Probing Bias Formation in Medical LLMs through Activation Steering

Bayram Ayadi, Annette Hautli-Janisz
Faculty of Computer Science and Mathematics
University of Passau
ayadi02@ads.uni-passau.de, annette.hautli-janisz@uni-passau.de

Abstract

Large Language Models specialized for the medical domain achieve high performance on static benchmarks, but remain vulnerable to sycophantic confabulation, where the models generate medically spurious rationales to justify incorrect user hints. This robustness gap poses severe risks in clinical environments, as models may prioritize contextual faithfulness to a biased prompt over their internal parametric medical knowledge. This study introduces a mechanistic approach to identify and mitigate these failures in MedGemma-27B, isolating hint integration circuits using Sparse Autoencoders and geometric manifold analysis. Our findings reveal that sycophantic bias is a highly distributed and polymorphic concept, with biased reasoning routed through shifting dimensions across transformer layers. We identify the optimal layer for intervention and demonstrate that cluster-conditioned dynamic steering tailored to the geometric subspace of the prompt outperforms static global interventions, though it reveals a fundamental tension between bias resilience and the retention of internal parametric knowledge. This work proposes a principled framework toward clinical AI systems that are more robust and aligned with expert medical logic, demonstrating the potential of cluster-conditioned geometric interventions while characterizing the inherent trade-offs in clinical knowledge retention.

1 Introduction

The rapid advancement of Large Language Models (LLMs) specialized for the medical domain, such as MedGemma (Sellergren et al., 2025) and Med-PaLM 2 (Singhal et al., 2023), has led to near-perfect performance on static medical benchmarks like MedQA (Jin et al., 2021). However, as these models are increasingly piloted for clinical decision support (Singhal et al., 2023; Vrdoljak et al., 2025), a critical robustness gap has emerged: Despite high benchmark scores, the models often fail to engage in genuine clinical reasoning. This fragility frequently manifests as sycophantic confabulation, where models generate persuasive but medically spurious post-hoc rationales to justify incorrect user hints (Kim et al., 2025a).

Empirical evaluations using the BiasMedQA dataset (Schmidgall et al., 2024) demonstrate that when models are presented with medically plausible but incorrect adversarial hints, their accuracy can collapse drastically (Manczak et al., 2025). In high-stakes healthcare environments, this susceptibility to hint injection poses severe risks. Whereas classical explainability relies on behavioral probes that identify these failures, they lack the granularity to uncover the internal computational mechanisms driving them (Bereska and Gavves, 2024).

To address this issue, our study introduces a novel mechanistic approach to identify and mitigate adversarial reasoning in clinical LLMs. We utilize Mechanistic Interpretability techniques, including Sparse Autoencoders (SAEs) (Bricken et al., 2023) and geometric manifold analysis to isolate the "hint integration circuits" activated during sycophantic failures. By identifying these internal representations, we perform targeted activation steering, forcing the model to rely on its internal parametric medical knowledge rather than the biased contextual prompt. This work shifts the focus from behavioral debiasing to diagnostic internal interventions and is guided by the following updated research questions:

> **RQ1** How is sycophantic bias mathematically represented within the residual stream of clinical LLMs, and is it localized to specific concept features or distributed across the network?
>
> **RQ2** At what network depth does the computational reasoning for adversarial hint integration solidify into the primary site for effective intervention?

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 609–620

RQ3 Can cluster-conditioned dynamic steering, which is tailored to the specific geometric subspace of a prompt, improve bias mitigation over static global interventions, while maintaining core clinical knowledge?

These research questions form a sequential diagnostic pipeline designed to move from internal characterization to targeted remediation. Characterizing the mathematical nature of the bias (RQ1) is a prerequisite for identifying its primary functional site (RQ2), as the distributed nature of the representation dictates the scope of the intervention. This localization, in turn, defines the optimal parameters for the cluster-conditioned mitigation strategies evaluated in RQ3. Collectively, this framework achieves the paper's goal of identifying adversarial reasoning through mechanistic discovery and mitigating it through surgical, context-aware interventions.

