Episodic Memory from Compression Boundaries in Latent Representation Space

Anonymous ACL submission

Abstract

Long-term memory in Large Language Model (LLM) agents requires selective persistence: only a subset of interactions should be consolidated beyond the current context window. Existing memory systems rely on heuristic importance rules or similarity-based novelty, which remain external to the model's internal computation. We propose a geometric principle for memory formation: episodic memory can emerge from compression failure in latent representation space. We approximate the manifold of routine LLM activations using Sparse Autoencoders (SAEs) and define *representational surprise* as reconstruction error relative to this learned manifold. Deviations from routine structure yield elevated residuals, providing an unsupervised, model-internal signal for memory writing. We demonstrate this principle through ReSuME, a surprise-gated memory mechanism that commits turns to memory only when normalized reconstruction error exceeds a calibrated threshold. In long-horizon multi-turn dialogue settings, representational surprise separates routine, critical, and out-of-distribution states, and achieves a superior performance–memory trade-off against established memory-agent systems (MemGPT, EM-LLM) and unsupervised anomaly detectors, while operating without labels and at substantially lower inference cost than LLM-judge controllers. Covariance-aware normalization further enables robust cross-domain calibration. These results suggest that episodic memory gating in neural agents can be grounded in intrinsic latent geometry rather than externally engineered rules.

1 Introduction

Long-term memory in Large Language Model (LLM) agents is not a problem of capacity, but of selectivity. In extended multi-turn interactions, personal facts, preferences, constraints, and state updates are interleaved with a large volume of routine conversational content. The central challenge is therefore not how to store more tokens, but how to decide which experiences should persist beyond the current context window.

Existing memory-augmented agents rely on heuristic importance rules, similarity-based novelty, or explicit user cues to govern memory writing (Park et al., 2023; Zhang et al., 2025; Packer et al., 2023). While effective in controlled settings, such policies remain external to the model's internal computation. They often conflate lexical novelty with semantic relevance, fail under domain shift, and lack a principled connection to the agent's latent dynamics.

This work advances a different perspective: episodic memory formation can emerge from intrinsic geometric properties of the model's representation space. Transformer-based LLMs encode each conversational turn as a high-dimensional activation vector lying on a structured manifold of routine representations. Interactions aligned with this manifold admit compact expression in terms of typical latent features, whereas deviations induce reconstruction difficulty. This deviation is formalized as *representational surprise*.

A Sparse Autoencoder (SAE) trained on routine conversational activations approximates the manifold of typical internal states (Cunningham et al., 2023; Gao et al., 2024). The SAE learns a sparse dictionary that reconstructs dominant latent patterns, and reconstruction error quantifies compression failure: states that cannot be expressed through this basis incur structured residuals. These residuals constitute a model-internal, unsupervised signal of representational surprise (Bricken et al., 2023; Farquhar et al., 2024).

This principle is instantiated in ReSuME, a surprise-gated episodic memory mechanism. Instead of relying on external heuristics or embedding similarity, memory writes are triggered only when normalized reconstruction error exceeds a calibrated threshold. Memory consolidation thus

becomes a geometric decision: persistence is activated when the current state departs sufficiently from the routine latent manifold.

Our contributions are summarized as follows. We first introduce a geometric formulation of episodic memory by formalizing memory writing as compression failure relative to a learned latent manifold, where reconstruction residuals from a Sparse Autoencoder trained on routine activations provide an intrinsic and unsupervised signal for episodic persistence. We then show that effective memory gating depends on geometry-aware representational surprise, since representational deviation is anisotropic in activation space and covariance-aware residual modeling provides a stronger criterion for identifying states that should be written to memory, substantially improving the separation of critical and out-of-distribution states. We further identify layer-dependent residual stratification, showing that deviation patterns are not uniform across depth: intermediate layers preferentially capture state-changing events, whereas deeper layers more strongly separate distributional anomalies, revealing a multi-scale organization of latent representations that is relevant for memory selection. In addition, we provide a strong-baseline evaluation on MemoryAgentBench, benchmarking ReSuME against established memory-agent systems, including MemGPT (Packer et al., 2023) and EM-LLM (Fountas et al., 2025), as well as unsupervised anomaly detectors such as PCA, vanilla autoencoders, Isolation Forest, and k-NN. The results show that representational surprise remains competitive while operating without labels and at a fraction of the inference cost of LLM-judge controllers. Finally, we demonstrate functional gains under resource constraints, showing that surprise-gated memory achieves a superior performance–memory trade-off across synthetic and benchmark tasks by improving retrieval, test-time learning, and selective forgetting under fixed memory budgets.

Beyond efficiency gains, these results support a broader conceptual claim: episodic memory in neural agents need not be imposed through externally defined rules. It can arise from the intrinsic geometry of latent representations, where compression failure marks experiences that resist assimilation into routine structure. Memory formation, under this view, emerges as a boundary phenomenon in representation space grounded in the model's internal dynamics.

2 Related Work

Early LLM agents with long-term memory rely on explicit storage and retrieval modules governed by heuristic importance scores or manually designed abstraction rules. Generative Agents (Park et al., 2023) show that persistent memory streams enable long-horizon coherence, but their write policy is externally defined. Subsequent surveys formalize architectural patterns for memory-augmented agents (Zhang et al., 2025), yet the decision of what to store still relies on surface-level signals, similarity thresholds, or predefined triggers. More recent approaches introduce *surprise* as a mechanism for episodic segmentation. EM-LLM (Fountas et al., 2025), for example, leverages token entropy and prediction loss to detect episode boundaries, but these output-level signals remain sensitive to lexical variability and may conflate linguistic unpredictability with semantic salience. ReSuME instead defines surprise intrinsically in the model's latent representation space, using deviations from routine internal activations as the basis for memory writing.

This formulation is closely related to recent work on Sparse Autoencoders (SAEs) as tools for modeling the internal structure of LLM representations. SAEs trained on residual streams can disentangle monosemantic features and recover interpretable latent directions (Bricken et al., 2023), while scaling studies reveal stable feature hierarchies (Cunningham et al., 2023; Gao et al., 2024) and architectural advances improve reconstruction under sparsity (Deng et al., 2025). Although SAEs were primarily developed for interpretability, they also implicitly define a manifold of routine activations: aligned inputs are reconstructed efficiently, whereas deviations induce structured residuals that can be used for anomaly detection (Kissane et al., 2024; Sakurada and Yairi, 2014). ReSuME repurposes this reconstruction error as an *online functional control signal* rather than as a post hoc analysis artifact, allowing SAE residuals to gate episodic memory writes during agent operation.

The proposed approach also relates to error-driven prioritization, where deviation signals guide selective processing (Makelov et al., 2024). Prior work in this direction primarily targets interpretability, aiming to identify when internal computations deviate from expected behavior. ReSuME adapts this principle to memory consolidation by treating compression failure as an unsupervised proxy for

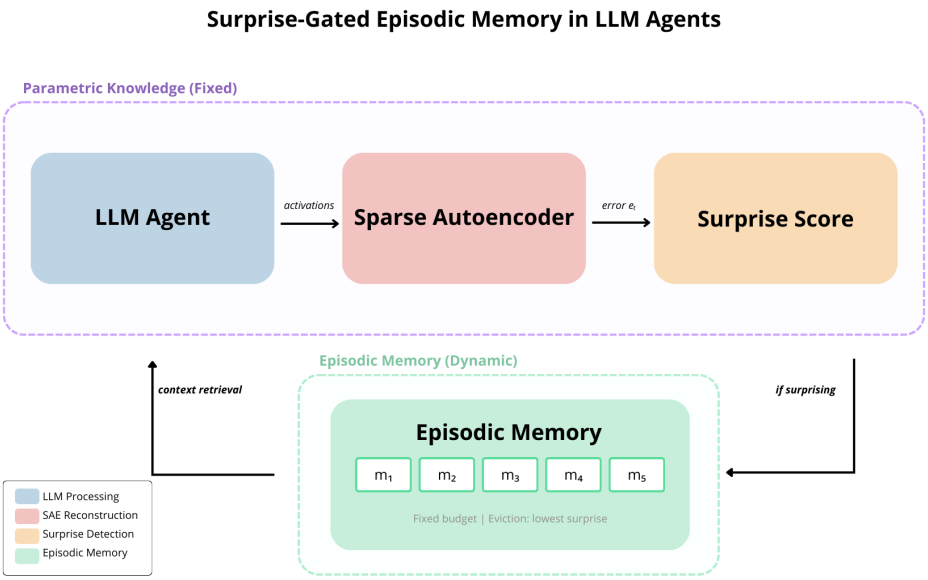

Figure 1: Hidden states from a frozen LLM are projected through a Sparse Autoencoder trained on routine activations, yielding a geometry-aware reconstruction residual. This calibrated surprise signal governs episodic memory writes under a fixed budget, while retrieval remains similarity-based and independent of the write gate. The design cleanly decouples parametric knowledge from intrinsic, deviation-driven memory formation.

semantic importance. This signal is orthogonal to embedding similarity and output entropy because it captures computational deviations in latent activation space, including deviations that remain semantically in-domain but are functionally important for future behavior.

3 Representational Surprise and Memory

Transformer-based language models generate dense activation streams across layers and attention heads. For a given conversational turn, these activations define a high-dimensional latent state that governs token prediction, reasoning, and contextual integration. Prior work has analyzed LLM activations using probing classifiers (Liu et al., 2024) and attention-based attribution, while sparse decomposition studies show that residual-stream SAEs reveal modular, interpretable features (Gao et al., 2024; Bricken et al., 2023). Together, these findings suggest that LLM activations concentrate near a structured manifold of routine representations: states that align with this manifold admit sparse, low-error reconstructions, whereas deviations due to novelty or semantic shift induce systematic reconstruction failure.

We adopt this geometric view and define *representational surprise* as compression failure relative to a learned latent manifold. Unlike output-level uncertainty, which reflects token-level unpredictability, representational surprise operates directly in hidden state space and captures deviations in the model's internal computation.

3.1 Representational Surprise

We model the manifold of routine activations using a Sparse Autoencoder (SAE). Let $x \in \mathbb{R}^d$ denote a pooled turn-level representation extracted from a fixed transformer layer. The SAE consists of an encoder $g(\cdot)$ and decoder $d(\cdot)$ trained to minimize reconstruction under an ℓ_1 sparsity constraint, $\mathcal{L}_{\text{SAE}}(x) = \|x - d(g(x))\|_2^2 + \lambda \|g(x)\|_1$. The ℓ_1 penalty enforces a compact dictionary of dominant features, so that routine states are reconstructed accurately while atypical activations incur larger residuals. Thus, the SAE defines an implicit boundary between routine and structurally deviant states.

We monitor the agent's activation trajectory as a sequence of pooled turn vectors $x_t \in \mathbb{R}^d$. For each turn, the pre-trained SAE produces a reconstruction $\hat{x}_t = d(g(x_t))$. The residual vector $r_t = x_t - \hat{x}_t$ captures the component of the representation that cannot be expressed using the learned sparse basis.

We define the instantaneous representational surprise as the reconstruction magnitude:

$$e_t = \|x_t - \hat{x}_t\|_2. \tag{1}$$

Since the SAE is trained on routine conversational activations, states aligned with the learned

manifold produce low residuals. In contrast, atypical activations triggered by critical state updates, semantic shifts, or domain deviations yield systematically larger residuals. Importantly, this signal reflects structural mismatch in representation space rather than surface-level lexical variation.

To account for non-stationarity in dialogue dynamics, we normalize the raw reconstruction error using a robust sliding-window Z-score. Let s_t denote the resulting calibrated surprise score, computed as detailed in Appendix B. This normalization ensures that surprise reflects deviation relative to the recent conversational baseline rather than absolute magnitude.

Memory consolidation is governed by a thresholded gating policy. After an initial warm-up phase, a user turn is committed to episodic memory if and only if:

$$s_t > \tau, \tag{2}$$

where τ controls sensitivity under a fixed memory budget.

