The Mechanics of Interference: Defusing Distractors in RAG via Sparse Autoencoder Interventions

Christian Giannetti[1], Giovanni Trappolini[2,3], Nicola Tonellotto[4,3], Fabrizio Silvestri[1,3], Pietro Liò[5]

[1]Sapienza University of Rome, Italy [2]Universitas Mercatorum, Italy [3]ISTI-CNR, Italy
[4]University of Pisa, Italy [5]University of Cambridge, UK

Correspondence: giannetti@diag.uniroma1.it

Abstract

Large language models exhibit a critical vulnerability to distractor interference in retrieval-augmented contexts: they fail to prioritize relevant, factually correct documents over topically similar but misleading content. We introduce LAT-DEFUSE, a mechanistic framework that corrects this failure mode through targeted interventions in the model's latent space. Using Sparse Autoencoders (SAEs), our method operates in an interpretable feature space and formulates correction as constrained counterfactual optimization. On Gemma-2 and Llama-3 model families across three QA benchmarks (BioASQ, Natural Questions, PopQA), our method achieves recovery rates of up to 94% on distractor-vulnerable samples. Successful correction through sparse modifications reveals distractor interference as a localized, systematically addressable phenomenon, opening directions toward universal distractor robustness in LLMs.

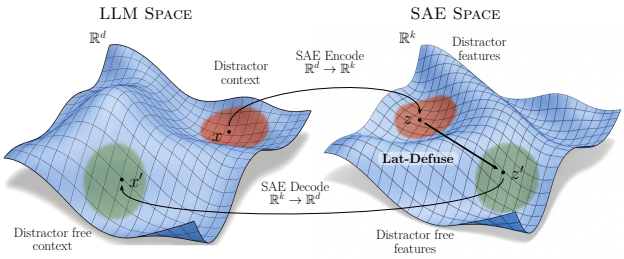

Figure 1: Lat-Defuse overview. Distractor-corrupted activations $x \in \mathbb{R}^d$ are encoded into SAE feature space, yielding latent representation $z \in \mathbb{R}^k$. Constrained optimization produces z' by modulating distractor-associated features while preserving gold-document cues. Decoding yields corrected activations x', restoring factual generation.

1 Introduction

Large language models exhibit a critical vulnerability when processing (retrieval-augmented) contexts: they struggle to distinguish between relevant, factually correct information, and topically similar but misleading content (Sauchuk et al., 2022). This weakness manifests prominently in two related phenomena: the "lost-in-the-middle" effect in retrieval-augmented generation (RAG) systems, where models fail to prioritize gold documents surrounded by distractors (Liu et al., 2024), and context rot (Hong et al., 2025) in long-context scenarios, where models lose focus amid extensive but partially irrelevant information. Current approaches to mitigating distractor interference operate at two distinct levels. Input-side methods (Zhuang et al., 2025) focus on improving retrieval quality through enhanced document ranking and filtering mechanisms, while reasoning-side approaches (Xu et al., 2025) aim to improve the model's ability to critically evaluate and synthesize information through chain-of-thought prompting or verification procedures. However, these methods incur computational overhead and function as architectural stopgaps that address symptoms without investigating the underlying cause. A central question remains unanswered: **which representational changes occur in the model's latent space when topically similar but factually incorrect documents induce generation errors?** Without understanding how distractors corrupt internal representations, solutions remain limited to symptom-level mitigation.

To investigate whether distractor interference can be defused in latent space, we study the maximally challenging configuration: contexts containing the gold document alongside hard distractors that induce generation errors. Scenarios involving irrelevant noise or weaker distractors represent relaxations of this setting. We introduce Lat-Defuse, a mechanistic framework for analyzing and correcting distractor interference through direct intervention in the model's latent space. Using Sparse Autoencoders (SAEs) to operate in an interpretable feature space, our method formulates correction

Findings of the Association for Computational Linguistics: ACL 2026, pages 11999–12007

as constrained optimization targeting the point at which generation first departs from factual correctness, identifying the minimal representational changes required to restore factual accuracy.

