top x_i\|_2^2, \tag{A.2}$$

$$E(B) = \frac{\|XQ_B\|_F^2}{n} = 1 - \frac{\|X(I - P_B)\|_F^2}{n}. \tag{A.3}$$

Consequently, $0 \le r_i(B) \le 1$, $0 \le E(B) \le 1$, *and a core satisfying* $E(B) \ge \tau$ *has mean squared embedding reconstruction error at most* $1 - \tau$.

Proof. Because P_B is a symmetric idempotent projector, $P_B x_i$ is orthogonal to $(I - P_B)x_i$. Thus $\|x_i\|_2^2 = \|P_B x_i\|_2^2 + \|e_i(B)\|_2^2$. Orthonormality gives $\|P_B x_i\|_2^2 = \|Q_B^\top x_i\|_2^2$. Using $\|x_i\|_2^2 = 1$ proves the first identity. Summing it over all n blocks gives the second identity and the stated bounds. □

Equation A.2 explains the completion rule: the largest residual candidate contributes a direction that is poorly represented by the current selected span. Equation A.3 gives the coverage target its direct reconstruction interpretation. These quantities are evaluated on the original normalized vectors, matching the online selector; the centered spectral analysis is defined separately in Appendix C.

A.2 Coverage Gain and Incremental Basis Updates

For an eligible candidate j with $r_j(B) > 0$, modified Gram Schmidt constructs the new unit direction

$$q = \frac{x_j - Q_B(Q_B^\top x_j)}{\sqrt{r_j(B)}}, \qquad Q_{B\cup\{j\}} = [Q_B, q]. \tag{A.4}$$

Proposition A.2 (Monotone coverage completion) *For the update in Eq. A.4,*

$$P_{B\cup\{j\}} = P_B + qq^\top, \tag{A.5}$$

$$E(B \cup \{j\}) - E(B) = \frac{1}{n}\sum_{i=1}^{n} (q^\top x_i)^2, \tag{A.6}$$

$$r_i(B \cup \{j\}) = r_i(B) - (q^\top x_i)^2. \tag{A.7}$$

In particular, the coverage gain is at least $r_j(B)/n > 0$. *More generally,* $B \subseteq S$ *implies* $E(S) \ge E(B)$.

Proof. The normalized residual q is orthogonal to every column of Q_B. The projector onto the expanded span is therefore $P_B + qq^\top$. Substituting this expression into the definition of E gives Eq. A.6, and subtracting the new projected component gives Eq. A.7. All terms in the coverage gain are nonnegative, and the $i = j$ term equals $r_j(B)$. For nested sets $B \subseteq S$, extend Q_B by an orthonormal basis of the additional directions in $\mathrm{span}(S)$ and apply the same argument to each added direction. □

For m added independent directions $q_1, \ldots, q_m$, the increments sum to

$$E(B_m) = E(B_0) + \frac{1}{n} \sum_{k=1}^{m} \|Xq_k\|_2^2. \tag{A.8}$$

Thus geometric completion progressively covers additional directions after the evidence set has been protected. The residual update identity also allows candidate scores to be updated without rebuilding all previous projections at each iteration.

A.3 Evidence Preservation and Forwarding Invariants

The selector begins with $B_0 = \mathcal{P}_t$ and uses the expanded capacity $K_t = \max\{K, |\mathcal{P}_t|\}$. It only adds indices during completion. The forwarding stage starts from the full valid request and considers removing indices outside the completed core B_t.

Proposition A.3 (Protected and structurally valid forwarding) *For an accepted global rewrite,*

$$\mathcal{P}_t \subseteq B_t \subseteq S_t \subseteq [n], \qquad L(S_t) \geq (1-\delta)L([n]). \tag{A.9}$$

Every retained removable block contains its complete assistant call and result group, and the retained messages preserve their original chronological order. Between global reselections, under stable block identities, the append rule satisfies

$$S_{t+1} = S_t \cup \mathcal{N}_{t+1} \cup \mathcal{A}_{t+1}, \qquad S_t \subseteq S_{t+1}, \tag{A.10}$$

where $\mathcal{N}_{t+1}$ and $\mathcal{A}_{t+1}$ are newly available blocks and newly activated evidence, respectively.