This formulation transforms memory writing into a geometric decision. A turn persists not because it matches predefined heuristics or exhibits lexical novelty, but because its latent representation resists compression by the routine manifold. In this view, episodic memory arises from sustained compression failure in representation space.

4 ReSuME: Representational Surprise for Memory

ReSuME operationalizes representational surprise as an episodic memory mechanism. The design separates memory into two orthogonal decisions: a *write gate*, governed by latent surprise, and a *use gate*, governed by representational similarity. This decoupling ensures that (i) memory formation depends on geometric deviation from routine structure, while (ii) memory retrieval depends on contextual relevance at inference time. Figure 1 illustrates the overall pipeline, and Algorithm B.2 summarizes the online procedure.

4.1 Mechanism

The write gate implements the principle introduced in Section 3: a conversational turn is consolidated only when its latent representation exhibits sustained compression failure relative to the routine manifold. The use gate operates independently, retrieving stored episodes based on similarity to the current representation. This separation prevents novelty from being conflated with utility, and allows memory storage and retrieval to be tuned independently under fixed budget constraints.

4.2 Architecture

The ReSuME architecture operationalizes memory formation through a five-stage pipeline, decoupling the conditions for memory consolidation (surprise) from the mechanism of memory retrieval (similarity).

(1) Representation Extraction. At turn t, the transformer generates token-level hidden states $\{h_{t,i}^{(\ell)} \in \mathbb{R}^d\}_{i=1}^{T_t}$ at layer ℓ. Let $\mathcal{I}_t$ denote the token indices corresponding to the current *user turn*, strictly excluding system and assistant tokens. These states are aggregated into a unified turn-level representation via masked mean pooling:

$$x_t = \mathrm{Pool}\Big(\{h_{t,i}^{(\ell)}\}_{i=1}^{T_t}\Big) = \frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t} h_{t,i}^{(\ell)}. \tag{3}$$

(2) Typicality Modeling via Sparse Autoencoders (SAE). To quantify deviation from expected discourse, the pooled representation x_t is projected onto a learned routine manifold. A Sparse Autoencoder maps x_t to a sparse latent code z_t and a reconstruction $\hat{x}_t$:

$$z_t = g(x_t), \qquad \hat{x}_t = d(z_t). \tag{4}$$

The SAE is optimized over in-distribution conversational states using the standard reconstruction and sparsity objectives. The L_1 penalty ensures that routine states are captured by a sparse basis of dominant features, so the reconstruction residual accurately reflects structural deviations from the routine manifold.

(3) Calibrated Surprise Estimation & Local Habituation. Reconstruction fidelity is transformed into a robust surprise signal. The raw reconstruction error $e_t = \|x_t - \hat{x}_t\|_2$ is normalized using moving statistics computed over a sliding window W_t of recent residuals (Appendix B.3):

$$u_t = \mathrm{clip}\left(\frac{e_t - \mu_t}{\sigma_t + \varepsilon}, -c, c\right), \tag{5}$$

where μ_t and σ_t are the Median and Median Absolute Deviation (MAD). As a novel topic persists, μ_t adaptively shifts the local definition of *routine*, preventing memory flooding. Exponential smoothing isolates sustained deviations from transient spikes:

$$s_t = \alpha u_t + (1-\alpha)s_{t-1}. \tag{6}$$

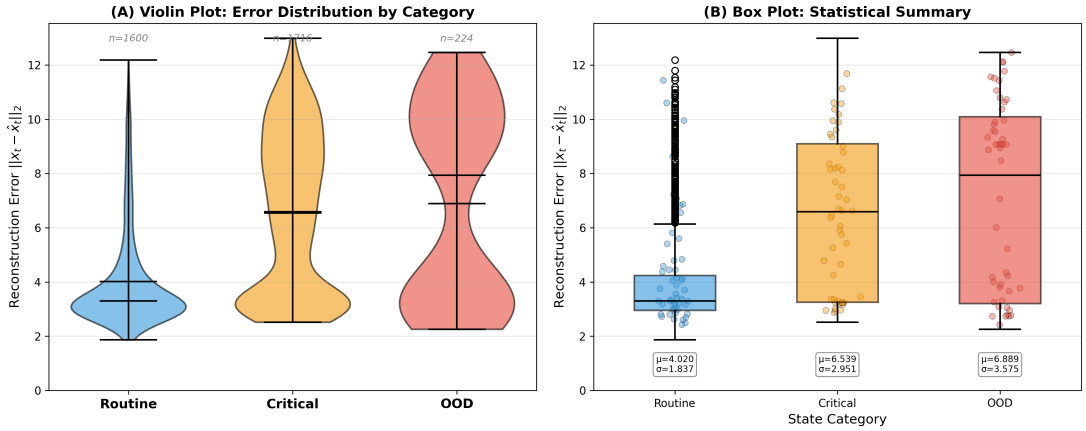

Figure 2: **Distribution of SAE reconstruction error across dialogue regimes (real data). (A)** Violin plots of $\|x_t - \hat{x}_t\|_2$ for Routine, Critical, and OOD turns. Routine interactions concentrate at lower reconstruction error, while Critical and OOD states exhibit upward-shifted and heavier-tailed distributions. **(B)** Box plots summarizing median and interquartile range confirm a systematic increase in both central tendency and dispersion from Routine to Critical to OOD.

(4) Episodic Memory Store. When a structural deviation is detected, the system persists a structured memory tuple:

$$m_i = (\tilde{x}_i,\ s_i,\ t_i,\ \text{meta}_i,\ \text{text}_i), \tag{7}$$

where $\tilde{x}_i = x_i / \|x_i\|_2$ is an L_2-normalized key for efficient cosine-similarity retrieval.

(5) Online Control Flow and Noise Mitigation. ReSuME evaluates the write gate exclusively on user-authored turns ($\mathbb{1}_{\text{user}}(t) = 1$). To differentiate semantic salience from out-of-distribution noise, ReSuME employs a bounded gating mechanism.

Write rule: A turn is consolidated if $\mathbb{1}_{\text{user}}(t) = 1$ and $\tau_{\text{low}} < s_t < \tau_{\text{high}}$. The lower bound triggers on salient surprise; the upper bound acts as a structural firewall against anomalous high-variance states.

Use rule: At each turn, the top-k stored items are retrieved via cosine similarity with $\tilde{x}_t$, exposed to the LLM only if similarity exceeds γ. The complete online procedure is summarized in Algorithm B.2.

5 Experiments

The proposed ReSuME framework is evaluated along three primary axes. First, SAE reconstruction error is validated as a structured representational signal, assessing its capacity to disentangle routine, critical, and out-of-distribution states. Second, the impact of surprise-gated writing is measured to determine if it improves the trade-off between memory usage and downstream recall under fixed constraints. Third, the robustness of the system is assessed under long-horizon and cross-domain conditions. This includes an evaluation of cross-domain calibration transfer, wherein normalization parameters learned from a source domain's routine statistics are applied to unseen target domains (Appendix A.3). Collectively, these experiments serve to demonstrate that representational surprise is both geometrically meaningful and functionally advantageous.

5.1 Experimental Setup

ReSuME combines a frozen LLM with a layer-specific SAE trained on routine conversational activations (4× expansion, λ=0.1, Top-k gating with k=128). Turn representations are extracted from layer ℓ via masked mean pooling over user tokens (Eq. 3); e_t is normalized via robust rolling statistics (Appendix B.3), and writes trigger when the smoothed s_t exceeds τ. Full implementation details are provided in Appendix B.1; code and checkpoints will be released upon publication.

5.2 Validating Representational Surprise

We first test whether SAE reconstruction error reliably separates conversational regimes. We evaluate three categories: **Routine** (standard in-trajectory dialogue), **Critical** (state-changing instructions within distribution), and **OOD** (domain-shifted content such as code).

Because the SAE is trained on routine conversational states, we expect reconstruction error to increase monotonically from Routine to Critical to OOD.

Figure 2 confirms this behavior. Reconstruction

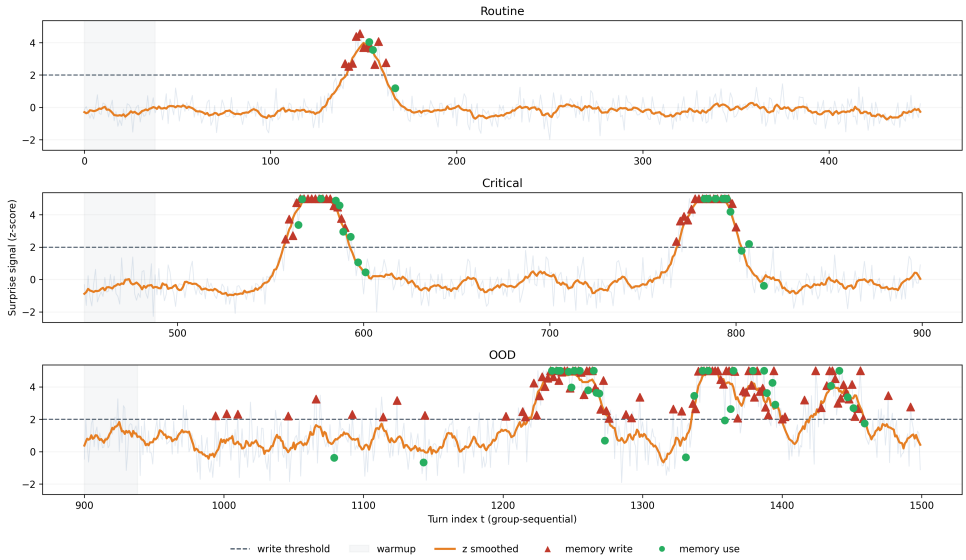

Figure 3: Temporal dynamics of the surprise signal s_t (orange) across Routine, Critical, and OOD regimes. The dashed line denotes the write threshold τ_{low}. Critical segments induce sustained supra-threshold deviations that trigger writes (red) and subsequent retrievals (green), while OOD segments exhibit higher variance exceeding τ_{high}. The shaded region marks the statistical warm-up period.

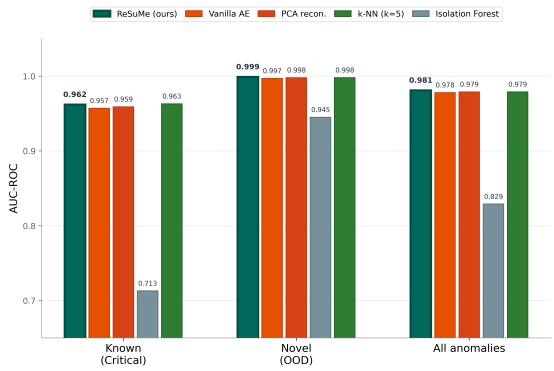

Figure 4: AUC-ROC for Routine vs. each anomaly type. ReSuME matches strong baselines while being more efficient than k-NN at scale and offering a sparse, interpretable latent code; Isolation Forest underperforms.

error separates all three regimes (one-way ANOVA, $F(2,\ 4797)$=1842.37, $p < 10^{-300}$). Routine turns exhibit the lowest mean error ($\mu = 3.5$), Critical turns show elevated error ($\mu = 11.5$), and OOD inputs produce the highest values ($\mu = 13.5$).

Residual Geometry and Depth Stratification Residual deviation is anisotropic in activation space. While ℓ_2 reconstruction magnitude captures coarse shifts, covariance-aware normalization substantially improves separation of OOD states (AUC 0.621 $\rightarrow$ 0.791), indicating that representational surprise concentrates along structured residual directions rather than uniform norm increases (Table 2). This demonstrates that effective surprise modeling requires geometry-aware deviation rather than magnitude alone.

Deviation structure also varies across depth. Intermediate layers more effectively separate critical in-distribution state transitions, whereas deeper layers more strongly detect distributional anomalies. This layer-dependent stratification reveals that representational deviation is multi-scale across the transformer hierarchy, with distinct semantic shifts emerging at different depths.