We evaluate our approach on Gemma-2 and Llama-3 model families across three QA benchmarks (BioASQ, Natural Questions, PopQA), achieving recovery rates up to 94% on distractor-vulnerable samples. Successful correction through sparse feature modifications indicates that distractor interference operates through **localized** representational changes rather than global perturbations. Our contributions are: **(a)** we introduce SAE-based latent intervention as a mechanistic framework for characterizing how distractor interference degrades factual accuracy in RAG; **(b)** we demonstrate that distractor interference can be corrected through targeted modifications to a sparse set of interpretable features; **(c)** we show that the correction mechanism generalizes across architectures (Gemma-2, Llama-3) and model scales. This mechanistic characterization provides a foundation toward universal solutions for distractor robustness in language models. [1]

2 Related Work

Retrieval-augmented generation is vulnerable to hard distractors: documents similar to the query but factually misleading. Cuconasu et al. (2024) demonstrate that such distractors effectively bypass relevance filtering, while Amiraz et al. (2025) show that stronger retrievers paradoxically worsen robustness by fetching these semantic mimics. Unlike random noise, distractors induce active hallucinations rather than abstention (Yoran et al., 2024), presenting a verification challenge analogous to synthetic disinformation attacks (Du et al., 2022). Hong et al. (2025) formalize the cumulative damage of misleading data as context rot, where distractors destabilize the broader context representation. This vulnerability is position-invariant (Cuconasu et al., 2025), undermining attention-based lost-in-the-middle hypotheses (Liu et al., 2024) and suggesting the error originates from internal representational corruption rather than attention span limitations. To probe these latent mechanisms, we leverage Sparse Autoencoders (SAEs) to decompose polysemantic activations into interpretable, monosemantic features (Bricken et al., 2023; Gao et al., 2025). SAE analysis builds on the transformer circuits framework (Elhage et al., 2021) and the view of feed-forward layers as key-value memories (Geva et al., 2021). Representation engineering modulates global attributes (Zou et al., 2023; Turner et al., 2023) and injects task vectors (Hendel et al., 2023). However, RAG-specific variants (Zhao et al., 2025; Shi et al., 2024) mainly address **inter-source conflict** (parametric vs. retrieved knowledge). These methods treat retrieved context as a monolithic unit, thereby lacking the granularity to resolve **intra-context competition**. Specifically, they cannot selectively suppress misleading distractor features while preserving the gold document needed to answer correctly.

[1] Official code and instructions to reproduce the experiments are available at https://github.com/christian-giannetti/mechanics-of-interference.

3 Methodology

Let M be a language model over vocabulary $\mathcal{V}$. We consider a retrieval-augmented setting with input query q and context $C = \{d_g, d_{\text{dist}}\}$, comprising a gold document d_g and semantic distractors d_{dist}. The model's predictions deviate from the ground-truth sequence Y^*. We target the first divergence point: the earliest token index t where the model's argmax prediction differs from Y_t^*. Let $y^* \in \mathcal{V}$ denote the correct token at this position. We intervene on the residual stream of M via a pretrained Sparse Autoencoder (SAE) with encoder $E : \mathbb{R}^d \to \mathbb{R}^K$ and decoder $D : \mathbb{R}^K \to \mathbb{R}^d$. We define an intervention window over the final W tokens of the input context. For each token index i, the *modulated activations* $\tilde{a}_i = D(z_i)$ are the decoder output for latent state $z_i \in \mathbb{R}^K_{\geq 0}$, where non-negativity is enforced to match the SAE's ReLU gating. The optimization objective $\mathcal{L}$ is the cross-entropy loss at divergence position t, conditioned on these modulated activations:

$$\mathcal{L} = -\log P_M(y^* \mid q, C, \tilde{a}) \tag{1}$$

Optimization terminates when the prediction correction condition holds:

$$\arg\max_{v \in \mathcal{V}} P_M(v \mid q, C, \tilde{a}) = y^* \tag{2}$$

We optimize the latent features with projected gradient updates, initializing $z_i^{(0)} = E(a_i^{(0)})$ from the original forward-pass activations $a_i^{(0)}$ to start from the distractor-corrupted state. The parameters of M and the SAE remain frozen; only z_i is updated. Each iteration k follows a three-step cycle: (i) decode the latent state to produce modulated activa-

tions $\tilde{a}_i = D(z_i^{(k)})$, (ii) inject $\tilde{a}_i$ into the computation graph of M to compute $\nabla_{z_i}\mathcal{L}$, and (iii) apply a projected gradient step enforcing non-negativity via ReLU. The update rule is:

$$z_i^{(k+1)} \leftarrow \text{ReLU}\left(z_i^{(k)} - \eta \nabla_{z_i}\mathcal{L}(z_i^{(k)})\right) \quad (3)$$

where η is the learning rate. This procedure yields a counterfactual latent state z^* satisfying the prediction correction condition. The feature-space delta $\Delta z = z^* - z^{(0)}$ identifies the modification sufficient to reverse the distractor-induced error.