Proof. The core is initialized to $\mathcal{P}_t$ and is modified only by insertion, so $\mathcal{P}_t \subseteq B_t$. Forwarding starts from $[n]$ and deletes only members of its complement to B_t, proving $B_t \subseteq S_t$. Initially the length constraint holds; accepting each proposed deletion only when it continues to hold proves the bound by induction. Parsing constructs complete call and result blocks, selection changes only whole block membership, and reconstruction sorts retained messages by their original indices. This preserves call and result groups and chronological order. Finally, the ordinary update is a set union with the saved actual forwarded set, proving Eq. A.10. □

Keeping the actual forwarded set is important: it preserves not only the core but also the additional original blocks retained by the request length guard. Protected block membership is therefore independent of whether a protected embedding introduces a new basis direction.

B Algorithm and Implementation Details

B.1 Complete Blocks and Selection Representations

A removable interaction block contains an assistant message with one or more tool calls and all corresponding tool result messages, matched by call ID. The parser checks identifier uniqueness and complete matching between calls and results. The system/task prefix and messages outside removable blocks are preserved independently. When parsing or validation of the reconstructed request fails, the original request is returned.

The selection representation concatenates assistant reasoning, assistant text, tool names and arguments, and tool observations under the markers `ASSISTANT_REASONING`, `ASSISTANT_TEXT`, `ACTION`, and `OBSERVATION`. This encoder input is limited to 12,000 characters, retaining its beginning and end for overlength blocks. Only the selection representation is shortened; forwarded blocks use the complete, unchanged original message objects. The frozen encoder is Qwen3-Embedding-8B with 1,024 output dimensions.

This block level interface differs from compressed retrieved text (Xu et al., 2024) and learned compact representations such as gist tokens and in context autoencoding (Mu et al., 2023; Ge et al., 2023). In GEM, embeddings guide which original blocks to retain; the acting model continues to receive ordinary messages.

B.2 Evidence Protection and Shared Settings

The protected set is a deduplicated union

$$\mathcal{P}_t = \mathcal{P}_t^{\text{recent}} \cup \mathcal{P}_t^{\text{goal}} \cup \mathcal{P}_t^{\text{state}} \cup \mathcal{P}_t^{\text{error}} \cup \mathcal{P}_t^{\text{source}}. \tag{B.1}$$

Recent protection retains four complete blocks of immediate working context. Goal protection retains the eligible block most similar to the current task. State protection retains up to three relevant records using the state query, artifact paths, explicit editing calls, and supported Bash/Python write patterns. Error protection retains up to two recent error records within eight blocks. Source protection is activated when blocks resembling source records constitute at least half of the history and keeps recent evidence from distinct sources.

These rules distinguish the execution roles of records rather than treating semantic similarity as interchangeability. State related and persistent memory provide the relevant context for this distinction (Chen et al., 2026; Zhou et al., 2026; Wu et al., 2026; Hu et al., 2026). The numerical protection limits in Table B.1 are the shared GEM settings. Detected write patterns identify records of possible state changes for protection, while the original call and result messages preserve the associated execution evidence.

Table B.1: Shared implementation settings across Code, Office, Security, and Web.

Setting	**Value**
Minimum complete blocks	16
Coverage target τ / nominal core capacity K	0.90 / 16
Recent / Goal / State / Error limits	4 / 1 / 3 / 2
Error eligibility horizon	8 blocks
Source protection activation	Source fraction ≥ 0.50
Initial block protection	0
Candidate / fill policy	`semantic_evidence` / `max_residual`
Request character reduction limit δ	0.05
Global reselection interval	256 new blocks
Length of selection representation	12,000 characters
Embedding dimension / batch size	1,024 / 16
Recorded `budget_k`	8
Random completion seed	42
Model temperature / context window	0.0 / 1,000,000 tokens

The core capacity expands to include all required evidence. The setting `budget_k` is recorded separately; the protected set, core capacity, and request length guard determine the final forwarded count.

B.3 Global Selection and Incremental Forwarding

Every request is inspected. Requests with fewer than 16 complete blocks remain unchanged. Global selection is triggered at the first eligible request, after a session reset/rewrite or goal change, or after 256 newly accumulated blocks. It initializes the core with protected evidence and adds largest residual candidates until coverage, capacity, or candidate exhaustion terminates selection.

The final forwarded set starts from the full safe request. Only blocks outside the selected core are considered for deletion, with larger redundant blocks considered first. Each accepted deletion must preserve at least 95% of the original request's serialized character count. Ordinary append events keep the saved actual forwarded set, append new blocks and activated evidence, and preserve chronological order. Algorithm 1 combines these steps.