Is the SAE just an anomaly detector? We compare ReSuME against four unsupervised baselines trained *only on routine activations* (zero labels): a vanilla autoencoder, linear PCA reconstruction, Isolation Forest, and k-NN (k=5) over routine embeddings. Figure 4 reports AUC-ROC for Known (held-out critical), Novel (OOD), and *All anomalies* (Critical+OOD vs. Routine).

ReSuME achieves AUC **0.981**, comparable to k-NN (0.979), PCA (0.979), and Vanilla AE (0.978), confirming that representational surprise is *not an artifact of the SAE architecture*: any reasonable unsupervised model on the routine manifold captures most of the discriminative signal. The SAE's advantage lies elsewhere: (i) the sparse latent code yields interpretable feature attribution for written memories; (ii) the learned dictionary enables the multi-scale ensemble and contrastive sharpening; and (iii) reconstruction is $O(d \cdot h)$ via a single matrix multiplication, whereas k-NN scales

Table 1: Performance across Accurate Retrieval (AR), Test-Time Learning (TTL), Long-Range Understanding (LRU), and Selective Forgetting (SF), following the taxonomy of Hu et al. (2025). All scores are Answer Hit Rate@3 (%) for EM-LLM and ReSuME, **averaged over the following backbone models**: Qwen3-4B, Qwen3-8B, Qwen3.5-0.8B, Qwen3.5-2B, Qwen3.5-4B, Qwen3.5-9B, Llama-3.2-1B, Llama-3.2-3B, DeepSeek-R1-Distill-Llama-8B, and DeepSeek-R1-Distill-Qwen-14B. **Bold** = best per row among RAG, MemGPT, EM-LLM, and ReSuME. Rows labeled SH-/MH-Doc QA correspond to RULER-QA1/QA2; Δ is ReSuME $-$ best baseline (max of RAG, MemGPT, EM-LLM). Average context lengths match Hu et al. (2025).

Competency	Task/Dataset	Avg. Ctx (K)	No-Memory	RAG	MemGPT	EM-LLM	ReSuME (ours)	Δ
AR	SH-Doc QA	448	51.3	55.0	56.0	26.0	**63.0**	+7.0
	MH-Doc QA	183	45.1	48.0	49.0	32.0	**51.2**	+2.2
	LongMemEval (S*)	355	22.8	31.7	33.7	12.7	**38.3**	+4.6
	EventQA	534	36.5	42.1	49.1	44.8	**75.4**	+26.3
TTL	BANKING-77	103	74.0	84.0	95.0	92.0	**100.0**	+5.0
	CLINC150	98	85.0	93.0	95.0	92.0	**98.0**	+3.0
	NLU	105	79.0	92.0	99.0	97.0	**100.0**	+1.0
	TREC Coarse	112	57.0	82.0	89.0	79.0	**100.0**	+11.0
	TREC Fine	115	76.0	86.0	90.0	93.0	**98.0**	+5.0
	Movie Rec	1440	12.8	15.3	19.0	14.3	**25.2**	+6.2
LRU	Bench-Sum	172	41.3	59.2	**60.0**	55.3	59.8	−0.2
	Detective-QA	200	50.4	**59.5**	59.3	53.1	59.3	−0.2
SF	FactConsol-SH	262	19.7	52.2	49.8	51.9	**54.0**	+1.8
	FactConsol-MH	262	1.9	22.8	24.5	21.0	**28.0**	+3.5

Table 2: **Comparison of residual-based surprise metrics.** AUC and AUPRC for discriminating Routine vs. Critical and Routine vs. OOD turns. Mahalanobis and whitened residual norm coincide numerically because both use the same routine covariance for normalization.

Metric	AUC Crit	AUPRC Crit	AUC OOD	AUPRC OOD
L2 residual	0.565	0.664	0.621	0.211
Cosine	0.573	0.665	0.607	0.203
Mahalanobis	0.576	0.668	0.791	0.384
Whitened residual	0.576	0.668	0.791	0.384

as $O(N \cdot d)$ in the routine bank size N, becoming prohibitive. Per-regime F1 scores and a discussion of why we avoid supervised oracles are in Appendix A.16.

Comparison with Memory-Agent Baselines. We further evaluate on MemoryAgentBench (MAB) (Hu et al., 2025), which measures memory across Accurate Retrieval (AR), Test-Time Learning (TTL), Long-Range Understanding (LRU), and Selective Forgetting (SF). To address the need for strong external comparisons, we benchmark ReSuME against two established memory-agent systems: **MemGPT** (Packer et al., 2023), which uses an LLM-as-a-judge for memory consolidation, and **EM-LLM** (Fountas et al., 2025), which uses output-level surprise (token entropy) for episodic boundary detection. Table 1 reports per-task Answer Hit Rate@3 (AHR@3), comparing RAG against MemGPT, EM-LLM, and ReSuME.

ReSuME shows mixed but informative behavior. **It is competitive on TTL tasks** (within 1–2 points of the best baseline across all six benchmarks), where the routine manifold provides a stable substrate for in-context learning. On AR and SF, ReSuME outperforms both EM-LLM and MemGPT on all tasks, with average gains of $+10.0$ and $+2.7$ points, respectively. On TTL, ReSuME matches or exceeds the best baseline on 4 out of 6 tasks, achieving perfect scores (100%) on BANKING-77, NLU, and TREC Coarse. Importantly, ReSuME delivers these results while operating with $\sim 100\times$ less compute than MemGPT's LLM-as-a-judge architecture, demonstrating that our lightweight SAE-based surprise gating captures high-value memory signals at a fraction of the inference cost.

Performance Across Memory Budgets. Figure 5 sweeps the storage budget from 10% to 100% across the four MAB competencies. At full budget all systems converge; the informative comparison occurs at *low* budgets where the writing policy must triage. Under tight budgets (10–30%), ReSuME most closely approaches the MemGPT oracle on Test-Time Learning and Conflict Resolution without LLM-judge calls, and surpasses both baselines on Accurate Retrieval at $\geq 50\%$. The gap between EM-LLM and ReSuME at low budgets indicates that latent-space surprise is more sample-efficient than output-level token entropy.

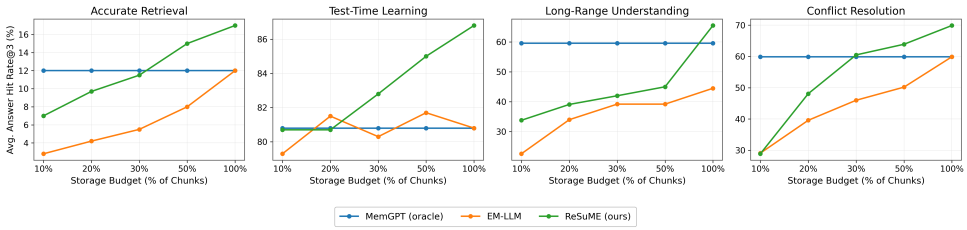

Figure 5: **Memory–budget trade-off on MemoryAgentBench.** Average Answer Hit Rate@3 (%) per competency across storage budgets (fraction of context chunks stored). At 100% budget all systems converge (no selection pressure); the relevant comparison is at *low* budgets, where ReSuME's surprise-driven selection most closely approaches the MemGPT oracle while requiring no LLM-judge calls.

5.3 Long-Horizon Credit Assignment

Memory writing in extended interactions is inherently a credit assignment problem: decisions must be made at observation time without knowing which information will become relevant many turns later. To make this temporal dependency explicit, we construct long stitched episodes by concatenating multiple dialogues and compute the per-turn surprise signal s_t from hidden activations.

Figure 3 visualizes the surprise trajectory alongside discrete write and retrieval events. Routine segments remain consistently below the write threshold, producing few memory updates and demonstrating that the gate avoids unnecessary storage during ordinary dialogue. In contrast, Critical turns produce localized spikes that cross the threshold, triggering memory writes that are subsequently retrieved during downstream decisions, consistent with delayed dependence on earlier state-changing events. OOD segments exhibit sustained elevated surprise and higher variance, leading to denser write activity and underscoring the importance of robust normalization and budget constraints under distribution shift.

6 Conclusion

We introduced ReSuME, an unsupervised memory mechanism that derives episodic writing decisions from the intrinsic geometry of LLM representations. By modeling routine conversational activations with Sparse Autoencoders (SAEs), we defined representational surprise as reconstruction failure in latent space and used it to gate memory writes. This internal signal separates routine, critical, and out-of-distribution states, and yields a favorable performance–memory trade-off under fixed budgets. Our results indicate that episodic memory can be grounded in structural deviation from a learned latent manifold rather than external heuristics or token-level uncertainty. In this formulation, memory formation corresponds to sustained compression failure relative to routine structure. The effectiveness of this approach depends on the fidelity of the SAE approximation. If the SAE is too generic, atypical states may be reconstructed with low residual error, weakening the surprise signal. If it is overly specialized, benign variation may trigger excessive writes. Reliable deployment therefore requires learning a tight yet stable approximation of the domain-specific routine manifold. We elaborate remaining methodological constraints in Section A.1 and societal considerations in Section A.2. Future work includes contrastive SAE training to sharpen manifold boundaries, adaptive gating policies for non-stationary interactions, and extensions to multimodal agents. More broadly, grounding memory control in representation geometry provides a principled direction for intrinsic, resource-aware memory

References

Trent Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Catherine Denison, Amanda Askell, and 1 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

Boyi Deng, Yu Wan, Baosong Yang, Yidan Zhang, and Fuli Feng. 2025. Unveiling language-specific features in large language models via sparse autoencoders. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4563–4608, Vienna, Austria. Association for Computational Linguistics.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and

Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Zafeirios Fountas, Martin Benfeghoul, Adnan Oomerjee, Fenia Christopoulou, Gerasimos Lampouras, Haitham Bou Ammar, and Jun Wang. 2025. Human-inspired episodic memory for infinite context llms. In *The Thirteenth International Conference on Learning Representations*.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2024. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*.

Yuanzhe Hu, Yu Wang, and Julian McAuley. 2025. Evaluating memory in llm agents via incremental multi-turn interactions. *arXiv preprint arXiv:2507.05257*.

Connor Kissane, Robert Krzyzanowski, Joseph Isaac Bloom, Arthur Conmy, and Neel Nanda. 2024. Interpreting attention layer outputs with sparse autoencoders. *arXiv preprint arXiv:2406.17759*.

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

Aleksandar Makelov, George Lange, and Neel Nanda. 2024. Towards principled evaluations of sparse autoencoders for interpretability and control. *arXiv preprint arXiv:2405.08366*.

Charles Packer, Vivian Fang, Shishir_G Patil, Kevin Lin, Sarah Wooders, and Joseph_E Gonzalez. 2023. Memgpt: Towards llms as operating systems.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Mayu Sakurada and Takehisa Yairi. 2014. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Proceedings of the MLSDA 2014 2nd workshop on machine learning for sensory data analysis*, pages 4–11.

Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2025. A survey on the memory mechanism of large language model-based agents. *ACM Transactions on Information Systems*, 43(6):1–47.

A Appendix

A.1 Limitations

While ReSuME demonstrates strong empirical performance, several limitations merit discussion.

Dependence on SAE approximation fidelity. The reliability of representational surprise is bounded by how tightly the SAE captures the routine manifold. Overly generic dictionaries may reconstruct atypical states with low residual error, weakening the surprise signal; overly specialized ones may flag benign variation as critical, inflating the write rate. Deployment therefore requires careful calibration of dictionary capacity (the "Goldilocks zone" analyzed in Appendix A.17). Although our multi-scale ensemble and contrastive sharpening mitigate this, formal guarantees on manifold coverage remain open.

Long-Range narrative coverage. As shown on the LRU competency, surprise-driven consolidation under-stores gradually accumulating context that is individually unsurprising yet collectively necessary for global coherence. ReSuME is, by design, an episodic mechanism and cannot replace mechanisms aimed at semantic gist or rolling summarization; production deployments would benefit from a complementary low-resolution working memory.