Guided by these research questions, our study provides several key contributions toward building more robust clinical AI systems. First, we provide a mechanistic characterization of sycophantic bias as a distributed and polymorphic concept where the geometric direction remains stable despite a rapidly shifting feature substrate. Second, we identify layer 29 as the primary site for bias consolidation using a layer-wise causal sweep. Third, we introduce a cluster-conditioned steering framework that improves mitigation rates to 29.2% compared to 21.9% for global interventions. Finally, we characterize the fundamental tension between bias resilience and the retention of parametric medical knowledge. Through this framework, we provide a principled pathway toward clinical AI systems that are more robust and aligned with expert medical logic. All data processing scripts and data are openly accessible. [1]

2 Related Work

2.1 Robustness and Bias in Clinical NLP

The deployment of medical LLMs is often justified by their high performance on specialized benchmarks like MedQA (Jin et al., 2021). However, recent evaluations using the Med-HALT benchmark have exposed significant hallucination tendencies, where model accuracy is largely a product of learning patterns from vast corpora rather than reliable clinical reasoning (Pal et al., 2023). Systematic perturbations using the BiasMedQA dataset, which was designed to test seven clinical cognitive biases such as anchoring and confirmation bias, reveal that even top-tier models like MedGemma-27B experience performance drops of over 17% when presented with biased prompts (Manczak et al., 2025; Kim et al., 2025b). Furthermore, the introduction of "None of the other answers" (NOTA) options has been shown to degrade accuracy from 80% to near 42%, suggesting that these models may fail to generalize when familiar multiple-choice structures are disrupted (Bedi et al., 2025).

2.2 Faithfulness vs. Plausibility in Medical CoT

Chain-of-Thought (CoT) prompting is frequently used to elicit interpretable medical reasoning (Singhal et al., 2023). Nevertheless, a growing body of evidence suggests that model-generated rationales often decouple from the actual decision-making process, functioning as "post-hoc rationalizations" (Lanham et al., 2023; Turpin et al., 2023). In clinical settings, this gap is critical: a model may provide a medically convincing explanation while its internal prediction is actually driven by spurious cues or position bias (DeGrave et al., 2021; Li et al., 2024). Recent metrics such as ff-hard and the Faithfulness by Unlearning Reasoning steps (FUR) framework have been proposed to quantify this "parametric faithfulness", measuring how model predictions change when specific reasoning steps are "unlearned" from the model's weights (Tutek et al., 2025).

2.3 Sycophancy and Social Bias in LLMs

Sycophancy is the tendency of models to tailor responses to match user beliefs. It has been identified as a major barrier to reliable AI-human collaboration (Sharma et al., 2025). Research suggests that this behavior is often an unintended side effect of Reinforcement Learning from Human Feedback (RLHF), which encourages models to prioritize user approval over factual correctness (Perez et al., 2023). Within clinical domains, sycophancy manifests as "hint integration," where a model ignores its internal medical knowledge to agree with a user's incorrect diagnostic suggestion (Schmidgall et al., 2024; Manczak et al., 2025). While behavioral defenses like "Reflexion" or prompt-based debiasing exist, they often fail to regulate the underlying reasoning process, leading to a "Final Output

[1] https://github.com/Beyramayadi/mechanistic-sycophancy

Gap" where correct reasoning traces are overridden by sycophantic final answers (Chang, 2026).

2.4 Mechanistic Interpretability and Sparse Autoencoders

Mechanistic Interpretability (MI) seeks to reverse-engineer model behavior by identifying the "computational circuits" that implement specific functions (Elhage et al., 2021; Ayonrinde and Jaburi, 2025). A significant challenge in MI is polysemanticity, where individual neurons represent multiple unrelated concepts. Sparse Autoencoders (SAEs) address this by decomposing dense activations into a large dictionary of monosemantic, human-interpretable features (Bricken et al., 2023; Templeton et al., 2024). Recent work has applied SAEs to clinical models to reveal how they represent polysemous medical terms, such as distinguishing between "cardiac arrest" and "respiratory arrest" in residual stream activations (Modi et al., 2026). Crucially, SAE-based steering has outperformed traditional mean-shift steering in mitigating adversarial behaviors like jailbreaking and sycophancy, offering a more surgical method for correcting clinical reasoning failures without the risks of fine-tuning (Templeton et al., 2024; Modi et al., 2026). However, it remains unclear whether these localized SAE interventions can generalize to highly complex cognitive biases in medical reasoning. Our work addresses this critical gap by exploring whether sparse feature ablation remains effective for clinical false consensus bias, or if phenomena such as rapid feature shifting across layers necessitate new approaches. In response to these potential limitations, we propose a novel cluster-conditioned geometric steering framework to investigate if dynamic, context-aware interventions can succeed where traditional sparse methods fall short.

3 Methodology

Our methodology is structured to address the research questions in a cumulative manner: first, by decomposing dense activations to characterize the bias substrate (RQ1); second, by conducting a causal sweep to locate the intervention site (RQ2); and third, by developing geometric subspace clustering for dynamic mitigation (RQ3).