Algorithm 1 GEM at an intercepted agent request

Input: request, goal, saved forwarded state, τ, K, δ
Parse complete blocks and preserve the original prefix
if the tool protocol is invalid or fewer than 16 complete blocks are available **then**
  return the original request
end if
Construct and deduplicate the protected set $\mathcal{P}_t$
if global reselection is triggered **then**
  Encode blocks; set $B \leftarrow \mathcal{P}_t$ and $K_t \leftarrow \max\{K, |\mathcal{P}_t|\}$
  Construct an orthonormal basis Q_B of the selected representations
  while $|B| < K_t$, $E(B) < \tau$, and an independent eligible candidate exists **do**
    Select $j \leftarrow \arg\max_{i \notin B}(1 - \|Q_B^\top x_i\|_2^2)$ over eligible candidates
    Add j to B and update the basis and residuals
  end while
  Initialize $S \leftarrow [n]$; save the original request length L_0
  for eligible blocks $i \notin B$ in the deletion priority order **do**
    if $L(S \setminus \{i\}) \geq (1 - \delta)L_0$ **then**
      $S \leftarrow S \setminus \{i\}$
    end if
  end for
else
  Keep the saved forwarded set and append new blocks and activated evidence
  Let S be the resulting index set in the current history
end if
Reconstruct complete original messages in chronological order
if the reconstructed request fails structural validation **then**
  return the original request
end if
Save the actual forwarded set and session/goal bookkeeping
return the reconstructed request

With n encoded blocks, dimension d, and final basis rank b, incremental projection updates in Eq. A.7 can be implemented in $O(ndb)$ arithmetic and $O(nd + db + n)$ working storage, excluding the original messages and encoder. These are operation count bounds for incremental updates; encoding and request serialization are accounted for separately.

C Additional Geometry and Evidence Analysis

C.1 Representations, Centering, and Matched Controls

The isolated block analysis uses 171 trajectories and 4,701 complete blocks, encoded independently with frozen Qwen3.5-9B mean layer 32 representations of dimension 4,096. The selector space analysis uses 32 trajectories and 1,162 block representations of dimension 1,024. These are the two representation populations used in Section 3.1.

For $X_c = X - \mathbf{1}\bar{x}^\top = U\Sigma V^\top$ with positive centered energy, define

$$p_j = \frac{\sigma_j^2}{\sum_k \sigma_k^2}, \qquad r_{\text{eff}} = \exp\left(-\sum_j p_j \log p_j\right), \qquad \bar{r}_{90} = \frac{1}{n}\min\left\{k : \sum_{j \leq k} p_j \geq 0.90\right\}. \tag{C.1}$$

The centered analysis isolates variation around the mean direction, complementing studies of representation anisotropy and dominant embedding directions (Ethayarajh, 2019; Mu et al., 2017; Gao et al., 2021).

The feature shift control independently circularly shifts each feature along the block axis, preserving its marginal values while disrupting cross feature alignment. The Gaussian control matches the trajectory matrix dimensions. Both comparisons are performed within trajectories before aggregation. Table C.1 gives the aggregate dimensionality statistics corresponding to the main text comparison.

Table C.1: Structured low dimensional structure of agent histories. Values are trajectory medians for centered representations; r_{90}/n is the fraction of available blocks required to explain 90% of centered energy.

Representation	Metric	Real	Feature shift	Gaussian
Isolated blocks	Effective rank	8.39	16.99	19.96
Isolated blocks	r_{90}/n	0.47	0.78	0.88
Selector space	Effective rank	14.04	23.54	30.07
Selector space	r_{90}/n	0.48	0.78	0.85

Real histories have lower effective rank than their feature shift controls in all 171 isolated block comparisons and all 32 selector space comparisons. The consistent pattern across the two spaces supports structured geometric redundancy as a basis for the completion stage.

C.2 Matched Budget Evidence Retention

The matched budget analysis holds the retained block count fixed at each checkpoint. Both selectors are measured on the same available history and recorded next action. The online compressor has access only to the current goal and available history; the recorded next action is used for offline evaluation.

For action representation a, the quantities in Eq. 4 distinguish projected action direction energy from retention of actual neighboring history records. The action projection is $\|Q_S^\top a\|_2^2$, and Top 3 retention is the fraction of the three history blocks most similar to a that remain in S. The 650 checkpoint comparison in Table 1 shows higher action related retention with evidence constraints at the same retained count, while the reported centroid similarity remains 0.98.