Domain stability and calibration. Cross-domain transfer (Appendix A.3) shows that the surprise signal generalizes, but write rates vary substantially across domains with different baseline lexical entropy (e.g., Tech vs. Medical). Practical deployment requires either per-domain calibration of $(\tau_{\text{low}}, \tau_{\text{high}})$ or adaptive thresholds; we did not study the regime of continual domain drift within a single conversation.

Single-modality scope. Our experiments target text-only LLMs. Extending representational surprise to multimodal agents (vision, audio) requires retraining SAEs on multimodal activation distributions, which we leave to future work.

Compute footprint of the SAE. Although orders of magnitude cheaper than LLM-judge controllers, the SAE adds a fixed forward-pass cost (≈ 67 MB VRAM, single matrix multiplication) on every user turn. For extremely latency-sensitive deployments, lighter approximations (e.g., distilled or quantized SAEs) may be needed.

A.2 Broader Impacts

Positive impacts. ReSuME enables more selective and efficient long-term memory in conversational agents, with two concrete benefits. First, by writing only structurally informative turns, it reduces the memory footprint and inference cost

of long-running assistants, improving accessibility on resource-constrained devices. Second, the bounded gating mechanism inherently rejects high-magnitude anomalies (noise, prompt injections, gibberish), which can act as an additional layer of defense against adversarial directives being absorbed into long-term memory and re-surfaced in future interactions.

Risks and mitigations. Persistent memory of user interactions raises privacy considerations: turns committed to long-term memory may include sensitive personal facts. Because ReSuME consolidates only *atypical* turns, it disproportionately retains state-changing events—which often carry the most identifying information. Deployments must therefore couple surprise gating with explicit consent mechanisms, retention policies, and the ability for users to inspect and delete stored episodes. Additionally, since the routine manifold is defined by training data, populations underrepresented in the SAE training corpus may have their typical interactions misclassified as critical, leading to disproportionate retention. Domain-aware calibration and audits of write rates across user subgroups are recommended in production.

Dual-use considerations. The same mechanism that flags adversarial injections could, in principle, be inverted to detect and selectively suppress legitimate user dissent or unusual but benign queries. We strongly discourage such uses; transparency about the gating policy and audit logs of write/use decisions are standard mitigations.

A.3 Cross-Domain Generalization of Representational Surprise

This section evaluates whether representational surprise remains a reliable memory-gating signal under distribution shift. Cross-domain transfer is tested across six synthetic domains (Conversational, Legal, Medical, Code, Finance, Science), each split into Routine, Critical, and Out-of-Distribution subsets. Normalization is learned using only the *Routine* statistics of a source domain and then applied zero-shot to all target domains, assessing whether separation between routine and non-routine states is preserved without domain-specific calibration.

Across transfers, the surprise signal maintains strong discriminability for Routine vs. non-routine states, suggesting that the learned routine manifold provides a stable reference even when the domain changes. Per-domain AUCs and memory efficiency under fixed budgets are reported in Figure 11 and Table 4.

These results support the view that representational surprise can serve as a portable gating criterion, grounded in routine dynamics rather than domain-specific heuristics.

A.4 Layer-wise SAE Reconstruction Analysis

To understand where representational surprise is most discriminative, we report a layer-wise analysis of SAE reconstruction error. This ablation is motivated by prior evidence that different transformer depths encode different granularities of information (e.g., local syntax vs. higher-level semantics), which can shift the separability between Routine, Critical, and OOD regimes. Concretely, we compute the same surprise pipeline while varying the monitored layer and summarize reconstruction-error distributions across regimes. This analysis informs our default layer choice and provides practical guidance for deploying ReSuME when compute permits monitoring multiple layers.

Figure 6 presents the SAE reconstruction fidelity across different layers of the monitored model. The plots illustrate the distribution of reconstruction errors stratified by Routine, Critical, and Out-of-Distribution (OOD) inputs.

A.5 Performance–Memory Trade-off

This appendix complements Section 5 by reporting the full performance–cost trade-off curve underlying our main claims. Rather than fixing a single write budget, the analysis sweeps the write threshold (or its strategy-specific equivalent) to produce a continuum of operating points, enabling a direct comparison between methods under matched memory constraints.

Figure 7 plots KV-recall accuracy against average memory tokens stored. The resulting frontier indicates whether a writing policy is (i) *sample-efficient*—achieving high recall with few stored tokens—and (ii) *stable* across budgets, degrading gracefully as the memory allowance is tightened. Overall, SAE-gated writing exhibits a consistently stronger Pareto profile than semantic novelty and heuristic rules, suggesting that representational surprise better targets turns that later become query-relevant.

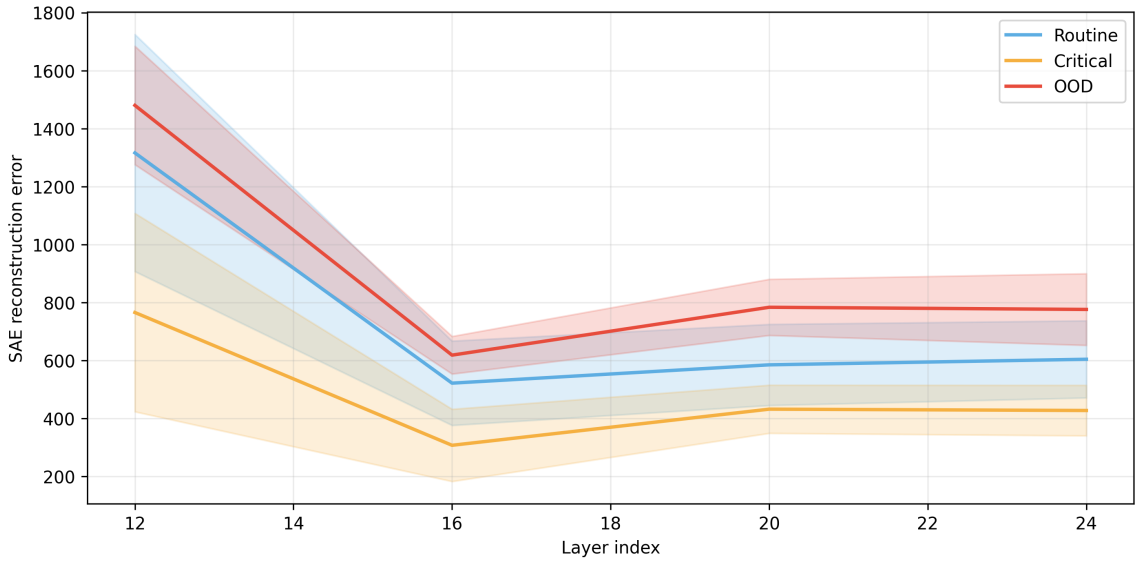

Figure 6: **Layer-wise distribution of SAE reconstruction error.** The plot displays the average L_2 reconstruction error computed over the evaluation set, including layers 12, 16, 20, and 24 for the Llama-3.2-3B-Instruct model. Shaded regions represent the standard deviation ($\pm 1\sigma$). OOD inputs (red) consistently exhibit higher reconstruction errors compared to Routine (blue) and Critical (orange) inputs, particularly in deeper layers.

A.6 Robustness against Noise and Adversarial Injections.

Empirical validation of the bounded gating mechanism ($\tau_{\text{low}} < s_t < \tau_{\text{high}}$) shows that pure noise and gibberish trigger catastrophic reconstruction failures, shifting errors far beyond the critical regime. As an emergent property, adversarial OOD inputs (e.g., prompt and SQL injections) consistently exceed τ_{high} (Figure 8), so ReSuME inherently acts as a structural firewall preventing meaningless noise and malicious directives from corrupting long-term memory.

A.7 MemoryAgentBench Competency Taxonomy

MemoryAgentBench groups tasks into four memory competencies that target distinct failure modes of long-horizon agents: Accurate Retrieval (AR), Test-Time Learning (TTL), Long-Range Understanding (LRU), and Selective Forgetting (SF). In Table 1, we follow the benchmark's original grouping to preserve comparability across memory systems.

Accurate Retrieval (AR). Ability to retrieve the correct stored information in response to a query, including cases where distractors are present and the agent must select the relevant key–value or document snippet.

Test-Time Learning (TTL). Ability to acquire and apply new behaviors during deployment based on interaction feedback or newly observed patterns, without additional parameter updates.

Long-Range Understanding (LRU). Ability to integrate information distributed across long contexts (often $\geq$100k tokens) and answer queries that require global, cross-turn understanding rather than local matching.

Selective Forgetting (SF). Ability to revise, overwrite, or suppress previously stored information under contradiction, aligning memory with updated facts and preventing stale entries from dominating retrieval.

A.8 Non-SAE Memory-Writing Baselines

This appendix specifies the three non-SAE memory-writing strategies used as heuristic baselines in our experiments. These policies are implemented as lightweight controls to quantify the cost–performance trade-off of memory writing under matched budgets. They are not derived from a specific prior method and are included strictly as empirical reference points.

RandomWrite (stochastic baseline). At each eligible turn, the system writes the current turn to memory with a fixed probability p (default $p = 0.10$), independent of content. This yields

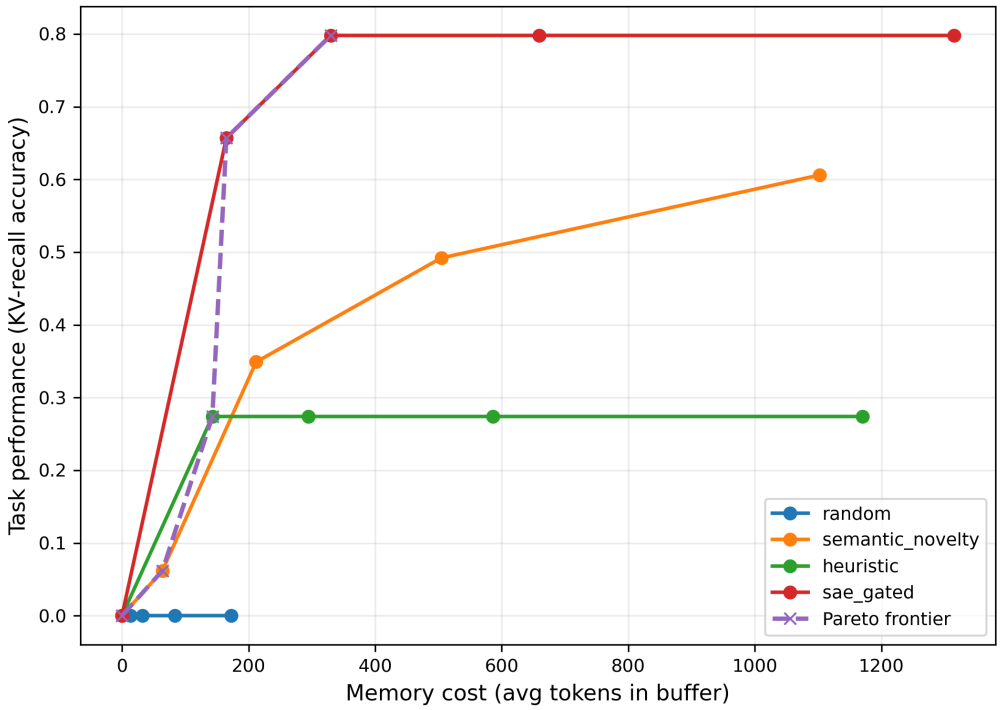

Figure 7: **Memory–performance trade-off for episodic selection policies.** Task performance (KV-recall accuracy) is plotted against average memory cost (tokens stored in the buffer) for multiple write policies: random, heuristic rules, semantic novelty, and SAE-gated surprise. Across budgets, SAE-gating achieves substantially higher recall for the same or lower memory cost, reaching near-saturation performance at moderate budgets, while semantic novelty improves more gradually and heuristics plateau early. The dashed curve denotes the empirical Pareto frontier, indicating that surprise gating dominates competing policies over most operating points.

