EPISODIC MEMORY FROM COMPRESSION BOUNDARIES IN LATENT REPRESENTATION SPACE

David O. C. Ferreira, Priscila R. M. F Ribeiro, Emanuel B. Passinato,
Diogo F. C. Silva & Arlindo R. Galvão Filho
Advanced Knowledge Center for Immersive Technologies (AKCIT)
Federal University of Goiás
{oneil, priscila.maia, diogo_fernandes}@discente.ufg.br
emanuel.passinato@egresso.ufg.br arlindogalvao@ufg.br

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 compared to heuristic and similarity-based baselines under fixed budgets. 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 compres-

sion 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:

- **A geometric formulation of episodic memory.** We formalize memory writing as compression failure relative to a learned latent manifold. Reconstruction residuals from a Sparse Autoencoder trained on routine activations provide an intrinsic, unsupervised signal for episodic persistence.
- **Geometry-aware representational surprise.** Effective memory gating requires covariance-aware residual modeling. Representational deviation is anisotropic in activation space, and geometry-aware normalization substantially improves separation of critical and out-of-distribution states.
- **Layer-dependent residual stratification.** Deviation patterns are not uniform across depth: intermediate layers preferentially capture state-changing events, while deeper layers more strongly separate distributional anomalies, revealing multi-scale organization of latent representations.
- **Improved performance under memory constraints.** Across controlled KV-recall settings and MemoryAgentBench tasks, surprise-gated memory achieves a superior performance–memory trade-off compared to heuristic and similarity-based baselines.
- **Functional gains under resource constraints.** We show that surprise-gated memory achieves a superior performance–memory trade-off across synthetic and benchmark tasks, improving retrieval, test-time learning, and selective forgetting under fixed memory budgets.

Empirical evaluation reveals three core findings. First, representational surprise exhibits structured separation across conversational regimes, increasing monotonically from routine dialogue to critical state transitions and to out-of-distribution inputs. Second, under fixed memory budgets, surprise-gated writing achieves a superior performance–cost trade-off compared to heuristic and similarity-based policies. Third, robust normalization enables cross-domain calibration, indicating that the signal captures structural deviation rather than surface-level lexical novelty.

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

Memory Writing in LLM Agents. Early LLM agents equipped 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) demonstrate that persistent memory streams enable long-horizon behavioral coherence, but their write policy is externally defined and not grounded in the model's internal computation. Subsequent surveys formalize architectural patterns for memory-augmented agents (Zhang et al., 2025), yet the decision of what to store typically remains based on surface-level signals, similarity thresholds, or predefined triggers.

More recent approaches introduce *surprise* as a mechanism for segmenting experience into episodic units. Output-level formulations operationalize surprise using token-level uncertainty or prediction error. For example, EM-LLM (Fountas et al., 2025) leverages entropy and prediction loss to detect episode boundaries. While effective for segmentation, such signals remain tied to output uncer-

tainty and are sensitive to lexical variability. Crucially, they conflate linguistic unpredictability with semantic salience, limiting their reliability as a general-purpose memory writing policy.

In contrast, our approach defines surprise intrinsically in the model's latent representation space, decoupling memory formation from token-level uncertainty and external heuristics.

Sparse Autoencoders and Latent Structure. Sparse Autoencoders (SAEs) have recently emerged as powerful 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). Scaling studies demonstrate that sparse decompositions reveal stable, reusable feature hierarchies across prompts and domains (Cunningham et al., 2023; Gao et al., 2024). Advances such as top-k and gated SAEs further improve reconstruction fidelity under extreme sparsity constraints (Deng et al., 2025).

Although primarily developed for interpretability, SAEs implicitly define a manifold of routine activations. Inputs aligned with dominant latent features are reconstructed efficiently, whereas deviations induce structured residuals. Prior work suggests that these residuals contain information not captured by the learned basis, enabling anomaly detection in high-dimensional representations (Kissane et al., 2024; Sakurada & Yairi, 2014).

ReSuME builds on this observation but repurposes SAE reconstruction error as a functional control signal rather than an analysis artifact. Instead of using SAEs to interpret representations post hoc, we use reconstruction failure online as a write gate for episodic memory.

Error-Driven Prioritization and Intrinsic Signals. Our formulation also relates to error-driven prioritization mechanisms, where deviation signals guide selective processing or resource allocation (Makelov et al., 2024). However, prior work focuses on interpretability evaluation or representation analysis rather than memory consolidation in agents.

We treat compression failure in latent space as an unsupervised proxy for semantic importance. This framing shifts memory writing from externally imposed heuristics to an intrinsic geometric property of the model's representational dynamics. Importantly, our signal is orthogonal to embedding similarity and output entropy, capturing computational deviations that remain semantically in-domain.

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. Although these representations encode rich semantic structure, they remain entangled and lack explicit control mechanisms for downstream decisions such as memory consolidation.

Prior work has analyzed LLM activations using probing classifiers (Liu et al., 2024) and attention-based attribution, uncovering partial semantic organization. More recent advances in sparse decomposition demonstrate that residual stream representations admit low-dimensional structure. In particular, monosemantic neurons and residual-stream Sparse Autoencoders (SAEs) (Gao et al., 2024) reveal that LLM activations can be approximated using a sparse set of interpretable features. These features correspond to modular directions that recur across prompts and layers, including entities, syntactic roles, and high-level semantic attributes.