Let f_{jc} denote a checkpoint metric for checkpoint c of trajectory j, and let m_j be that trajectory's checkpoint count. The pooled statistic is

$$\bar{f}_{\text{pooled}} = \frac{\sum_{j=1}^{32}\sum_{c=1}^{m_j} f_{jc}}{\sum_{j=1}^{32} m_j}, \qquad \sum_{j=1}^{32} m_j = 650. \tag{C.2}$$

This specifies the weighting used for pooled checkpoint means; trajectories remain the independent experimental units. The comparison evaluates geometric coverage and action evidence retained at the sampled history checkpoints.

C.3 Low Energy Subspace and Controlled Replacement

Let V_{low} contain the centered history singular vectors after the 90% energy cutoff, and write $\tilde{x}_i = x_i - \bar{x}$. The block's low energy fraction and its alignment with the action are

$$L_i = \frac{\|V_{\text{low}}^\top \tilde{x}_i\|_2^2}{\|V^\top \tilde{x}_i\|_2^2}, \tag{C.3}$$

$$A_i = \cos\big(V_{\text{low}} V_{\text{low}}^\top \tilde{x}_i, V_{\text{low}} V_{\text{low}}^\top a\big)\,. \tag{C.4}$$

These quantities are evaluated when their denominators and cosine norms are nonzero. The targeted condition removes a retained block with high A_i; the matched condition removes a low alignment block while matching representation similarity, recency, and serialized length. The two conditions retain the same number of blocks. These definitions specify the intervention reported in Table 2.

The spectral visualization groups ordered directions into rank bands, whereas the intervention uses the cumulative energy cutoff above. These definitions distinguish visualization by spectral rank from intervention by accumulated energy and specify the basis directions used in each analysis.

D Ablation and Hyperparameter Analysis

D.1 Selectors at the Same Retained Count

Table D.1 compares selection strategies at the same retained count on 32 trajectories and 59 checkpoints. This comparison isolates which records are selected from how many records are retained. The evidence first selector achieves the highest reported final projection and next action Top 3 retention while retaining high geometric coverage.

Table D.1: Offline selector comparison at the same retained count on 32 trajectories and 59 checkpoints.

Selector	Action	Final	Goal	Top 3	Centroid
Recent K	0.96	0.88	0.82	0.76	0.97
Random K	0.92	0.92	0.95	0.45	0.99
Cosine Top K	0.86	0.91	0.97	0.53	0.96
Novelty Top K	0.93	0.93	0.96	0.33	0.99
GEM	0.95	0.95	0.96	0.78	0.99

D.2 Contribution of Evidence Protection

Table D.2 reports the component ablation at each variant's natural retained fraction. Relative to geometry only selection, the complete protection bundle retains more action related information. Removing State or Error protection reduces Top 3 retention, supporting their role within the combined evidence set. The natural budget comparison complements the fixed count comparison above; their retained fractions and checkpoint protocols are reported separately.

Table D.2: Offline component ablation at each variant's natural retained fraction (32 trajectories, 59 checkpoints).

Variant	Retention	Action	Top 3	Centroid
Full constraints	0.46	0.95	0.62	0.99
Geometry only	0.39	0.91	0.25	0.98
Recent + Goal only	0.45	0.95	0.57	0.98
Without State	0.45	0.95	0.53	0.98
Without Error	0.45	0.95	0.56	0.98
All candidates	0.44	0.95	0.59	0.99
Maximum basis $K = 12$	0.44	0.95	0.61	0.98
Maximum basis $K = 20$	0.47	0.95	0.63	0.99

These results use each variant's natural retained fraction. The same K comparison above controls retained count to assess selection quality at a common budget.

For the online ablation in Table 5, **Random completion** retains the protected set and candidate pool and fills remaining slots by seeded sampling to match the greedy reference count. **Geometry only** removes the semantic protection groups while retaining the candidate policy, residual rule, and request validity bookkeeping. These definitions identify the evidence protection and geometric completion components varied in the online outcome comparison.

D.3 Budget and Coverage Sensitivity

The exact K sweep uses 28 trajectories and 525 checkpoints common to all budgets. Table D.3 gives the complete tested range $K = 8, \ldots, 20$, including all four selectors. Evidence constrained selection exceeds geometry only Top 3 retention throughout this range.

The coverage sweep uses 32 trajectories and 59 checkpoints. Table D.4 shows the retained fraction, next action Top 3 retention, and action projection for every tested threshold. Increasing τ progressively adds geometric coverage and retains additional task evidence. The $\tau = 0.90$ row is the operating point highlighted in the main text.

Table D.3: Exact retained count sweep. Values are next action Top 3 retention on 28 trajectories and 525 checkpoints common to all K.