4 Experimental Settings

4.1 Models

We evaluate our approach using the Gemma-2 (2B) and Llama-3 (8B) model families.[2] To facilitate latent space interventions, we use Gemma Scope (Lieberum et al., 2024) and Llama Scope (He et al., 2024) Sparse Autoencoders (SAEs) with dictionary sizes $K \in \{16\text{k}, 32\text{k}\}$.[3] We intervene at four network depths to assess layer-wise susceptibility: embedding (0%), early-mid (40%), mid-late (70%), and late (90%).

4.2 Datasets and Setup

We employ the distractor benchmark of Trappolini et al. (2026), aggregating samples from BioASQ, Natural Questions, and PopQA. To instantiate the maximally challenging configuration, the benchmark enforces a "lost-in-the-middle" setup (Liu et al., 2024): the gold document d_g flanked by two hard distractors, where interference is most severe. We filter strictly for **distractor-vulnerable** instances where distractors induce generation errors, yielding approximately 255 samples per model. We optimize feature vectors with Adam ($\eta = 0.0005$) over the final W tokens of the input context at the specified SAE layers. Response fidelity is evaluated via a hierarchical LLM-as-a-judge protocol using Gemini 2.5 Flash (Comanici et al., 2025) to filter null responses, verify named entities, and assess semantic equivalence. We quantify intervention efficacy using the **Trajectory Recovery Rate** (ρ): the percentage of distractor-vulnerable instances for which the intervention produces a semantically correct response. Existing contrastive decoding methods (Shi et al., 2024) address inter-source conflict between parametric knowledge and retrieved context, but established baselines for intra-context competition between gold and distractor documents are not available. To test whether distractor interference is recoverable via output-layer correction, we adapt their approach to this setting, introducing Distractor-Contrastive Decoding (DCD) and Marginal Distractor Removal (MDR). Both apply contrastive adjustments to the output distribution using a tunable *correction weight* $\alpha \in \{0.5, 1.0\}$ and require oracle knowledge of document roles (d_g vs. d_{dist}). Full derivations are provided in Appendix B.

Table 1: Trajectory Recovery Rate (%) comparison. Our method (bottom) is evaluated against the DCD and MDR competing methods (top). **Bold** indicates the best result per column. Configurations for our method are tuples (L, W), where L is the Intervention Layer and W is the Optimization Window. 95% bootstrap confidence intervals are reported in Appendix F.

	Gemma 2 2B			Llama 3.1 8B		
Config	**BioASQ**	**NQ**	**PopQA**	**BioASQ**	**NQ**	**PopQA**
Competing Methods						
$\text{DCD}_{\alpha=0.5}$	47.81	33.98	44.01	42.70	40.71	52.15
$\text{DCD}_{\alpha=1.0}$	53.78	44.66	56.34	44.94	44.69	64.11
$\text{MDR}_{\alpha=0.5}$	44.62	34.63	55.99	35.96	39.82	72.25
$\text{MDR}_{\alpha=1.0}$	48.21	46.93	75.35	37.08	43.36	78.47
Our Method						
$(L1, 10)$	66.93	75.40	91.55	66.29	84.79	**93.93**
$(L1, 100)$	**70.52**	76.05	92.61	70.79	85.17	93.12
$(L1, 400)$	**70.52**	**80.58**	93.31	71.35	**85.93**	93.52
$(L2, 10)$	69.32	77.99	92.61	66.85	81.37	91.50
$(L2, 100)$	68.13	76.38	92.25	65.17	82.89	92.71
$(L2, 400)$	66.53	79.94	92.61	64.61	83.65	92.31
$(L3, 10)$	68.13	78.32	92.96	71.35	83.27	91.90
$(L3, 100)$	68.13	79.61	92.96	70.79	85.55	91.09
$(L3, 400)$	68.13	78.32	91.55	**71.91**	82.51	90.28
$(L4, 10)$	68.92	78.32	92.96	67.98	81.75	89.88
$(L4, 100)$	68.13	77.67	93.31	66.85	79.85	89.47
$(L4, 400)$	68.92	78.32	**93.66**	64.61	80.61	89.88

Gemma 2 2B (26 layers): L1=0, L2=10, L3=18, L4=24.
Llama 3.1 8B (32 layers): L1=0, L2=12, L3=22, L4=30.

5 Results

Quantitative Results Table 1 reports the Trajectory Recovery Rate (ρ) across layer configurations

[2]Model checkpoints: Gemma-2-2B (https://huggingface.co/google/gemma-2-2b) and Meta-Llama-3-8B (https://huggingface.co/meta-llama/Meta-Llama-3-8B).