Together, these findings suggest that LLM activations concentrate near a structured manifold of routine representations. This manifold reflects the model's learned expectations over typical linguistic and conversational patterns. States that align with this manifold admit sparse, low-error reconstructions. In contrast, states that deviate from it, due to novelty, semantic incongruity, or domain 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. This signal forms the basis of our memory gating mechanism.

3.1 Sparse Autoencoders as Manifold Models

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 a reconstruction objective under a sparsity constraint:

$$\mathcal{L}_{\text{SAE}}(x) = \|x - d(g(x))\|_2^2 + \lambda \|g(x)\|_1. \tag{1}$$

The ℓ_1 penalty enforces a sparse latent code, encouraging the model to represent routine activations using a compact dictionary of features. Under this constraint, the SAE approximates the dominant subspace of in-distribution conversational states. Inputs that remain within this routine manifold admit accurate reconstruction. Inputs that activate combinations of features underrepresented in the learned dictionary incur larger residuals.

Thus, the SAE defines not only a reconstruction model but an implicit boundary between routine and structurally deviant states.

3.2 Representational Surprise

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:

$$e_t = \|x_t - \hat{x}_t\|_2. \tag{2}$$

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:

$$s_t > \tau, \tag{3}$$

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 1 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

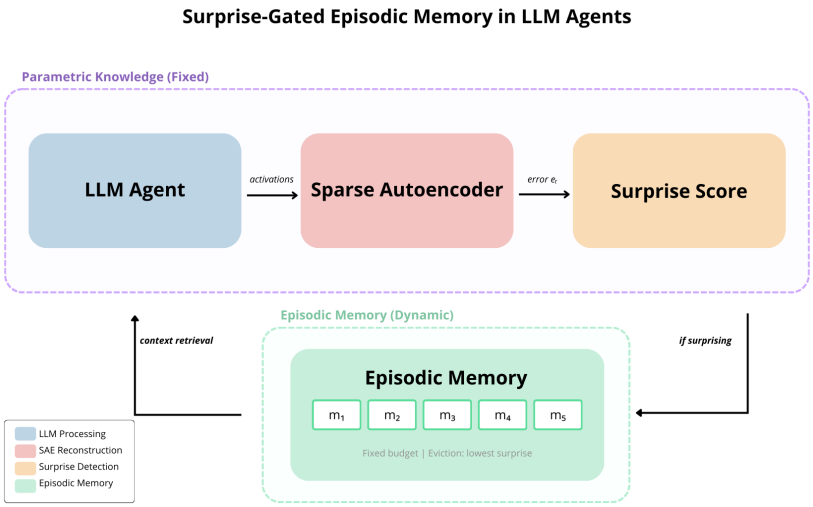

Figure 1: Hidden states from a frozen LLM are projected through a Sparse Autoencoder trained on routine activations, yielding a geometry-aware reconstruction residual. This calibrated surprise signal governs episodic memory writes under a fixed budget, while retrieval remains similarity-based and independent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

(1) Representation extractor. At turn t, the transformer produces 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*, excluding system and assistant tokens. We aggregate these states into a single turn-level representation via masked mean pooling:

$$x_t \;=\; \mathrm{Pool}\Big(\{h_{t,i}^{(\ell)}\}_{i=1}^{T_t}\Big) \;=\; \frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t} h_{t,i}^{(\ell)}. \tag{4}$$

This pooling step produces a compact state vector that summarizes the latent configuration associated with the user's contribution. Alternative pooling strategies, including last-token pooling or mean pooling over the final n user tokens, are evaluated as ablations.

(2) Typicality model (SAE). The pooled representation x_t is projected onto the learned routine manifold using a Sparse Autoencoder:

$$z_t = g(x_t), \qquad \hat{x}_t = d(z_t). \tag{5}$$

The SAE is trained on in-distribution conversational states to minimize:

$$\mathcal{L}_{\mathrm{SAE}}(x) = \|x - d(g(x))\|_2^2 + \lambda\|g(x)\|_1. \tag{6}$$

The sparsity constraint enforces a compact latent code, forcing routine states to be expressed through a limited set of dominant features. As a result, reconstruction residuals reflect structured deviations from this routine manifold rather than arbitrary noise.

Competency	Task/Dataset	Avg. Ctx (K)	No-Memory	RAG	Ours (SAE-gated)	Δ
AR	SH-Doc QA	448	51.3	79.3	**83.0**	+3.7
	MH-Doc QA	183	45.1	68.1	**71.2**	+3.1
	LongMemEval (S*)	355	42.8	61.9	**68.3**	+6.4
	EventQA	534	36.5	71.6	**75.4**	+3.8
TTL	BANKING-77	103	74.0	89.1	**91.0**	+1.9
	CLINC150	98	85.0	88.6	**93.0**	+4.4
	NLU	105	79.0	85.4	**88.0**	+2.6
	TREC Coarse	112	57.0	91.8	**94.0**	+2.2
	TREC Fine	115	76.0	83.5	**86.0**	+2.5
	Movie Rec	1440	12.8	19.0	**25.0**	+6.0
LRU	Bench-Sum	172	41.3	**59.2**	58.5	-0.7
	Detective-QA	200	50.4	59.5	**55.6**	-3.9
SF	FactConsol-SH	262	19.7	37.6	**43.8**	+6.2
	FactConsol-MH	262	1.9	2.7	**3.5**	+0.8

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). Each task is reported using its official metric (higher is better). We compare a no-memory baseline, a retrieval-augmented (RAG) baseline, and ReSuME (SAE-gated). Δ denotes the absolute improvement of ReSuME over RAG.