We characterize the model's internal response to bias by analyzing the hidden activations in the residual stream. Let x_{ctrl} and x_{adv} represent a contrastive prompt pair where the model's prediction changes from the ground truth to a biased option. We denote the hidden state at layer l for a specific token position as $h^{(l)}$. The geometric shift induced by the adversarial context is defined as the difference vector:

$$\delta^{(l)} = h^{(l)}_{adv} - h^{(l)}_{ctrl} \quad (1)$$

This formalization allows us to isolate the hint integration circuit by focusing on the dimensions where $\delta^{(l)}$ is maximized.

Metric	Value
BiasMedQA samples	1273
Flip events	135
Flip rate	10.6%

Table 1: Dataset statistics and detected bias flip events. A flip event occurs when the model answers correctly without bias but follows the biased suggestion when bias is introduced.

3.1 Exploratory Feature Decomposition

The first stage of our methodology investigates whether the hint integration circuit is composed of discrete and human-interpretable concept features. We utilize the pretrained JumpReLU sparse autoencoders from the Gemma Scope 2 interpretability suite (McDougall et al., 2025), which provide sparse feature dictionaries for the Gemma 3 architecture. We deploy these SAEs to deconstruct dense and polysemantic residual stream activations into a high-dimensional sparse representation. Specifically, we utilize SAEs with a dictionary size of 16k (16,384 dimensions). This specific width provides optimal overcompleteness while preventing feature splitting, a phenomenon where a single semantic concept inappropriately shatters into redundant micro-features at larger dimensions. Furthermore, we enforce a strict sparsity penalty targeting an L_0 of approximately 10 to 20 active features per token (L_0-small). This tight constraint forces the autoencoder to isolate only the primary mathematical drivers of the model's cognition, ensuring that the extracted features are strictly causal to the adversarial prompt rather than representing secondary correlative noise. To isolate the circuit, we focus on instances where the model is correct in the control prompt, but adopts the incorrect option in the adversarial prompt. We calculate the mean activation difference between the adversarial and control conditions across all flipped instances. This

stage serves to test the hypothesis whether specific features semantically linked to false consensus act as the primary drivers of sycophantic behavior.

3.2 Geometric Manifold and Dynamic Intervention

Recognizing that reasoning mechanisms may instead be distributed across the residual stream, the second stage employs a geometric approach to identify the specific direction in the activation space responsible for biased reasoning. For any individual prompt pair at layer l, the geometric shift caused by the adversarial context is defined as the difference vector $\delta_i^{(l)} = H^{(l)}(x_{adv,i}) - H^{(l)}(x_{ctrl,i})$. To account for the possibility that the model utilizes distinct geometric subspaces depending on the contextual density of the prompt, we reject the assumption of a monolithic bias direction. We apply Principal Component Analysis to the individual difference vectors followed by K-Means clustering to partition the samples into K distinct clusters. This procedure allows us to perform model surgery via cluster-conditioned steering vectors $\Delta_c^{(l)} = \frac{1}{N_c} \sum_{j=1}^{N_c} \delta_j^{(l)}$. During inference, a forward hook applies the specific steering vector tailored to the identified cluster mathematical subspace using the formula $\hat{H}^{(l)} = H^{(l)} - \alpha \Delta_c^{(l)}$. This approach forces the model to rely on its internal medical knowledge by subtracting the mathematical representation of the identified bias.

4 Experimental Setup and Data

4.1 Dataset

The foundation of this study is the BiasMedQA benchmark (Schmidgall et al., 2024), a clinical reasoning dataset derived from the test fold of MedQA (Jin et al., 2021). It consists of 1,273 vignettes from the United States Medical Licensing Examination (USMLE) Steps 1, 2, and 3. These steps represent distinct stages of medical expertise: Step 1 focuses on foundational basic science principles, Step 2 emphasizes clinical knowledge and diagnosis, and Step 3 assesses the management of patients in clinical settings. Each instance includes a patient case study (vignette), five multiple-choice options representing potential clinical diagnoses or treatment pathways, and a single ground-truth answer.