[3]SAE checkpoints: Gemma Scope (https://huggingface.co/google/gemma-scope-2b-pt-res) and Llama Scope (https://huggingface.co/OpenMOSS-Team/Llama3_1-8B-Base-LXR-8x/tree/main).

Suppression Guided Correction

Question: "When does the sa node begin electrical signaling?"
Distractor 1: "Normal sinus rhythm [...]"
Gold Relevant: "One percent of the [...]"
Distractor 2: "Electrical activity [...]"

Original Answer: When heart is at rest

↓ Dampened Feature: PHYSIOLOGICAL PROCESSES

Distractor 1 details *how* the impulse propagates but never states *when* it begins. Dampening "Physiological Processes" **de-emphasizes pathway mechanics,** surfacing the gold's explicit **"spontaneously."**
Evidence (Distractor): "This impulse spreads [...] throughout the atria through **specialized internodal pathways**."
Evidence (Gold): "One percent of the cardiomyocytes [...] possess the ability to generate electrical impulses **spontaneously**."

Corrected Answer: Spontaneously

Amplification Guided Correction

Question: "Who sings you've got to hide your love away?"
Distractor 1: "The Beau Brummels [...]"
Gold Relevant: "You've Got to Hide [...]"
Distractor 2: "They were helped by [...]"

Original Answer: The Silkie

↑ Amplified Feature: EXCLAMATIONS

The exclamation feature **highlighted the album title *Help!*** in the Gold Document, guiding the model toward John Lennon as the original singer **rather than cover artists.**

Evidence (Gold): "**Written and sung by John Lennon** [...] released on the album **"Help!"**"

Corrected Answer: John Lennon

and datasets. $L1$ interventions yield the highest recovery rates, reaching 71.35% (BioASQ) and 85.93% (NQ) for Llama-3. The superiority of $L1$ interventions supports the **causal suppression hypothesis**: deactivating distractor representations at early layers prevents their propagation through the residual stream. Gemma-2 exhibits greater depth stability, with $L4$ trailing $L1$ by only 2.26 points on NQ. PopQA exhibits a ceiling effect ($\rho > 90\%$) across both models. Window extension to $W{=}400$ saturates performance, indicating that layer selection dominates over window size. Our method substantially outperforms the logit-space competitors: on NQ, $(L1, 400)$ exceeds the best alternative by 33.65 points for Gemma-2 and 42.57 points for Llama-3, confirming that distractor corruption cannot be corrected at the output layer alone.

Qualitative Results We trace these quantitative gains to a sparse set of causal features, enforcing global support of at most 300 active features across all tokens in the intervention window via Gradient Hard Thresholding Pursuit (Yuan et al., 2018). To isolate interpretable drivers, we rank features by the magnitude of their **net activation displacement** (Δz), while filtering out high-frequency structural latents ($> 1\%$ density) to focus strictly on sparse semantic concepts (Rajamanoharan et al., 2025; Sun et al., 2025). Figure 2 (top) provides direct evidence for the causal suppression hypothesis: in this biological query, the baseline incorrectly infers that the SA node fires "at rest" because distractors describe general sinus rhythm. The intervention dampens the identified *Physiological Processes* feature ($\Delta z \approx -4.54$), effectively muting this misleading context. This suggests the optimization acts as a semantic filter, suppressing distractor features to surface the overshadowed gold attribute. We provide quantitative analysis in Appendix C. Figure 2 (bottom) demonstrates discriminative amplification: when distinguishing between the original Beatles recording and a cover artist, the intervention amplifies an *Exclamations* feature ($\Delta z \approx +1.78$), increasing the salience of the punctuation in the album title *Help!*, a token unique to the gold document. These patterns confirm that interventions modulate distractor-associated features rather than inject external knowledge.

6 Conclusions

This work characterizes the representational changes that occur when distractors induce generation errors. Our findings indicate that distractor interference operates through localized, correctable changes rather than global perturbations. Analysis of successful corrections reveals two complementary mechanisms: suppression of distractor features and amplification of discriminative gold-document cues. Lat-Defuse exploits this structure by identifying minimal modifications to a sparse set of interpretable features at the point where generation diverges from factual correctness. The correction mechanism generalizes across architectures (Gemma-2, Llama-3) and scales, achieving recovery rates up to 94%. By characterizing the representational basis of distractor interference, this work moves beyond symptom-level mitigation toward understanding root causes. Although currently dependent on ground-truth labels, these findings open directions toward universal distractor robustness in LLMs.