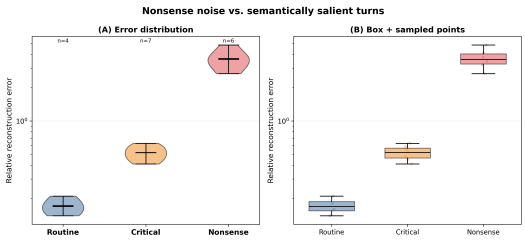

Figure 8: **Robust Gating and Extreme Anomalies.** Density estimation of representational surprise (s_t) extending to adversarial classes. While salient events trigger moderate surprise, adversarial or purely noisy inputs cause extreme compression failure, pushing the error beyond τ_{high}. The shaded region defines the bounded gating zone ($\tau_{\text{low}} < s_t < \tau_{\text{high}}$), ensuring that extreme anomalies are discarded.

an expected-cost baseline that is agnostic to semantic relevance.

HeuristicWrite (keyword triggers). Writing is evaluated only on user turns. A turn is written to memory if it matches a predefined set of lexical triggers (e.g., "`please remember`", "`my code is`", "`my token is`", "`my id is`", "`password`"). In the KV-recall protocol, ground-truth *update* turns are written unconditionally (see Appendix B for the gating procedure and Appendix A.5 for the task definition).

SemanticNoveltyWrite (embedding novelty). Writing is conditioned on semantic redundancy with respect to recently stored items. Let v_t denote the current turn embedding and let $\mathcal{B}_t$ denote a lookback window of the most recent stored embeddings (default $|\mathcal{B}_t| = 50$). We compute cosine similarities $\{\cos(v_t, v_j)\}_{v_j \in \mathcal{B}_t}$ and write the turn if

$$\max_{v_j \in \mathcal{B}_t} \cos(v_t, v_j) < \tau_{\text{nov}}, \tag{8}$$

with default novelty threshold $\tau_{\text{nov}} = 0.75$. As with HeuristicWrite, ground-truth *update* turns in

KV-recall are written unconditionally.

Remark. These baselines are internal heuristic policies formulated for this experimental setup, without attribution claims. They are introduced solely to contextualize the cost–performance behavior of the SAE-gated writing mechanism under identical evaluation conditions.

A.9 Full novelty detection table

Table 3 reports the complete set of novelty-detection and efficiency metrics omitted from the main paper for space. We evaluate representational surprise (SAE reconstruction error with robust normalization) against semantic novelty and heuristic baselines on (i) Routine vs. Critical and (ii) Routine vs. OOD discrimination. In addition to AUC/PRC, we include runtime and memory overhead, as well as write-related statistics (writes per conversation and write rate), to make the performance–cost trade-off explicit.

A.10 ROC Analysis (One-vs-Rest)

Figure 9 presents the Receiver Operating Characteristic (ROC) curves for a One-vs-Rest classification setup. In this analysis, we evaluate the ability of three signals—*Surprise* (ours), *Semantic Similarity* (TF-IDF baseline), and a *Keyword Heuristic*—to distinguish one specific data group (positive class) from the remaining two (negative class).

The results highlight the orthogonal strengths of the baselines:

- **Semantic Similarity** (Green) excels at detecting distributional shifts. It achieves near-perfect separation for *Routine* (AUC 0.97) and *OOD* (AUC 1.00) tasks, as the OOD vocabulary ("galaxy", "quantum") is disjoint from the routine context. However, it fails completely on the *Critical* task (AUC 0.47), indicating that critical messages are semantically close to routine business operations despite their urgency.

- **Keyword Heuristic** (Orange) is highly specialized for the *Critical* class (AUC 0.89), effectively capturing specific triggers (e.g., "urgent", "failure"). Conversely, it performs poorly on *OOD* detection (AUC 0.30), where such keywords are absent.

The *Surprise* signal (Blue) functions as a general-purpose anomaly detector. While it does not outperform the specialized baselines in their respective niches (e.g., specific keyword triggers), it maintains consistent detectability across distribution shifts (OOD AUC 0.71) and routine deviations (Routine AUC 0.75), without relying on predefined vocabulary lists.

A.11 Sensitivity to Threshold and Temporal Smoothing

Figure 10 analyzes the trade-off between the memory write rate (a proxy for storage cost, controlled by threshold T) and the detection performance (F1 score) for identifying non-routine events (Critical and OOD). We evaluate the impact of the exponential moving average (EMA) smoothing factor α on the robustness of the surprise signal.

Two key behaviors emerge from the analysis:

- **Noise Reduction via Smoothing:** The raw surprise signal ($\alpha = 0.00$, blue line) is volatile. It suffers a significant performance drop at moderate write rates ($\alpha = 0.85$, red line), indicating that lowering the threshold without smoothing introduces disproportionate false positives (noise) before capturing true positives.

- **Stability at Scale:** Higher smoothing ($\alpha = 0.85$, red line) yields a monotonic improvement in F1 score. By aggregating the surprise signal over time, the system effectively filters out transient token-level fluctuations. This allows it to achieve the highest peak performance (F1 ≈ 0.78), albeit requiring a higher write budget to fully capture the sustained anomalies.

A.12 Structural Divergence in Residual Space

Figure 14 visualizes the principal components (PCA) of the SAE residual vectors ($\mathbf{r} = \mathbf{x} - \hat{\mathbf{x}}$) across different model layers. This analysis determines whether the "surprise" signal is merely unstructured noise or if it contains identifiable semantic directionality.

The projections reveal a layer-wise evolution in representational distinctiveness:

- **Early-Mid Layers (Layer 12):** The residual distributions for Routine (blue), Critical (orange), and OOD (red) are largely entangled. At this stage, the model's representations—and consequently the SAE's reconstruction errors—lack the specificity to distin-

Table 3: Novelty detection performance of representational surprise, including efficiency metrics.

Method	Routine vs Critical		Routine vs OOD		Efficiency Metrics						
	AUC ↑	PRC ↑	AUC ↑	PRC ↑	Time(ms)↓	Mem(Mb)↓	Write/Conv↓	WriteRate↓	Bank@50↓	Retrieval(ms/turn)↓	P@5
SAE Surprise	**0.998**	**0.999**	**1.000**	**1.000**	**0.5**	**49.5**	**0.289**	**0.31**	**0.3**	**1**	**0.795**
Sem. Simil.	0.885	0.894	0.486	0.213	1.2	61.2	1.0	1.00	1.0	2	0.773
Heuristic	0.954	0.930	0.965	0.782	2.5	0.0	1.0	1.00	1.0	2	0.773
Random	0.495	0.517	0.495	0.125	0	0.0	1.0	1.00	1.0	0	0.763

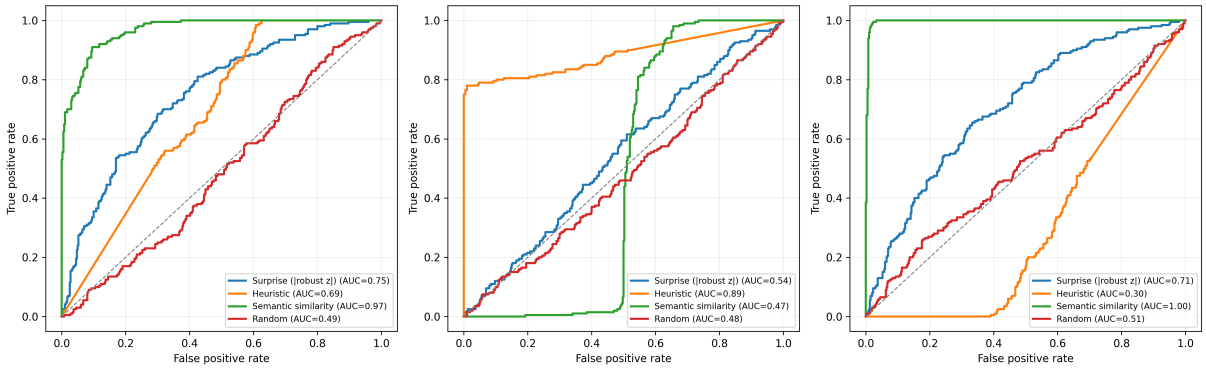

Figure 9: **One-vs-Rest ROC Analysis of Salience Signals.** Performance comparison between the Surprise signal (Robust z-score) and baselines (Heuristic, Semantic Similarity). *Left:* Detection of Routine samples vs. others. *Center:* Detection of Critical samples vs. others. *Right:* Detection of Out-of-Distribution (OOD) samples vs. others. While baselines exhibit high variance—failing in tasks outside their design scope (e.g., Semantic on Critical)—the Surprise signal offers generalized detection capabilities across different types of anomalies.

guish between routine operational anomalies and distinct semantic domains.

- **Deeper Layers (Layer 24):** A clear structural separation emerges. The OOD samples (red) occupy a distinct region of the residual space, orthogonal to the Routine cluster. Notably, while Critical samples (orange) remain closer to the Routine distribution (reflecting their shared business context), they form tight, identifiable sub-clusters.

This structural divergence confirms that high surprise scores are driven by systematic shifts in the activation space, rather than random magnitude fluctuations.

A.13 Surprise Distributions Under Domain Shift

Figure 11 illustrates the distribution of the Surprise z-score across various semantic domains (e.g., Medical, Financial, Tech, Gaming). It is important to note that the samples within these domains are not a separate dataset; they consist of the same underlying *Routine*, *Critical*, and *OOD* categories evaluated in previous sections, now aggregated by thematic context to assess the robustness of the surprise signal against domain shifts.

The box plots reveal that the baseline surprise inherently varies depending on the topical domain:

- **Structured Domains:** Fields such as *Medical* and *Financial* exhibit lower median surprise scores and tighter interquartile ranges. The language and conversational flows in these domains tend to be highly structured, formal, and predictable, aligning closely with standard routine representations.

- **High-Variance Domains:** Conversely, domains like *Tech* and *Gaming* show significantly higher median surprise scores and wider distributions. The prevalence of specialized jargon, colloquialisms, and rapidly evolving terminology naturally induces higher reconstruction errors in the SAE, shifting the overall surprise distribution upward even for non-critical inputs.

These findings highlight that while the SAE's reconstruction error is a strong proxy for salience, different topics possess distinct baseline "surprise" levels. This justifies the use of dynamic or domain-aware calibration (such as the robust z-score relative to a local reference) to prevent high-variance domains from indiscriminately triggering memory writes.

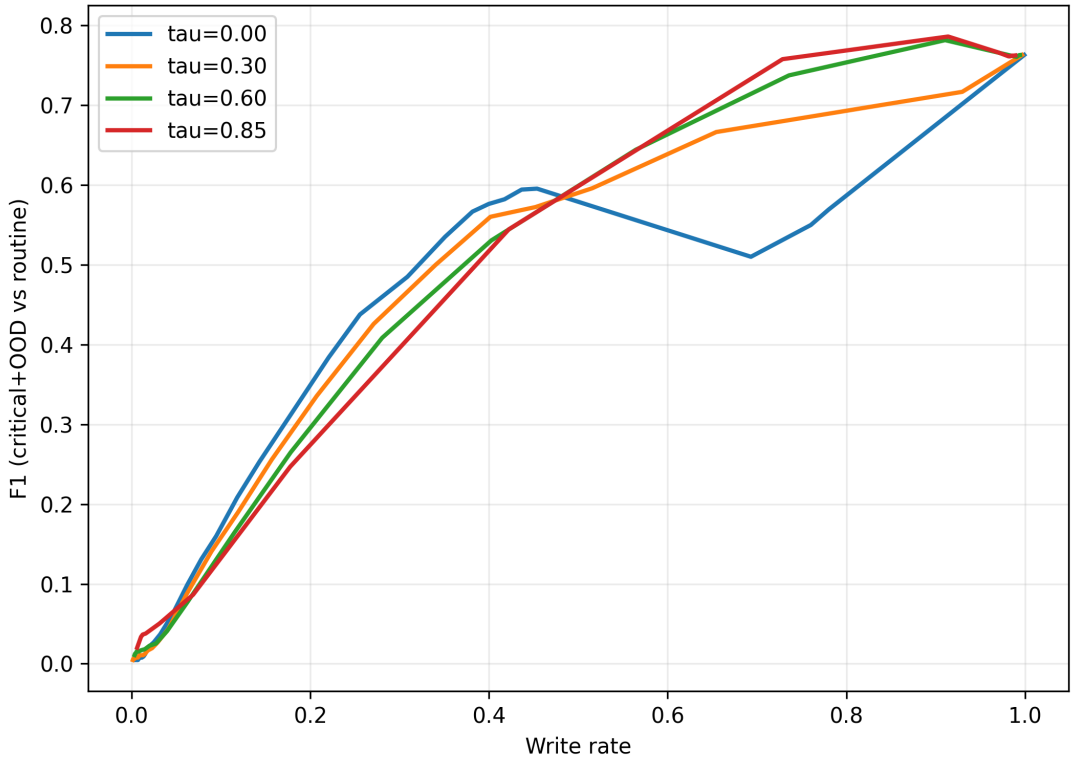

Figure 10: **Impact of Smoothing on Write Sensitivity.** Performance (F1) vs. Cost (Write Rate) for varying smoothing factors α. While the unsmoothed signal ($\alpha = 0$) degrades at intermediate thresholds, strong smoothing ($\alpha = 0.85$) ensures consistent performance gains as the write budget increases, validating the hypothesis that critical events manifest as sustained periods of high surprise.