(3) Surprise estimator. ReSuME transforms reconstruction fidelity into a calibrated surprise signal. The raw reconstruction magnitude is computed as $e_t = \|x_t - \hat{x}_t\|_2$. Because conversational dynamics are non-stationary, we normalize this signal using a robust sliding-window statistic:

$$u_t = \mathrm{clip}\left(\frac{e_t - \mu_t}{\sigma_t + \varepsilon}, -c, c\right), \tag{7}$$

where μ_t and σ_t are median and MAD-based estimates computed over recent residuals.

To emphasize sustained deviations over transient spikes, we apply exponential smoothing:

$$s_t = \alpha u_t + (1 - \alpha) s_{t-1}. \tag{8}$$

The smoothed score s_t constitutes the operational surprise signal. A memory write is triggered when $s_t > \tau$.

(4) Episodic memory store. When the write gate activates, the system stores a structured memory tuple:

$$m_i = (\tilde{x}_i,\ s_i,\ t_i,\ \mathrm{meta}_i,\ \mathrm{text}_i), \tag{9}$$

where $\tilde{x}_i = x_i / \|x_i\|_2$ is a normalized key used for similarity-based retrieval. Storing both the latent key and associated metadata allows efficient matching while preserving contextual content.

(5) Online control flow (write and use gates). ReSuME evaluates the write gate exclusively on user-authored turns. Let $\mathbb{1}_{\text{user}}(t)$ indicate whether turn t originates from the user. The rolling normalization baseline and surprise statistics are updated only when $\mathbb{1}_{\text{user}}(t) = 1$. This design prevents assistant-generated tokens from distorting the routine baseline or populating memory directly.

Write rule: store the current turn if $\mathbb{1}_{\text{user}}(t) = 1$ and $s_t > \tau$.

Use rule: at each turn, retrieve the top-k stored items ranked by cosine similarity with $\tilde{x}_t$. Retrieved memories are exposed to the agent only when similarity exceeds a threshold γ.

Algorithm 1 summarizes the complete online procedure. By separating surprise-driven consolidation from similarity-driven retrieval, ReSuME ensures that memory formation reflects structural deviation, while memory use reflects contextual relevance.

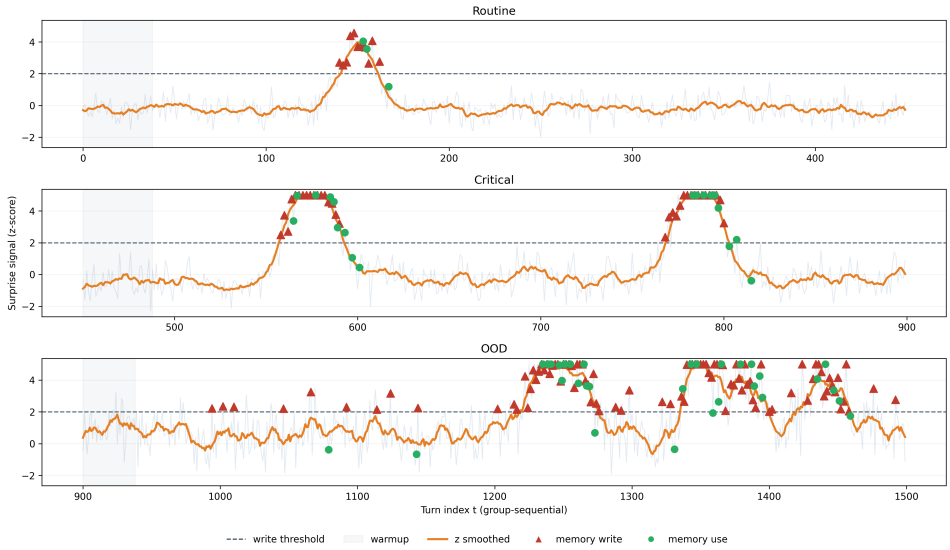

Figure 2: Temporal dynamics of the surprise signal s_t (orange) across Routine, Critical, and OOD regimes. The dashed line denotes the write threshold τ. Critical segments induce sustained supra-threshold deviations that trigger writes (red) and subsequent retrievals (green), while OOD segments exhibit higher variance. The shaded region marks the warm-up period.

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.1). Collectively, these experiments serve to demonstrate that representational surprise is both geometrically meaningful and functionally advantageous.

6 Experiments

6.1 Experimental Setup

ReSuME combines a frozen Large Language Model (LLM) with a layer-specific Sparse Autoencoder (SAE) trained to model routine conversational activations. The SAE reconstructs frozen LLM states using a $4\times$ expansion (d=4096 $\rightarrow$ h=16384) with an ℓ_1 sparsity penalty (λ=0.1). Training is performed on flattened multi-turn conversations using Adam, with Top-k gating (k=128) to stabilize sparse activations.