K	Geometry only	Evidence constrained	Recent K	Recent+Goal
8	0.18	0.56	0.50	0.51
9	0.21	0.56	0.53	0.52
10	0.24	0.57	0.55	0.53
11	0.27	0.59	0.54	0.54
12	0.31	0.60	0.56	0.56
13	0.33	0.62	0.58	0.57
14	0.36	0.62	0.60	0.59
15	0.42	0.64	0.63	0.61
16	0.46	0.67	0.66	0.63
17	0.50	0.68	0.68	0.66
18	0.54	0.71	0.69	0.67
19	0.59	0.74	0.71	0.72
20	0.66	0.78	0.74	0.74

Table D.4: Coverage threshold sweep on 32 trajectories and 59 checkpoints.

τ	Retained fraction	Top 3	Action projection
0.85	0.38	0.58	0.94
0.86	0.39	0.58	0.94
0.87	0.41	0.58	0.95
0.88	0.43	0.61	0.95
0.89	0.44	0.62	0.95
0.90	0.48	0.64	0.96
0.91	0.48	0.64	0.96
0.92	0.49	0.64	0.96
0.93	0.50	0.64	0.96
0.94	0.51	0.64	0.96
0.95	0.53	0.65	0.97

These sweeps provide the tabulated values corresponding to Figure 5. They evaluate representation level evidence retention under controlled compression settings.

E Evaluation Protocol and Efficiency Analysis

E.1 Task Units and Aggregate Metrics

The online evaluation uses WorkBuddyBench (Team et al., 2026) with DeepSeek-V4-Flash (Xu et al., 2026) at temperature zero. A complete task is the unit for reward and token aggregation. For a domain d with N_d tasks, define

$$\mathrm{Score}_d = \frac{100}{N_d} \sum_{i \in d} R_i, \qquad \mathrm{Tokens}_d = \frac{1}{N_d} \sum_{i \in d} T_i. \tag{E.1}$$

Overall Full260 means weight domains by task count. The fixed 40 panel contains ten tasks per domain, and the 20 task online ablation contains five per domain. All three use the same domain order: Code, Office, Security, and Web.

The trajectory level geometric analyses evaluate a different question from end to end reward. Long context understanding is likewise studied through context use analyses and dedicated benchmarks (Liu et al., 2024a; Bai et al., 2024; 2025); here the primary online outcome is the reward of a completed agent task.

E.2 Token Accounting and Full260 Efficiency

For each counted call c in task i, let I_{ic} and O_{ic} denote model input and output tokens, and let T_i^{enc} be embedding usage. Combined tokens are

$$T_i = \sum_c (I_{ic} + O_{ic}) + T_i^{\text{enc}}. \tag{E.2}$$

When usage is partitioned, $I_{ic} = I_{ic}^{\text{uncached}} + I_{ic}^{\text{cached}}$ and $O_{ic} = O_{ic}^{\text{normal}} + O_{ic}^{\text{reasoning}}$. These component counts partition the model input and output totals. Combined tokens equal the sum of the four model components and the selector's embedding usage.

Table E.1: Full260 token composition, mean per task. K and M denote thousands and millions of tokens. Values use the same reporting precision as the main text.

Token category	**Base agent**	**GEM**
Uncached input	48.70K	201.48K
Cached input	2.60M	1.82M
Normal output	12.40K	15.68K
Reasoning output	25.38K	22.81K
Model total	2.69M	2.06M
Embedding	0	46.97K
Combined	2.69M	**2.11M**

Table E.1 explains the net reduction in Table 3: combined usage decreases from 2.69M to 2.11M per task after accounting for embedding cost. The reduced cached input volume more than offsets the additional embedding and uncached input volume. The category breakdown accounts for every component of the combined total.

Token accounting complements studies of physical memory management in LLM serving (Kwon et al., 2023). The combined token measure quantifies processed token volume, including cached and uncached input, model output, and embedding usage, and provides the basis for the token efficiency comparison.

E.3 Reference Panel and Online Ablation Protocol

The fixed 40 method comparison is reported in Table 4; the online ablation with 20 tasks is reported in Table 5. This appendix specifies the aggregation and method definitions for both evaluations, with task means and domain summaries reported in their corresponding tables.

The fixed 40 panel combines results from distinct evaluation snapshots. The GEM row is extracted from Full260 using the reference task identifiers, while the reference methods come from the fixed 40 runs. The online ablation uses the same protected set definitions, candidate policy, and valid request structure, with the completion or protection component varied as described in Appendix D.