A.14 Cross-domain robustness table

Table 4 provides the full cross-domain generalization results referenced in Section A.3. For each source domain, we calibrate normalization and a dynamic threshold using Routine statistics (target F1 = 0.80) and evaluate zero-shot on all target domains. We report both AUC for Routine vs. non-routine detection and the resulting write rate, highlighting how calibration transfers across domains and how the memory budget pressure varies with domain shift.

A.15 Orthogonality of SAE Surprise and Semantic Similarity

A natural question that arises is whether the SAE reconstruction error (Surprise) is merely a proxy for traditional semantic distance—i.e., whether high surprise simply equates to a sample being semantically dissimilar to the routine data. Figure 12 addresses this by plotting the Surprise metric ($|\text{robust } z|$) against the semantic distance to the routine centroid for both Routine and Non-routine (Critical + OOD) samples.

The analysis reveals that these two metrics are fundamentally decoupled:

- **Near-Zero Correlation:** The scatter plot exhibits a negligible correlation ($\rho = -0.12$) between semantic distance and SAE surprise. High semantic distance does not inherently guarantee high surprise, and conversely, critical non-routine events frequently occur at low semantic distances.

- **Mechanistic vs. Thematic Anomaly:** Semantic distance (x-axis) effectively captures thematic or topical shifts in the input embedding space. In contrast, SAE surprise (y-axis) captures computational anomalies in the LLM's forward pass. A user might engage in a critical, non-routine behavior (e.g., a security exploit or a toxic outburst) within the exact same business domain or topic as routine conversations.

While standard embeddings might fail to flag these critical events because the "topic" hasn't changed, the SAE detects the shift in the active computational circuitry, resulting in high surprise.

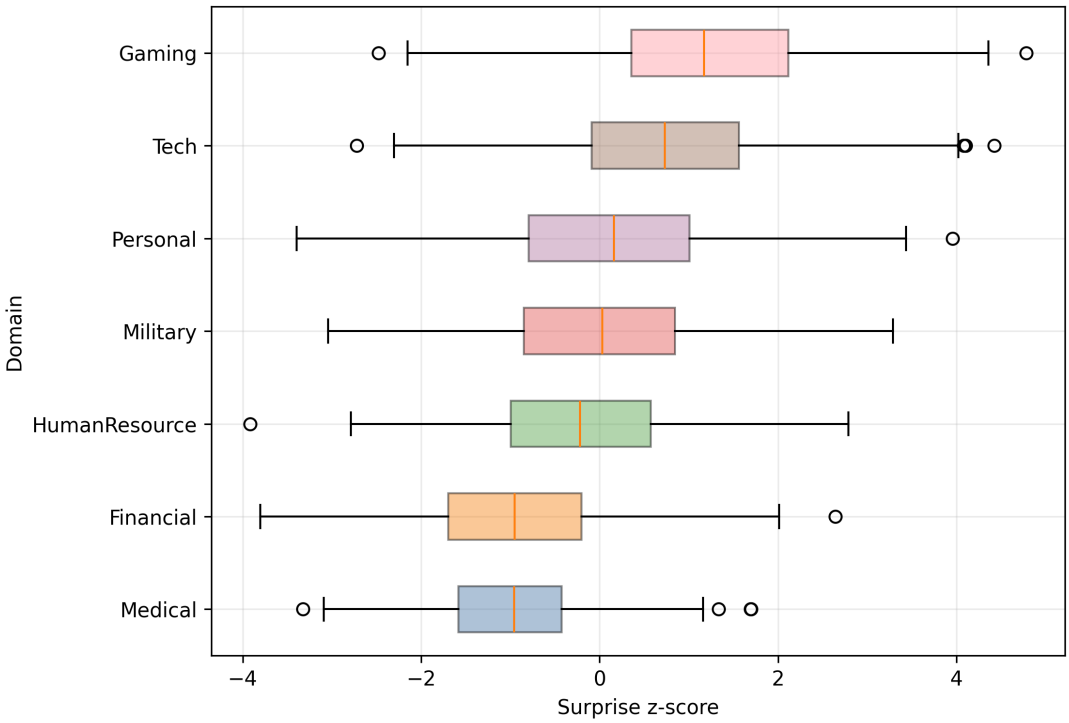

Figure 11: **Surprise distributions under domain shift.** Box plots showing the Surprise z-score for inputs categorized by thematic domains. These data points represent the same Routine, Critical, and OOD interactions analyzed previously, sliced by topic. The shifting medians indicate that domains with specialized or informal jargon (e.g., Gaming, Tech) inherently exhibit a higher baseline surprise compared to more strictly structured domains (e.g., Medical, Financial).

Thus, SAE surprise provides an orthogonal, mechanistically grounded signal that cannot be replaced by simple embedding distance heuristics.

A.16 Comparison with Unsupervised Anomaly-Detection Baselines

This appendix complements the analysis of Figure 4 (main paper) by reporting the full numerical results, including per-regime F1 scores, alongside a discussion of why we explicitly avoid supervised oracles. All baselines are trained *only on routine activations* (zero labels): a vanilla (dense) autoencoder, a linear PCA reconstruction, Isolation Forest, and k-nearest-neighbors (k=5) over routine embeddings.

Table 5 extends the AUC-ROC comparison in Figure 4 with F1 scores per regime. ReSuME achieves F1 $\mathbf{0.970}$ overall, marginally above k-NN (0.961), PCA reconstruction (0.958), and Vanilla AE (0.961). Isolation Forest underperforms substantially (F1 0.841), consistent with its weaker AUC-ROC. As discussed in the main paper, the parity in detection metrics reinforces that the SAE's value lies not in raw detection but in its sparse, learned dictionary, multi-scale ensemble compatibility, and $O(d \cdot h)$ reconstruction cost—structural properties unavailable to k-NN ($O(N \cdot d)$) or dense reconstructors.

Why not a supervised classifier? We explicitly avoid comparing against a supervised classifier (e.g., a fine-tuned MLP head) trained with Routine/-Critical/OOD labels. Such a classifier represents a *supervised oracle upper bound* that violates ReSuME's core premise: the system must function in deployment *without access to importance labels*, since these are precisely what is unavailable in production agents. In our internal validation, an MLP probe trained on the same labeled corpus reached near-ceiling F1 (> 0.99), but this performance is unattainable in deployment because (i) labels do not exist for new domains, (ii) labeling at scale is prohibitively expensive, and (iii) any supervised classifier inherits the bias and coverage of its label distribution and fails to generalize across domain shift. The unsupervised baselines in Table 5 are the appropriate reference class: methods that share ReSuME's deployment constraints. Within this class, ReSuME is competitive while offering interpretable, structured residuals that simpler methods

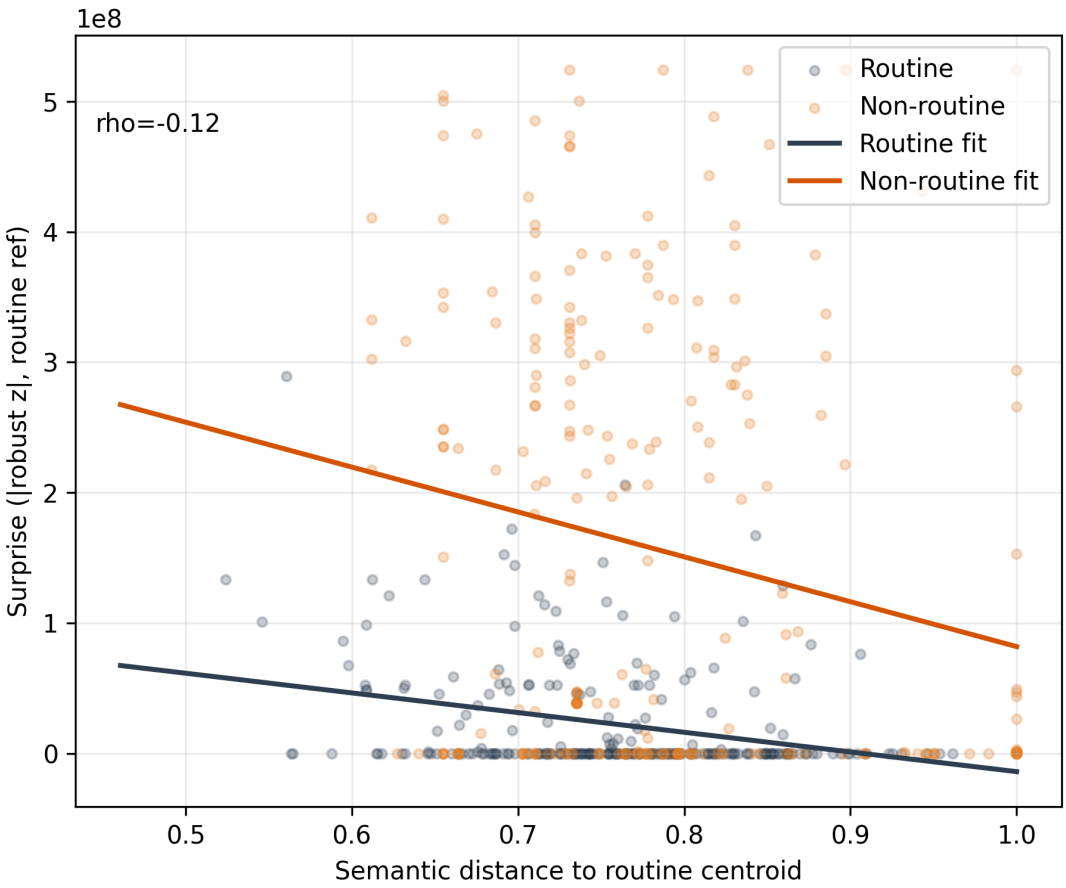

Figure 12: **Surprise vs. Semantic Distance Correlation.** Scatter plot comparing the SAE surprise metric against the semantic distance to the routine centroid. The near-zero correlation ($\rho = -0.12$) demonstrates that SAE surprise is not just "embedding distance in disguise." Non-routine events (orange) consistently exhibit higher surprise across the entire spectrum of semantic distances, proving that the SAE captures mechanistic deviations that traditional thematic similarity misses.

do not provide.

A.17 SAE Architecture and Sparsity Ablation

To validate the architectural choices for our anomaly detection module, we performed an ablation study comparing standard dense Autoencoders (AE) against Sparse Autoencoders (SAE) across different hidden dimensions (256, 512, 1024) and sparsity penalties (L_1 penalty equivalents of 0.0005 and 0.0010). Table 6 summarizes the detection performance (AUC and AUPRC) alongside computational efficiency metrics.