Turn representations x_t are extracted from layer ℓ via masked mean pooling over user tokens (Eq. 4). Reconstruction error $e_t = \|x_t - \hat{x}_t\|_2$ is normalized using robust rolling statistics (Appendix B.3), and memory writes are triggered when the smoothed score s_t exceeds threshold τ. Rolling statistics are updated only on user turns. Alternative pooling strategies are evaluated as ablations.

All implementation details required to reproduce our results, including the full online algorithm, signal processing pipeline, and hyperparameter configurations, are provided in Appendix B.1. Code and model checkpoints will be released upon publication.

6.2 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).

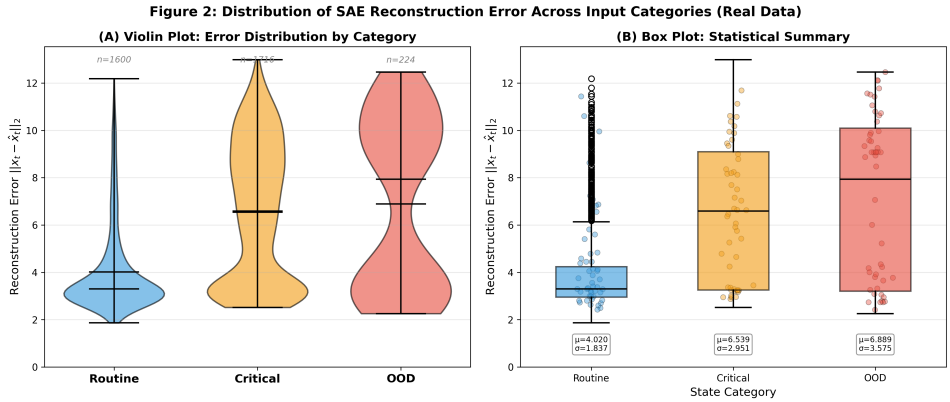

Figure 3: **Distribution of SAE reconstruction error across dialogue regimes (real data). (A)** Violin plots of $\|x_t - \hat{x}_t\|_2$ for Routine, Critical, and OOD turns. Routine interactions concentrate at lower reconstruction error, while Critical and OOD states exhibit upward-shifted and heavier-tailed 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.

Table 2: **Comparison of residual-based surprise metrics.** AUC and AUPRC (higher is better) for discriminating Routine vs. Critical (Crit) and Routine vs. OOD turns using ℓ_2 residuals, cosine deviation, Mahalanobis distance, and covariance-whitened residuals.

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

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

Figure 3 confirms this behavior. Reconstruction error separates all three regimes (one-way ANOVA, $F(2, 4797) =$ [valor], $p < 0.001$). 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 ($\mu = 13.5$). This monotonic progression indicates that representational surprise captures structured latent deviations rather than surface-level variation.

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.

Together, these results indicate that representational surprise is structured both geometrically and hierarchically, reinforcing its role as an intrinsic memory signal rather than a surface-level anomaly measure.

6.3 Memory Efficiency and Task Performance

We evaluate the trade-off between memory usage and downstream recall. Memory cost is measured as average tokens stored, and task performance is KV-recall accuracy: the fraction of evaluation turns where the correct key–value pair appears among retrieved memory items.

The KV-recall protocol injects update statements of the form "my {slot} = {gold}" into dialogue, followed later by evaluation queries. Memory writing is controlled either by SAE-gated surprise or heuristic baselines (Appendix A.5).

Memory cost	Random	Semantic novelty	Heuristic	SAE-gated	Pareto frontier
0	0.00	0.00	0.00	0.00	0.00
100	0.00	0.33	0.26	**0.65**	**0.65**
200	0.00	0.34	0.27	0.66	0.66
400	0.00	0.49	0.27	0.79	0.79
600	0.00	0.52	0.27	0.80	0.80
800	0.00	0.56	0.27	0.80	0.80
1000	0.00	0.60	0.27	0.80	0.80
1200	0.00	0.80	0.27	0.80	0.80

Table 3: **KV-recall accuracy versus memory cost.** Accuracy is reported against average stored tokens for Random, Semantic Novelty, Heuristic, and SAE-gated (ReSuME) writing strategies. KV-recall measures whether the correct key–value update is retrieved at evaluation time (higher is better). The Pareto frontier denotes the best accuracy achieved at each budget.

SAE-gated writing consistently dominates the Pareto frontier, reaching 80% recall with only 600 stored tokens. Unlike heuristic policies, it selectively captures latent deviations corresponding to state-changing updates.

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). ReSuME improves AR, TTL, and SF, which depend on preserving discrete distribution-shifting events. Performance slightly decreases on LRU, which requires retaining globally coherent but individually low-surprise content. This trade-off reflects the distinction between surprise-driven consolidation and retrieval-heavy strategies.

6.4 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 2 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.

We additionally evaluate cross-domain transfer of surprise calibration under zero-shot normalization. Across synthetic domains, the normalized signal preserves separability between routine and non-routine states, although write pressure varies with domain shift (details in Appendix A.1).

Further analyses of residual-space geometry, including covariance-aware metrics and layer sensitivity, are provided in Appendix A.17.

7 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. 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 mechanisms in neural agents.

ACKNOWLEDGMENTS

This work has been fully funded by the project Research and Development of Gênese Digital: Scaling Interactive and Culturally Adapted Digital Humans with Generative AI, supported by the Advanced Knowledge Center in Immersive Technologies (AKCIT), with financial resources from the PPI IoT/Manufatura 4.0 / PPI HardwareBR of the MCTI, grant number 057/2023, signed with EMBRAPII.