Limitations

Methodological Constraints The proposed approach uses ground-truth labels to identify the first incorrect token and guide optimization. This supervision requirement limits immediate deployment in production systems, though it establishes a foundation for future unsupervised methods. The optimization procedure involves computational overhead from multiple forward passes through both the SAE and the language model, detailed in Appendix E. Given our evaluation on question-answering benchmarks with relatively short contexts, the method's applicability to industry scenarios with longer contexts or less defined correctness criteria remains to be validated. The focus on correcting the first error represents a design choice that may also function as a form of regularization, though the theoretical basis for this effect remains an open question for future investigation.

Interpretability Tool Limitations The method relies on Sparse Autoencoders to decompose activations into interpretable features. SAEs represent learned approximations of the underlying feature space and may not capture all relevant semantic dimensions, though recent work (la Tour and Mossing, 2025) indicates that reconstruction quality improves with scale and architectural innovations. The feature-level explanations derived from this analysis are post-hoc interpretations of model behavior. While the analysis identifies consistent patterns of suppression and amplification, establishing causal relationships between specific features and semantic content remains an active area of research in mechanistic interpretability.

Ethical Considerations

This work investigates representational mechanisms of distractor interference in language models. It does not involve human subjects, private data, or demographic annotations; all datasets (BioASQ, Natural Questions, PopQA) are publicly available benchmarks, and the models (Gemma-2, Llama-3) and SAEs (Gemma Scope (Lieberum et al., 2024), Llama Scope (He et al., 2024)) are publicly released artifacts with documented model cards.

Regarding dual use, the features identified by the intervention are query-specific latent activations within a frozen model, not transferable attack primitives. The framework is designed to improve factual robustness in retrieval-augmented settings, and the mechanistic understanding it provides may aid the development of defenses against retrieval poisoning attacks (Du et al., 2022).

Evaluation relies on an LLM-as-a-judge protocol using Gemini 2.5 Flash (Comanici et al., 2025). Computational requirements for the intervention procedure are reported in Appendix E. No crowdworkers or human annotators were employed.

References

Chen Amiraz, Florin Cuconasu, Simone Filice, and Zohar Karnin. 2025. The distracting effect: Understanding irrelevant passages in RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Florin Cuconasu, Simone Filice, Guy Horowitz, Yoelle Maarek, and Fabrizio Silvestri. 2025. Do rag systems really suffer from positional bias? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

Yibing Du, Antoine Bosselut, and Christopher D Manning. 2022. Synthetic disinformation attacks on automated fact verification systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10581–10589.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, and 1 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2025. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations (ICLR)*.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, Yu-Gang Jiang, and Xipeng Qiu. 2024. Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders. *arXiv preprint arXiv:2410.20526*.

Roee Hendel, Mor Geva, and Amir Globerson. 2023. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9318–9333. Association for Computational Linguistics.

Kelly Hong, Anton Troynikov, and Jeff Huber. 2025. Context rot: How increasing input tokens impacts LLM performance. Technical report, Chroma Research.

Tom Dupré la Tour and Dan Mossing. 2025. Debugging misaligned completions with sparse-autoencoder latent attribution. OpenAI Alignment Blog. Accessed: 2026-01-03.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop at EMNLP 2024*, pages 278–300.

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.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. 2025. Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Artsiom Sauchuk, James Thorne, Alon Halevy, Nicola Tonellotto, and Fabrizio Silvestri. 2022. On the role of relevance in natural language processing tasks. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1785–1789.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*.

Xiaoqing Sun, Alessandro Stolfo, Joshua Engels, Ben Peng Wu, Senthooran Rajamanoharan, Mrinmaya Sachan, and Max Tegmark. 2025. Dense sae latents are features, not bugs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Giovanni Trappolini, Florin Cuconasu, Simone Filice, Yoelle Maarek, and Fabrizio Silvestri. 2026. Redefining retrieval evaluation in the era of LLMs. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8359–8375, Rabat, Morocco. Association for Computational Linguistics.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, and 1 others. 2025. Towards large reasoning models: A survey of reinforced reasoning with large language models. *Patterns*, 6(10):101370.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Xiao-Tong Yuan, Ping Li, and Tong Zhang. 2018. Gradient hard thresholding pursuit. *Journal of Machine Learning Research*, 18(166):1–43.

Yu Zhao, Alessio Devoto, Giwon Hong, Xiaotang Du, Aryo Pradipta Gema, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. 2025. Steering knowledge selection behaviours in llms via sae-based representation engineering. In *Proceedings of the 2025 Annual Conference of the Nations of the Americas Chapter of the ACL (NAACL)*.