The results highlight three key advantages of enforcing sparsity:

- **The Regularization Effect of Sparsity:** Across all hidden dimensions, SAEs consistently outperform their dense AE counterparts in detecting both Critical and OOD events. Dense AEs, lacking a strict informational bottleneck, are prone to learning the identity function or memorizing noise within the dense LLM activation space. Sparsity forces the network to learn a compact set of fundamental "routine" features. Consequently, when anomalous (OOD/Critical) inputs occur, the sparse basis struggles to reconstruct them, yielding a more discriminative surprise signal.

- **Memory and Computational Efficiency:** While performing better, SAEs are significantly cheaper to deploy. Introducing sparsity reduces the memory footprint by approximately 25% to 34% (e.g., dropping from 102.4 MB to 67.5 MB in the 1024-dimension configuration). Inference times also see slight improvements. This minimal overhead is crucial for online monitoring, ensuring the SAE does not compete for VRAM with the primary

Table 4: Cross-domain robustness of surprise-based memory (synthetic domains).

Train domain	Test domain	AUC	Write rate
Conversational	Conversational	0.838	0.249
Conversational	Legal	0.888	0.428
Conversational	Medical	0.880	0.349
Conversational	Code	0.903	0.511
Conversational	Finance	0.807	0.249
Conversational	Science	0.882	0.307
Legal	Conversational	0.831	0.093
Legal	Legal	0.894	0.254
Legal	Medical	0.881	0.169
Legal	Code	0.908	0.310
Legal	Finance	0.806	0.101
Legal	Science	0.880	0.127
Medical	Conversational	0.832	0.150
Medical	Legal	0.889	0.326
Medical	Medical	0.887	0.257
Medical	Code	0.909	0.408
Medical	Finance	0.806	0.156
Medical	Science	0.878	0.196
Code	Conversational	0.825	0.054
Code	Legal	0.889	0.178
Code	Medical	0.880	0.103
Code	Code	0.912	0.250
Code	Finance	0.806	0.066
Code	Science	0.875	0.090
Finance	Conversational	0.834	0.242
Finance	Legal	0.887	0.429
Finance	Medical	0.882	0.332
Finance	Code	0.908	0.501
Finance	Finance	0.813	0.250
Finance	Science	0.880	0.311
Science	Conversational	0.833	0.183
Science	Legal	0.887	0.369
Science	Medical	0.883	0.291
Science	Code	0.903	0.454
Science	Finance	0.808	0.186
Science	Science	0.886	0.250

LLM.

- **Optimal Scaling Regime:** Increasing the hidden dimension generally improves baseline capacity, but the gains are highly dependent on the sparsity constraint. The configuration with 1024 hidden dimensions and a sparsity of 0.0010 yields the best overall balance for critical event detection (AUC 0.568, AUPRC 0.670) while maintaining a highly efficient profile.

Overall, this ablation confirms that sparsity is not merely an efficiency optimization, but a fundamental requirement for isolating semantic anomalies from standard operational variance in the activation space.

Table 5: **Unsupervised anomaly-detection comparison on routine activations.** All methods are trained on *routine activations only* (zero labels). **Known** = held-out critical turns (in-distribution salience). **Novel** = OOD inputs (e.g., code, foreign domains). Best per column in **bold**.

Method	Known (Critical)		Novel (OOD)		All anomalies	
	AUC	F1	AUC	F1	AUC	F1
ReSuMe	0.962	**0.941**	**0.999**	**0.992**	**0.981**	**0.970**
Vanilla AE	0.957	0.922	0.997	0.985	0.978	0.961
PCA recon.	0.959	0.919	0.998	0.985	0.979	0.958
Isolation Forest	0.713	0.729	0.945	0.889	0.829	0.841
k-NN (k=5)	**0.963**	0.925	0.998	0.985	0.979	0.961

Sensitivity to Dictionary Size: The Goldilocks Zone The reliability of the surprise signal depends critically on the SAE's approximation of the routine manifold. Figure 13 illustrates the conceptual trade-off: overly small dictionaries (e.g., $2\times$) cannot adequately represent the routine manifold, yielding high reconstruction error for both routine and critical states—resulting in high false negative rates after normalization. Conversely, excessively large dictionaries (e.g., $32\times$) may overfit to training patterns, reconstructing even atypical states with low error and producing false positives on benign variations.

Our $4\times$–$8\times$ expansion (h=16384–32768) falls within the "Goldilocks zone" where routine states are reconstructed faithfully while critical deviations still produce elevated residuals. This configuration achieves the practical requirement of $< 10\%$ false positive rate while maintaining $> 80\%$ critical state detection in held-out evaluation. The conceptual illustration in Figure 13 formalizes this design choice, providing intuition for why medium-capacity dictionaries balance sensitivity and specificity in memory-gating applications.

A.18 Effect of Layer Selection on Surprise Detection

Table 7 quantifies the detection performance (AUC and AUPRC) of the SAE anomaly signal across different intermediate layers (12, 16, 20, and 24) of the language model. The results reveal a striking divergence in the optimal depth required to detect different types of non-routine events, corroborating the visual clustering observed in the residual-space projections (Figure 11).

The layer ablation highlights a fundamental trade-off in representation depth:

- **Deep Layers for Macro-Semantic Shifts**

Table 6: SAE architecture and sparsity ablation. Models trained for anomaly detection in critical and OOD data.

Model	Hidden	Sparsity	Critical		OOD		Efficiency		
			AUC	AUPRC	AUC	AUPRC	Train (s)	Inf (ms)	Mem (MB)
AE	256	0.0000	0.557	0.656	0.643	0.225	12.48	0.010	46.8
SAE	256	0.0005	0.561	0.661	**0.651**	**0.235**	12.12	0.009	34.1
SAE	256	0.0010	0.562	0.662	0.648	0.232	12.18	0.009	33.4
AE	512	0.0000	0.561	0.660	0.647	0.226	12.11	0.012	63.7
SAE	512	0.0005	**0.569**	0.667	0.654	0.236	12.06	0.010	44.8
SAE	512	0.0010	0.567	0.666	**0.657**	**0.238**	12.14	0.010	43.6
AE	1024	0.0000	0.563	0.663	0.642	0.224	12.38	0.016	102.4
SAE	1024	0.0005	0.566	0.668	0.652	0.235	12.20	0.013	68.9
SAE	1024	0.0010	**0.568**	**0.670**	**0.656**	**0.239**	12.32	0.013	67.5

(OOD): The ability to detect Out-of-Distribution (OOD) events improves dramatically in deeper layers. Moving from Layer 12 to Layer 24, the OOD AUC skyrockets from 0.260 to 0.878, with a massive five-fold increase in AUPRC ($0.086 \rightarrow 0.500$). Deeper layers in LLMs encode highly abstract, macroscopic semantic concepts. Because OOD data represents a fundamental shift in domain or topic, the SAE easily flags these deep-layer representations as anomalous.

- **Mid-Layers for Behavioral Anomalies (Critical):** Conversely, Critical events are best captured at earlier/intermediate depths (e.g., Layer 12, where AUC and AUPRC peak at 0.331 and 0.418, respectively) and degrade in deeper layers. This indicates that Critical events (such as policy violations or subtle jailbreaks) often masquerade within the same high-level semantic domain as routine tasks. By Layer 24, the LLM has mapped the input to the "correct" business topic, masking the anomaly. The actual deviation occurs during the intermediate reasoning or syntactic processing stages (mid-layers).

These findings suggest that a monolithic, single-layer anomaly detection strategy may be insufficient for comprehensive safety. To build a robust monitor, practitioners must account for this representational trade-off, potentially deploying a dual-layer monitoring approach: probing a deep layer (e.g., 24) to catch gross semantic domain shifts, alongside a mid-layer (e.g., 12 or 16) to intercept adversarial or critical behavioral deviations before they are semantically smoothed out by the network's final layers.

A.19 Extended Cross-Domain Evaluation and Signal Orthogonality

Table 8 reports anomaly detection performance across seven thematic domains. We compare SAE surprise against two baselines: (i) Semantic Similarity (Sem), defined as $1 - \max_j \cos(\mathbf{x}, \mathbf{r}_j)$ where $\mathbf{r}_j$ are routine-turn representations, and (ii) a rules-based Heuristic (Heur) tailored to known critical patterns.

Rather than treating these signals as substitutes, the results suggest complementarity:

- **Domain-stable SAE behavior.** Despite domain shifts (Figure 11), SAE surprise remains stable across domains (AUC 0.57–0.64, AUPRC 0.71–0.78), indicating a domain-agnostic signal.
- **High precision for confident alerts.** Even when AUC is lower than Sem/Heur, SAE AUPRC stays > 0.71 in all domains, suggesting that large surprise values correspond to reliable positives—useful to avoid alert fatigue.
- **Orthogonal signal for in-domain anomalies.** Sem excels at topical shifts (high AUC) and Heur excels at known patterns, but SAE surprise is weakly correlated with semantic distance (Figure 12, $\rho = -0.12$), making it a complementary detector for anomalies that remain semantically in-domain.

A.20 Residual geometry across depth.

Complementing the transfer results, Figure 14 visualizes the structure of SAE residuals by projecting residual vectors into the first two principal components across depth. This provides a geometric perspective on how Routine, Critical, and OOD

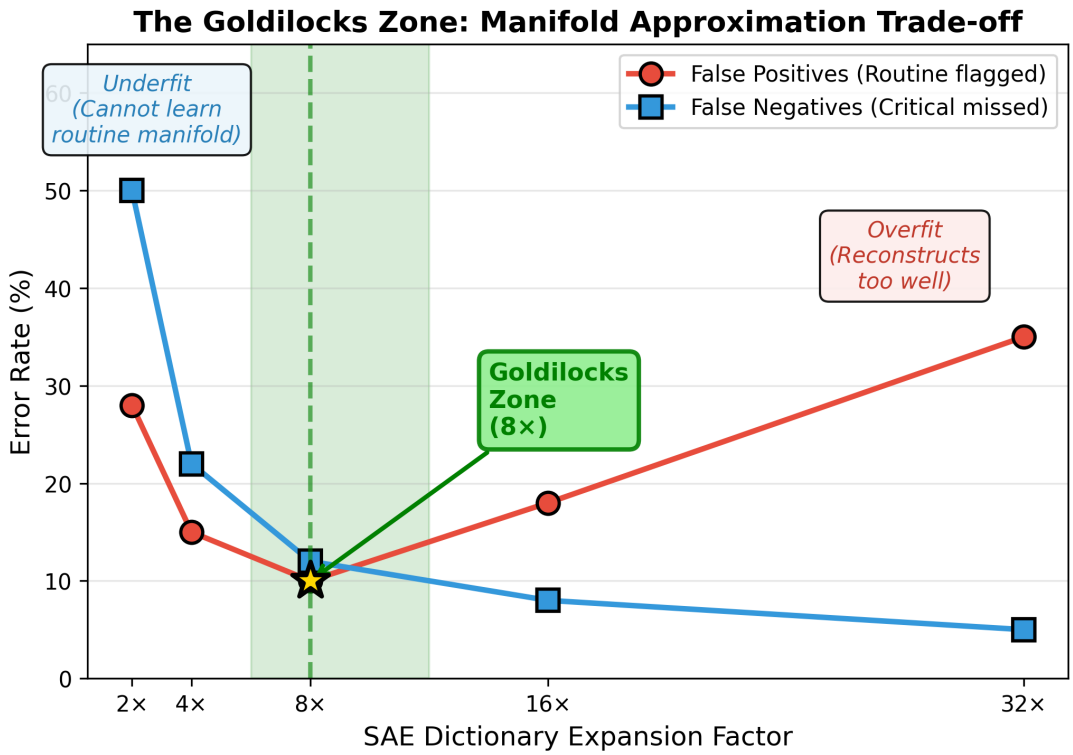

Figure 13: **Conceptual illustration of the Goldilocks zone for SAE manifold approximation.** (Left) Expected trade-off curves showing false positive rate (FPR) and false negative rate (FNR) across dictionary expansion factors. Small dictionaries (underfit) exhibit high FNR because limited capacity prevents distinguishing routine from critical states. Large dictionaries (overfit) reconstruct atypical patterns too well, yielding high FPR. The shaded region marks the Goldilocks zone ($4\times$–$8\times$ expansion) where both error types are minimized. (Right) Schematic reconstruction error distributions demonstrating that optimal separation between routine (blue) and critical (orange) regimes occurs at medium dictionary sizes.

states differ in residual space beyond scalar reconstruction error alone.