REFERENCES

Trent Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Catherine Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

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

Boyi Deng, Yu Wan, Baosong Yang, Yidan Zhang, and Fuli Feng. Unveiling language-specific features in large language models via sparse autoencoders. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4563–4608, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.229. URL https://aclanthology.org/2025.acl-long.229/.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

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

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

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

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

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, 2024.

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

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

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.

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

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

A Appendix

A.1 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 8 and Table 5.

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.2 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 4 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.3 Performance–Memory Trade-off

This appendix complements Section 6.3 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 5 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.

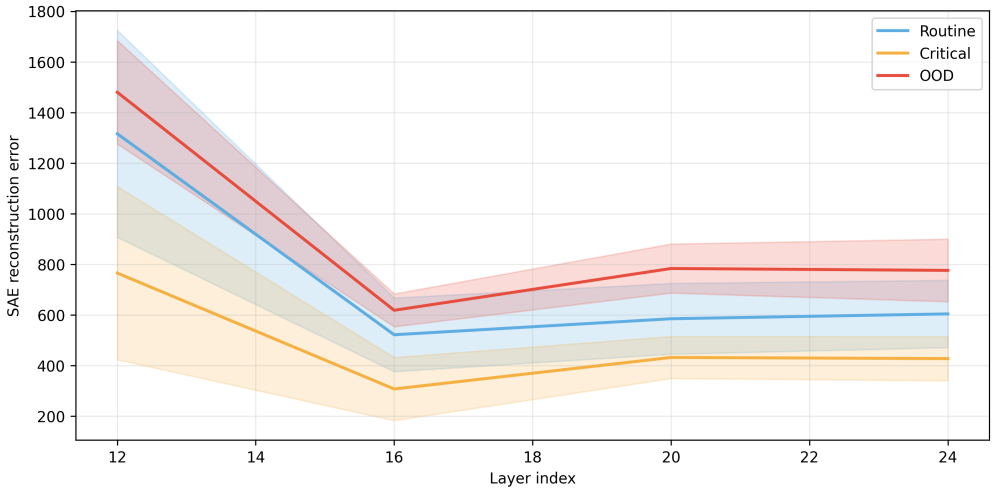

Figure 4: **Layer-wise distribution of SAE reconstruction error.** The plot displays the average L_2 reconstruction error computed over the evaluation set, including layers 12, 16, 20, and 24 for the Llama-3.2-3B-Instruct model. Shaded regions represent the standard deviation ($\pm 1\sigma$). OOD inputs (red) consistently exhibit higher reconstruction errors compared to Routine (blue) and Critical (orange) inputs, particularly in deeper layers.

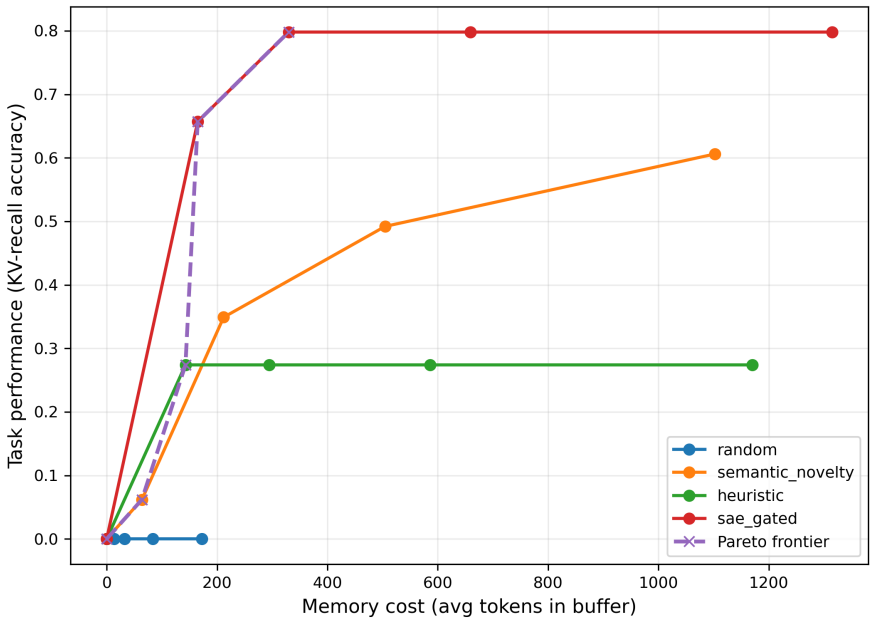

Figure 5: **Memory–performance trade-off for episodic selection policies.** Task performance (KV-recall accuracy) is plotted against average memory cost (tokens stored in the buffer) for multiple write policies: random, heuristic rules, semantic novelty, and SAE-gated surprise. Across budgets, SAE-gating 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.

A.4 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.5 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 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 Section 6.3 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

$$\max_{v_j \in \mathcal{B}_t} \cos(v_t, v_j) < \tau_{\text{nov}}, \tag{10}$$

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.6 Full novelty detection table

Table 4 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.7 ROC Analysis (One-vs-Rest)

Figure 6 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*

Table 4: 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

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.