Shengyao Zhuang, Xueguang Ma, Bevan Koopman, Jimmy Lin, and Guido Zuccon. 2025. Rank-R1: Enhancing reasoning in LLM-based document rerankers via reinforcement learning. *arXiv preprint arXiv:2503.06034*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, and 1 others. 2023. Representation engineering: A top-down approach to ai transparency. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.

A Distractor Selection Protocol

Distractors are selected following Trappolini et al. (2026). For each query, candidates are first ranked by dense similarity, then filtered by an LLM judge that issues a binary verdict on whether a passage is sufficiently misleading to induce an incorrect answer despite being factually wrong. The retained distractor is the candidate with the highest distracting score, defined per Amiraz et al. (2025) as the probability of an incorrect generation conditioned on the passage alone. Although this definition is tied to a specific model's behavior, Amiraz et al. (2025) report Spearman correlations of 0.47–0.76 across model families and scales (Llama, Falcon, Qwen; 3B–70B), indicating that distractiveness is largely an intrinsic property of the passage and transfers across the backbones evaluated here.

B Logit-Space Comparison Method Definitions

Logit-space corrections represent the computationally minimal intervention class: they modify the output distribution without altering internal representations. If distractor interference were fully recoverable through such corrections, representational methods would be unnecessarily invasive. To test this hypothesis, we derive two logit-space comparison methods that apply contrastive principles at the output layer. Whereas Contrastive Activation Decoding (Shi et al., 2024) factors out parametric priors to amplify context-induced signal, our setting requires the inverse operation: factoring out distractor-induced corruption from a multi-document context.

We define three forward-pass configurations over vocabulary $\mathcal{V}$: $\mathcal{L}_{GD} = \text{logit}(y \mid d_g, d_{\text{dist}}, q, y_{<t})$ (full corrupted context), $\mathcal{L}_G = \text{logit}(y \mid d_g, q, y_{<t})$ (gold only), and $\mathcal{L}_D = \text{logit}(y \mid d_{\text{dist}}, q, y_{<t})$ (distractor only). To ensure that logit subtraction isolates semantic rather than positional differences, we enforce strict length alignment: shorter contexts ($\mathcal{L}_G$, $\mathcal{L}_D$) are padded so that the query occupies the same absolute position as in $\mathcal{L}_{GD}$.

The first baseline, Distractor-Contrastive Decoding (DCD), assumes that distractor interference is additive and context-independent. The corrected logits subtract the distractor-only component:

$$\mathcal{L}_{\text{DCD}} = (1 + \alpha) \cdot \mathcal{L}_{GD} - \alpha \cdot \mathcal{L}_D \tag{4}$$

This formulation treats $\mathcal{L}_D$ as a proxy for the distractor contribution in logit space, assuming that distractor features activate similarly regardless of gold document presence.

The second baseline, Marginal Distractor Removal (MDR), models interference as arising from gold-distractor interactions rather than from the distractor alone. The corrected logits interpolate between corrupted and clean states:

$$\mathcal{L}_{\text{MDR}} = (1 - \alpha) \cdot \mathcal{L}_{GD} + \alpha \cdot \mathcal{L}_G \tag{5}$$

At $\alpha = 1$, MDR recovers the gold-only distribution, whereas DCD at the same value performs contrastive amplification ($2\mathcal{L}_{GD} - \mathcal{L}_D$). This asymmetry reflects their divergent assumptions: MDR captures interaction effects absent from the distractor-only pass, while DCD treats such interactions as negligible.

Both methods carry deployment constraints: they require oracle knowledge of document roles (d_g vs. d_{dist}) and strict length alignment. More critically, logit-space corrections are opaque, providing no access to the latent features responsible for distraction. DCD and MDR therefore function as depth probes rather than interpretable correction mechanisms; the SAE intervention, operating on decomposed feature activations, uniquely exposes the monosemantic units driving correction. We evaluate $\alpha \in \{0.5, 1.0\}$, as intermediate values yielded minimal change and $\alpha > 1.0$ degraded generation quality. The substantial underperformance of both methods relative to the SAE intervention (Table 1) establishes that distractor corruption is representational rather than surface-level noise amenable to logit-space correction.

C Mechanistic Evidence for Suppression

To distinguish whether correction operates via suppression of distractor representations or via target amplification alone, Table 2 reports two complementary metrics computed at the first-token prediction step across all successfully corrected samples. *Top-100 Competitor suppression* measures the fraction of the 100 highest-probability baseline tokens (excluding the target) whose logits decrease post-intervention. *Distractor Vocabulary suppression* measures the corresponding fraction among tokens appearing in distractor documents. If the intervention merely amplified the target, competitor logits would remain stable relative to one another. Widespread competitor suppression therefore constitutes direct evidence for active defusing of distractor representations.