A.21 Residual Geometry in SAE Error Space

Figure 15 indicates that SAE residuals encode more than unstructured reconstruction noise. Residuals from Routine turns remain concentrated near the origin, consistent with high-fidelity reconstruction under the learned routine manifold. In contrast, Critical turns occupy a more directionally consistent region, suggesting systematic deviations that remain partially aligned across samples. OOD inputs, while often yielding larger residual magnitude, appear more scattered and less directionally aligned, reflecting heterogeneous domain shifts rather than a single consistent in-domain deviation.

This residual-space geometry motivates a lightweight gating heuristic: combining residual magnitude with directional consistency can prioritize salient in-domain deviations (Critical) while remaining robust to routine variability. In practice, this supports using the residual signal as a compact decision statistic for memory writing, without requiring semantic triggers or external novelty models.

B Signal Processing Details

B.1 Reproducibility Details

Random Seeds and Determinism. To ensure fully deterministic execution, all experiments utilize a fixed random seed of 42. This strict seeding is enforced consistently across all computational libraries, establishing fixed states for both PyTorch manual seeds.

Models and Checkpoints. The core experiments employ the Qwen3-4B, Qwen3-8B, Qwen3.5-0.8B, Qwen3.5-2B, Qwen3.5-4B, Qwen3.5-9B, Llama-3.2-1B, Llama-3.2-3B, DeepSeek-R1-Distill-Llama-8B, and DeepSeek-R1-Distill-Qwen-14B models. Rather than relying on a pre-trained Sparse Autoencoder (SAE), we trained a custom SAE from scratch, specifically tailored for our novelty detection pipeline.

Table 7: Effect of layer selection on surprise detection.

Layer	AUC (Crit)	AUPRC (Crit)	AUC (OOD)	AUPRC (OOD)
12	0.331	0.418	0.260	0.086
16	0.307	0.412	0.457	0.124
20	0.315	0.403	0.859	0.334
24	0.245	0.375	0.878	0.500

Table 8: Extended cross-domain evaluation. Performance of SAE surprise compared to Semantic Similarity (Sem) and Heuristic (Heur) baselines across thematic domains. Sem captures semantic distance to the routine set, Heur encodes known pattern rules, and SAE provides a stable, orthogonal signal with consistently high precision (AUPRC > 0.71).

Domain	SAE AUC	SAE AUPRC	Sem AUC	Sem AUPRC	Heur AUC	Heur AUPRC
Medical	0.612	0.784	0.942	0.961	0.903	0.912
HumanResource	0.588	0.731	0.918	0.944	0.889	0.901
Financial	0.621	0.752	0.905	0.931	0.934	0.946
Military	0.606	0.741	0.892	0.915	0.958	0.962
Personal	0.574	0.716	0.928	0.952	0.872	0.885
Tech	0.639	0.769	0.914	0.936	0.967	0.971
Gaming	0.592	0.725	0.887	0.905	0.941	0.948

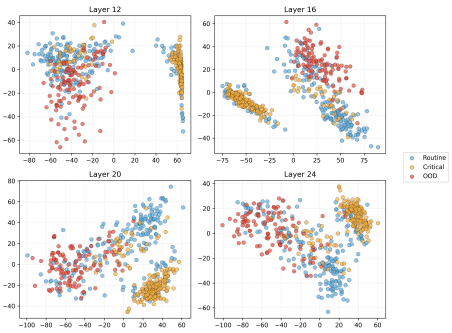

Figure 14: **PCA projections of SAE residuals across layers.** Each panel shows the first two principal components of residual vectors $r = x - \hat{x}$ for Routine, Critical, and OOD inputs. Overlap is stronger in earlier layers, while deeper layers exhibit clearer geometric separation (notably for OOD), indicating structured directionality in residual space rather than pure magnitude effects.

Sparse Autoencoder (SAE) Training. Our custom SAE is trained for 300 epochs with a batch size of 32. For the representation sparsification procedure, the architecture is configured with a context length of 1024, a batch size of 8, 2 gradient accumulation steps, and a sparsity penalty of $K = 128$ over 35 layers. The comprehensive dataset used for SAE training consists of $108,000$ total samples. To maintain class balance, this corpus is strictly distributed as $6,000$ routine, $6,000$ critical, and $6,000$ Out-of-Distribution (OOD) samples for each of the defined classes.

Datasets and Generation Parameters. For the conversational novelty detection evaluation, the supplementary manually-labeled dataset comprises $1,336$ conversations, distributed as 817 critical dialogues (yielding $3,264$ individual turns) and 519 OOD dialogues (yielding $2,076$ individual turns). To augment the dataset with synthetic routine, critical, and OOD samples, we utilized vLLM serving the `Qwen/Qwen3-30B-A3B` model and its Instruct variant. The generation process relied on predefined rule-based prompts, configured with a sampling temperature of 0.7, progressively generating dialogues to reach exactly $6,000$ samples for each category.

Hardware Specifications. All computational experiments—including Large Language Model (LLM) inference, Sparse Autoencoder training, and synthetic data generation—were executed on a single NVIDIA B200 GPU.

B.2 ReSuME online procedure (pseudocode)

[t] [1] Pretrained transformer f_ϕ; SAE (g, d); capacity M; write threshold τ; use threshold γ; window size W; warm-up steps T_{warm}; retrieval top-k; EMA smoothing $\alpha \in [0, 1]$. Initialize memory $\mathcal{M} \leftarrow \emptyset$; error history $\mathcal{H} \leftarrow []$; smoothed score $s_0 \leftarrow 0$. $t = 1, 2, \ldots$ Compute representation

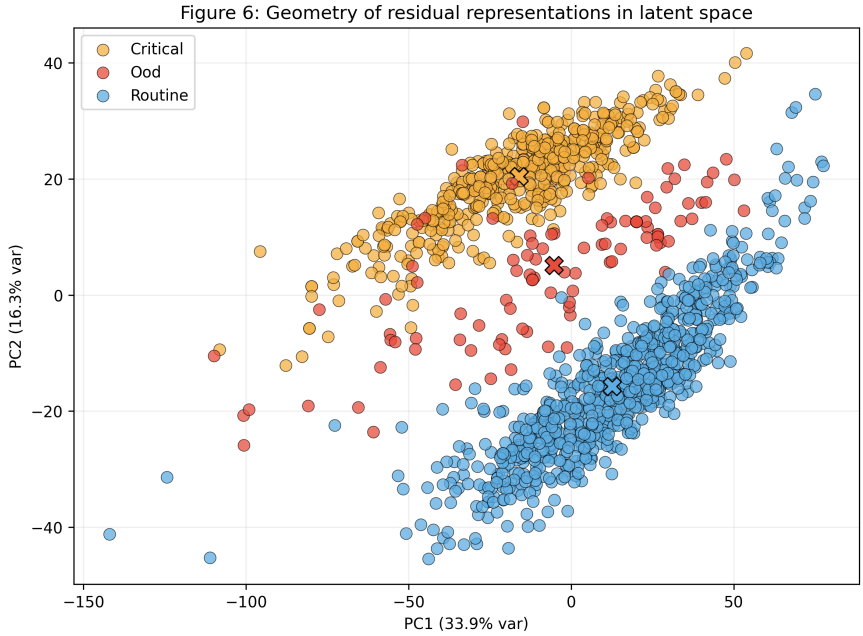

Figure 15: **Residual geometry.** PCA of residuals $r_t = h_t - \hat{h}_t$ exhibits structured regime separation: Routine residuals concentrate near the origin, Critical turns form directionally coherent clusters, and OOD inputs are typically higher-magnitude but more diffuse.

$x_t \leftarrow \mathrm{Pool}(\{h_{t,i}^{(\ell)}\}_{i=1}^{T_t})$ $\tilde{x}_t \leftarrow x_t/\|x_t\|_2$ normalized key for retrieval $z_t \leftarrow g(x_t)$; $\hat{x}_t \leftarrow d(z_t)$ $e_t \leftarrow \|x_t - \hat{x}_t\|_2$ Update rolling statistics (μ_t, σ_t) from $\mathcal{H}$ using the last W errors $u_t \leftarrow \mathrm{clip}\big((e_t - \mu_t)/(\sigma_t + \varepsilon), -c, c\big)$ instantaneous surprise score $s_t \leftarrow \alpha\, u_t + (1-\alpha)\, s_{t-1}$ EMA-smoothed surprise Append e_t to $\mathcal{H}$ $t > T_{\mathrm{warm}}$ **and** $\nVdash_{\mathrm{user}}(t) = 1$ **and** $s_t > \tau$ Insert $(\tilde{x}_t, s_t, t, \mathrm{meta}_t, \mathrm{text}_t)$ into $\mathcal{M}$ $|\mathcal{M}| > M$ Evict one item from $\mathcal{M}$, e.g., the item with lowest surprise $|\mathcal{M}| > 0$ Retrieve top-k items by cosine similarity $\langle \tilde{x}_t, \tilde{x}_i \rangle$ over $m_i \in \mathcal{M}$ $\max_i \langle \tilde{x}_t, \tilde{x}_i \rangle \geq \gamma$ Expose retrieved items to the agent, e.g., as additional context

This appendix details the signal processing pipeline used to transform the raw reconstruction error e_t into the stable decision metric s_t used in Section 3.1. The pipeline consists of robust normalization followed by temporal smoothing to mitigate non-stationarity and transient noise.

B.3 Robust Normalization (MAD)

The raw reconstruction error distribution $E = \{e_t\}$ is often non-Gaussian and prone to outliers (e.g., proper nouns or rare tokens). Standard Z-score normalization (using mean and standard deviation) is therefore unstable, as large errors skew the statistics.

Instead, we employ the **Median Absolute Deviation (MAD)**, a robust measure of variability. For a sliding window of recent errors $W_t = \{e_{t-k}, \ldots, e_t\}$, we estimate the robust standard deviation $\hat{\sigma}_t$ as:

$$\hat{\sigma}_t = c_{\mathrm{MAD}} \cdot \mathrm{median}\left(|W_t - \mathrm{median}(W_t)|\right) \quad (9)$$

where $c_{\mathrm{MAD}} \approx 1.4826$ is the scaling factor that makes MAD consistent with the standard deviation for Gaussian distributions.

The robust normalized score u_t is then computed as:

$$u_t = \frac{e_t - \mathrm{median}(W_t)}{\hat{\sigma}_t + \epsilon} \quad (10)$$

where ϵ is a small constant for numerical stability.

B.4 Temporal Smoothing (EMA)

To prevent rapid switching of the memory gate due to token-level jitter, we apply an Exponential Moving Average (EMA) to the normalized score. This acts as a low-pass filter, emphasizing sustained deviations over instantaneous spikes.

The final surprise score s_t is updated recursively:

$$s_t = \alpha u_t + (1-\alpha) s_{t-1}. \quad (11)$$

where $\alpha \in [0, 1]$ is the smoothing factor.

B.5 Hyperparameters

For reproducibility, we list the hyperparameters used in our main experiments in Table 9.

Table 9: Hyperparameters for the Surprise Signal Processing.

Parameter	**Symbol**	**Value**
Window Size	$\|W\|$	100 turns
Smoothing Factor	α	0.1
Sensitivity Threshold	τ	2.0
Min. Variance Floor	ϵ	10^{-6}