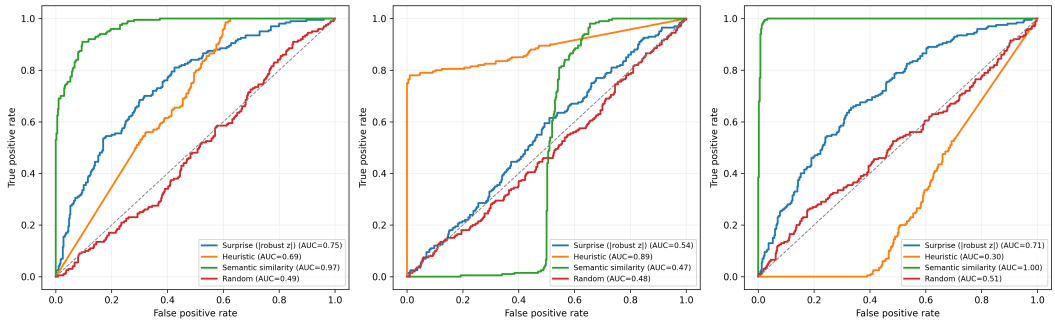

Figure 6: **One-vs-Rest ROC Analysis of Salience Signals.** Performance comparison between the Surprise signal (Robust z-score) and baselines (Heuristic, Semantic Similarity). *Left:* Detection of Routine samples vs. others. *Center:* Detection of Critical samples vs. others. *Right:* Detection of Out-of-Distribution (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.

A.8 Sensitivity to Threshold and Temporal Smoothing

Figure 7 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.

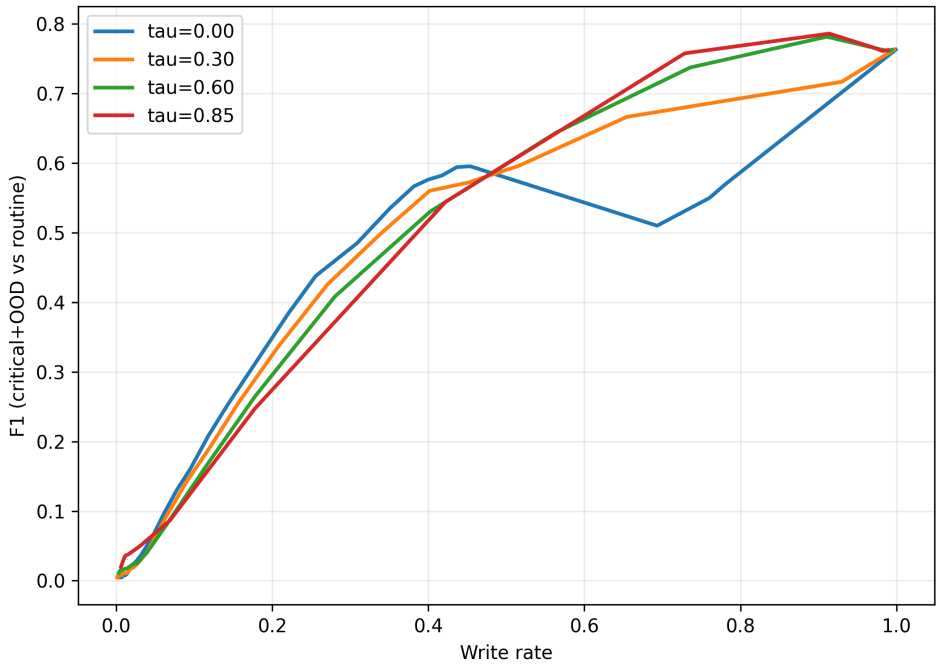

Figure 7: **Impact of Smoothing on Write Sensitivity.** Performance (F1) vs. Cost (Write Rate) for varying smoothing factors α. While the unsmoothed signal ($\alpha = 0$) degrades at intermediate thresholds, strong smoothing ($\alpha = 0.85$) ensures consistent performance gains as the write budget increases, validating the hypothesis that critical events manifest as sustained periods of high surprise.

A.9 Structural Divergence in Residual Space

Figure 10 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 distinguish 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.10 Surprise Distributions Under Domain Shift

Figure 8 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 do-

mains 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.