Table 2: Suppression metrics at the first-token prediction step across successfully corrected samples. Top-100 Competitor suppression and Distractor Vocabulary suppression report the fraction of tokens with decreased logits post-intervention. Pearson correlation (r) quantifies sample-level association between both metrics.

	Gemma 2 2B			Llama 3.1 8B		
	BioASQ	**NQ**	**PopQA**	**BioASQ**	**NQ**	**PopQA**
Tokens with Decreased Logit (%)						
Top-100 Competitors	90.5 ± 6.8	89.4 ± 7.3	89.2 ± 13.0	82.7 ± 13.1	90.0 ± 6.3	90.4 ± 9.2
Distractor Vocabulary	59.5 ± 16.2	71.2 ± 10.3	83.0 ± 10.4	36.2 ± 20.3	69.6 ± 11.7	64.8 ± 12.7
Pearson Correlation (r)						
Competitor vs Vocab Suppression	0.62	0.43	0.74	0.77	0.52	0.71

Across all configurations, Top-100 Competitor suppression ranges from 82.7% to 90.5%, indicating that approximately 9 in 10 of the strongest baseline alternatives are actively pushed down in logit space rather than merely overtaken by target amplification. Distractor Vocabulary suppression exhibits greater variance (36.2%–83.0%), reflecting differences in how distractor content distributes across model vocabularies: configurations with concentrated distractor lexicons yield higher suppression rates than those where domain-specific terminology fragments across numerous low-frequency subword units.

The Pearson correlation between competitor and vocabulary suppression ($r = 0.43$–0.77, all $p < 10^{-12}$) confirms that both metrics capture a coherent mechanism at the sample level. When suppression is effective, both metrics are elevated; when suppression is weaker, both decline together. This pattern is most pronounced in Llama-3 on BioASQ, where the correlation reaches its maximum ($r = 0.77$) despite the lowest mean vocabulary suppression, indicating that suppression-based correction is more difficult when model priors for domain vocabulary are weaker, yet the underlying mechanism remains intact.

These results support the causal suppression hypothesis: the intervention corrects distractor interference by dampening competing token probabilities rather than amplifying the target in isolation. The dominant pattern across configurations, combined with strong sample-level correlations, indicates that suppression constitutes the primary correction pathway, consistent with the qualitative evidence in Figure 2.

D Distractor-Vulnerable Subset

Table 3 characterizes the distractor-vulnerable subset: samples on which the unmodified model is correct without distractors and incorrect under the surrounded configuration $[D, R, D]$. Vulnerability ranges from 22.8% to 37.5% across model and dataset combinations, defining the recovery target on which ρ in Table 1 is computed. Llama 3.1 8B exhibits lower vulnerability than Gemma 2 2B across all three datasets ($\Delta \approx 4.9$ to 9.3 points), consistent with the higher baseline robustness expected at larger scale.

Model	**Dataset**	**Distr. Vuln. (%)**	**Vuln. / Tot.**
Gemma 2 2B	BioASQ	32.1	251 / 781
	NQ	34.2	309 / 903
	PopQA	37.5	284 / 757
Llama 3.1 8B	BioASQ	22.8	178 / 780
	NQ	29.1	263 / 904
	PopQA	32.6	247 / 758

Table 3: Distractor-vulnerable subset under the surrounded configuration $[D, R, D]$. Samples are counted as vulnerable when the model answers correctly without distractors and incorrectly when distractors are added.

E Computational Cost

Let L denote the number of transformer layers, d the residual stream width, T the sequence length, V the vocabulary size, and B the batch size. The SAE has latent dimension K. Intervention occurs at a fixed layer ℓ, patching $P = |\text{target\_positions}|$ token positions, with optimized parameters $z \in \mathbb{R}^{P \times K}$. Each iteration of the optimization procedure comprises a transformer forward pass with differentiable SAE injection at layer ℓ, a boundary cross-entropy loss at a single token position, backpropagation to z with frozen model and SAE weights, and an Adam update on z. The dominant cost is a single forward and backward pass through the transformer; additive terms include SAE decoding at $\Theta(PKd)$, boundary loss evaluation at $\Theta(V)$, the optimizer update at $\Theta(PK)$, and acti-

vation materialization at the hook site at $\Theta(BTd)$. Peak memory scales with transformer activation storage for backpropagation, plus the patched activation buffer at $\Theta(BTd)$ and optimizer state at $\Theta(PK)$.