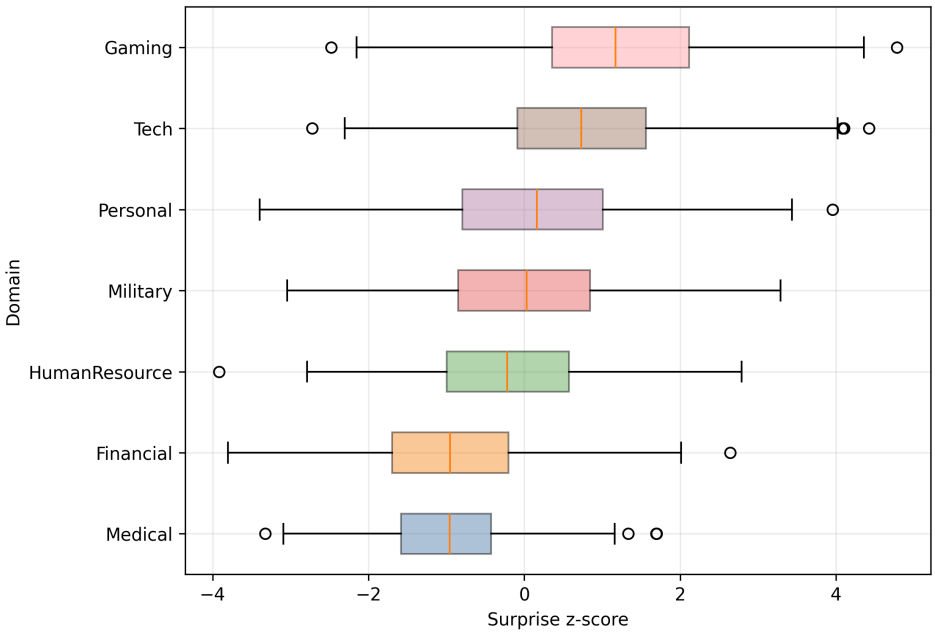

Figure 8: **Surprise distributions under domain shift.** Box plots showing the Surprise z-score for inputs categorized by thematic domains. These data points represent the same Routine, Critical, and OOD interactions analyzed previously, sliced by topic. The shifting medians indicate that domains with specialized or informal jargon (e.g., Gaming, Tech) inherently exhibit a higher baseline surprise compared to more strictly structured domains (e.g., Medical, Financial).

A.11 Cross-domain robustness table

Table 5 provides the full cross-domain generalization results referenced in Section A.1. 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.12 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 9 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

Table 5: 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

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

A.13 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.

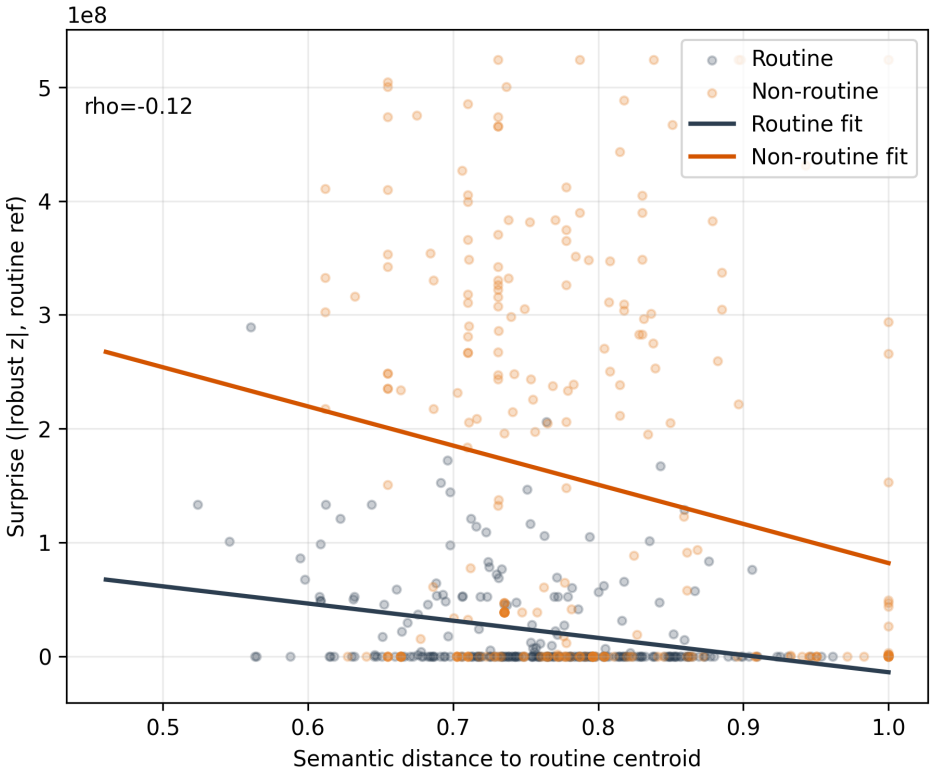

Figure 9: **Surprise vs. Semantic Distance Correlation.** Scatter plot comparing the SAE surprise metric against the semantic distance to the routine centroid. The near-zero correlation ($\rho = -0.12$) demonstrates that SAE surprise is not just "embedding distance in disguise." Non-routine events (orange) consistently exhibit higher surprise across the entire spectrum of semantic distances, proving that the SAE captures mechanistic deviations that traditional thematic similarity misses.

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 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 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

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

A.14 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 8).

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

- **Deep Layers for Macro-Semantic Shifts (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.15 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:

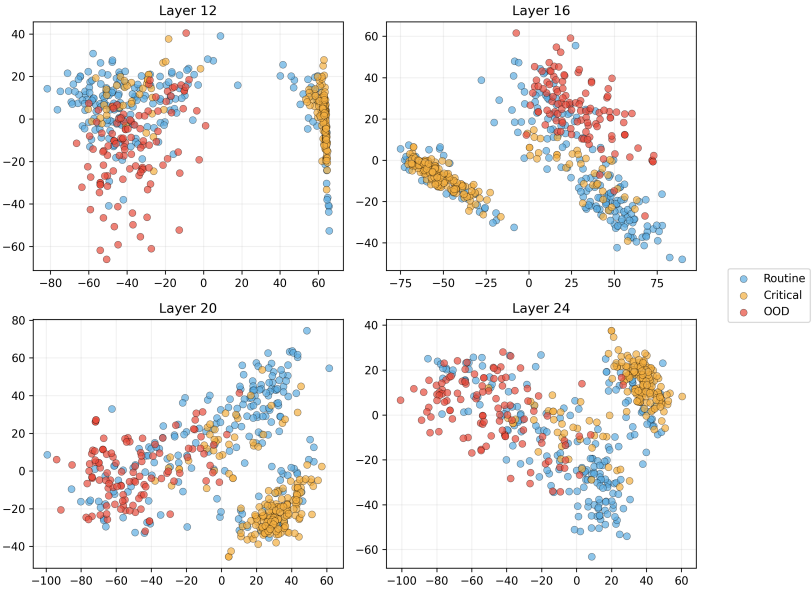

Figure 10: **PCA projections of SAE residuals across layers.** Each panel shows the first two principal components of residual vectors $r = x - \hat{x}$ for Routine, Critical, and OOD inputs. Overlap is stronger in earlier layers, while deeper layers exhibit clearer geometric separation (notably for OOD), indicating structured directionality in residual space rather than pure magnitude effects.

- **Domain-stable SAE behavior.** Despite domain shifts (Figure 8), 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 9, $\rho = -0.12$), making it a complementary detector for anomalies that remain semantically in-domain.

A.16 Residual geometry across depth.

Complementing the transfer results, Figure 10 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 states differ in residual space beyond scalar reconstruction error alone.

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

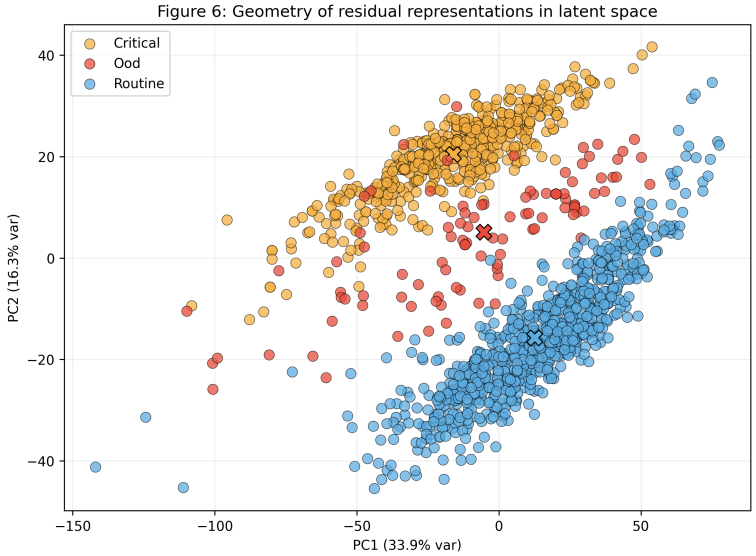

Figure 11: **Residual geometry.** PCA of residuals $r_t = h_t - \hat{h}_t$ exhibits structured regime separation: Routine residuals concentrate near the origin, Critical turns form directionally coherent clusters, and OOD inputs are typically higher-magnitude but more diffuse.

A.17 Residual Geometry in SAE Error Space

Figure 11 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 `meta-llama/Llama-3.2-3B-Instruct` and `Qwen/Qwen3-4B` 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.

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.

Algorithm 1 ReSuME: surprise-gated episodic memory (online)

Require: 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 $s_0 \leftarrow 0$.
for $t = 1, 2, \ldots$ **do**
  Compute representation $x_t \leftarrow \text{Pool}(\{h_{t,i}^{(\ell)}\}_{i=1}^{T_t})$
  $z_t \leftarrow g(x_t)$; $\hat{x}_t \leftarrow d(z_t)$
  $e_t \leftarrow \|x_t - \hat{x}_t\|_2$
  Update rolling statistics (μ_t, σ_t) from $\mathcal{H}$ (keep last W errors)
  $u_t \leftarrow \text{clip}\big((e_t - \mu_t)/(\sigma_t + \varepsilon), -c, c\big)$ ▷ instantaneous robust score
  $s_t \leftarrow \alpha\, u_t + (1 - \alpha)\, s_{t-1}$ ▷ EMA-smoothed surprise
  Append e_t to $\mathcal{H}$
  if $t > T_{\text{warm}}$ **and** $\nVdash_{\text{user}}(t) = 1$ **and** $s_t > \tau$ **then**
    Insert $(\tilde{x}_t, s_t, t, \text{meta}_t, \text{text}_t)$ into $\mathcal{M}$
    if $|\mathcal{M}| > M$ **then**
      Evict one item from $\mathcal{M}$ (e.g., lowest surprise)
    end if
  end if
  Retrieve top-k items by cosine similarity $\langle \tilde{x}_t, \tilde{x}_i \rangle$ over $m_i \in \mathcal{M}$
  if $\max_i \langle \tilde{x}_t, \tilde{x}_i \rangle \geq \gamma$ **then**
    Expose retrieved items to the agent (e.g., as additional context)
  end if
end for

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)

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 3.2. 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:

$$\hat{\sigma}_t = c_{\text{MAD}} \cdot \text{median}\left(|W_t - \text{median}(W_t)|\right) \tag{11}$$

where $c_{\text{MAD}} \approx 1.4826$ 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:

$$u_t = \frac{e_t - \text{median}(W_t)}{\hat{\sigma}_t + \epsilon} \tag{12}$$

where ϵ is a small constant for numerical stability.

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	$1e^{-6}$

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:

$$s_t = \alpha u_t + (1 - \alpha) s_{t-1}. \tag{13}$$

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.