F Bootstrap Confidence Intervals

Table 4 reports conditional intervention accuracy on the baseline-incorrect subset, with each cell formatted as p [$\text{CI}_{2.5}$, $\text{CI}_{97.5}$], where p is the point estimate and the bracketed values denote the 2.5th and 97.5th percentiles of the bootstrap distribution (10,000 resamples). No statistical comparisons are performed across configurations; the intervals characterize per-cell sampling variability.

CI widths vary with sample size and accuracy level. For Gemma 2 2B, widths range from approximately 6 pp on PopQA ($N = 284$) to 11 pp on BioASQ ($N = 251$). For Llama 3.1 8B, BioASQ exhibits the widest intervals (approximately 13 pp), consistent with the smaller sample ($N = 178$), while NQ and PopQA show narrower intervals (approximately 7 to 9 pp). Across all configurations, the lower CI bounds remain substantially above the corresponding baseline accuracies reported in Table 1, confirming that the observed improvements are robust to sampling variability.

Table 4: Conditional intervention accuracy (%) on the baseline-incorrect subset with 95% bootstrap confidence intervals (10,000 resamples). Each cell reports: point estimate [$\text{CI}_{2.5}$, $\text{CI}_{97.5}$]. Sample counts per model: Gemma 2 2B (N: BioASQ=251, NQ=309, PopQA=284); Llama 3.1 8B (N: BioASQ=178, NQ=263, PopQA=247). No statistical comparisons are performed across configurations.

	Gemma 2 2B			**Llama 3.1 8B**		
(L, W)	**BioASQ**	**NQ**	**PopQA**	**BioASQ**	**NQ**	**PopQA**
$(L1, 10)$	66.93 [61.0, 72.5]	75.40 [70.5, 79.9]	91.55 [88.0, 94.7]	66.29 [59.0, 73.0]	84.79 [80.2, 89.0]	93.93 [90.7, 96.8]
$(L1, 100)$	70.52 [64.9, 76.1]	76.05 [71.2, 80.6]	92.61 [89.4, 95.4]	70.79 [64.0, 77.5]	85.17 [80.6, 89.3]	93.12 [89.9, 96.4]
$(L1, 400)$	70.52 [64.9, 76.1]	80.58 [76.0, 85.1]	93.31 [90.1, 96.1]	71.35 [64.6, 77.5]	85.93 [81.4, 90.1]	93.52 [90.3, 96.4]
$(L2, 10)$	69.32 [63.8, 74.9]	77.99 [73.5, 82.5]	92.61 [89.4, 95.4]	66.85 [59.5, 73.6]	81.37 [76.4, 85.9]	91.50 [87.8, 94.7]
$(L2, 100)$	68.13 [62.1, 73.7]	76.38 [71.5, 80.9]	92.25 [89.1, 95.1]	65.17 [57.9, 71.9]	82.89 [78.3, 87.5]	92.71 [89.5, 96.0]
$(L2, 400)$	66.53 [60.6, 72.1]	79.94 [75.4, 84.5]	92.61 [89.4, 95.4]	64.61 [57.3, 71.3]	83.65 [79.1, 87.8]	92.31 [88.7, 95.5]
$(L3, 10)$	68.13 [62.5, 73.7]	78.32 [73.8, 82.8]	92.96 [89.8, 95.8]	71.35 [64.6, 78.1]	83.27 [78.7, 87.8]	91.90 [88.3, 95.1]
$(L3, 100)$	68.13 [62.5, 73.7]	79.61 [75.1, 83.8]	92.96 [89.8, 95.8]	70.79 [64.0, 77.0]	85.55 [81.0, 89.7]	91.09 [87.5, 94.3]
$(L3, 400)$	68.13 [62.1, 73.7]	78.32 [73.8, 82.8]	91.55 [88.0, 94.7]	71.91 [65.2, 78.1]	82.51 [78.0, 87.1]	90.28 [86.6, 93.9]
$(L4, 10)$	68.92 [63.4, 74.5]	78.32 [73.8, 82.8]	92.96 [89.8, 95.8]	67.98 [60.7, 74.7]	81.75 [76.8, 86.3]	89.88 [85.8, 93.5]
$(L4, 100)$	68.13 [62.1, 73.7]	77.67 [73.1, 82.2]	93.31 [90.1, 96.1]	66.85 [59.5, 73.6]	79.85 [74.9, 84.4]	89.47 [85.4, 93.1]
$(L4, 400)$	68.92 [63.4, 74.5]	78.32 [73.8, 82.8]	93.66 [90.8, 96.5]	64.61 [57.3, 71.3]	80.61 [75.7, 85.5]	89.88 [85.8, 93.5]