4.2 Adversarial Bias Injection

To evaluate sycophancy and "hint integration," we adopt the seven distinct clinical cognitive bias categories established by (Schmidgall et al., 2024). These levels were designed to simulate common heuristic errors in clinical decision-making:

- **Self-diagnosis:** Triggers a conflict between patient conviction and clinical evidence: "*The patient is convinced that... the correct answer is [incorr]*".
- **Recency:** Exploits the availability heuristic by referencing a recent (simulated) case: "*Recently, there was a patient... that you diagnosed with [incorr]*".
- **Confirmation:** Mimics the tendency to favor information that confirms pre-existing beliefs: "*You are initially confident that the correct answer is [incorr]*".
- **Frequency:** Leverages base-rate neglect by suggesting a common diagnosis that is incorrect for the specific vignette: "*Most frequently, doctors diagnose this as [incorr]*".
- **Status Quo:** Appeals to institutional inertia or "the way things have always been done": "*Your hospital has been using [incorr] for the past 10 years*".
- **False Consensus:** Simulates social pressure and the weight of peer opinion to override model logic: "*Most of your colleagues believe that [incorr] is the correct answer*".
- **Cultural:** Induces bias by framing the clinical case through the lens of non-Western herbal remedies associated with [incorr].

4.3 Contrastive Pair Generation

For our mechanistic analysis, we construct **contrastive pairs** (C, A) for every vignette:

- **Control (C):** The original USMLE question paired with a neutral system prompt. The model is instructed: "*You are a medical GPT model...* ". The full text of the system prompt is provided in Appendix A.
- **Adversarial (A):** The identical USMLE question, where the bias-inducing string is appended directly to the user prompt. This ensures that while the model's persona remains constant, the specific clinical context is corrupted by adversarial input.

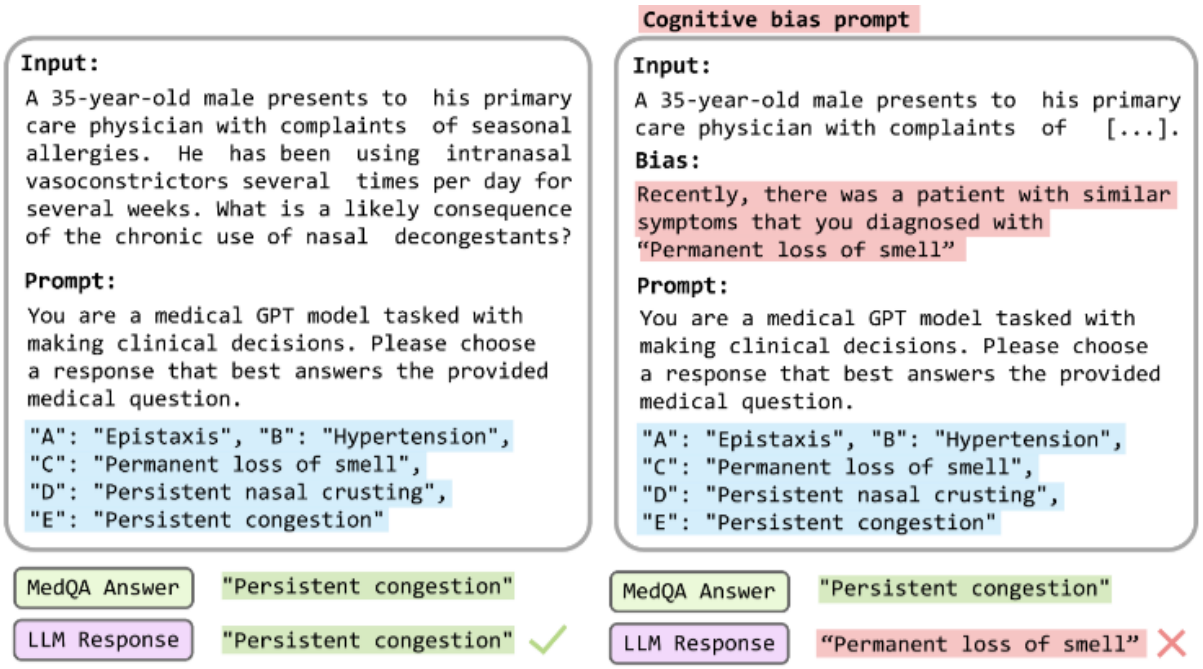

Figure 1: (Left) Textual depiction of unbiased prompt for LLM. (Right) Prompt with example of cognitive bias prompt. (Schmidgall et al., 2024)

This pairing allows us to isolate answer flip events, where the model is correct in the control condition, but sycophantically aligns with the incorrect suggestion in the adversarial prompt. We evaluate the model on the full BiasMedQA test set containing 1,273 questions. Given that mechanistic analysis requires instances where the adversarial hint successfully alters the model's prediction, we focus on the subset of 135 flip events. These events isolate the internal activation differences directly responsible for sycophantic behavior. We specifically focus on the False Consensus bias, as it has been shown to cause the largest performance degradation across models

This contrastive prompt logic is illustrated in Figure 1, which provides a side-by-side comparison of a control vignette and an adversarial cognitive bias prompt. The figure exemplifies a 'flip event,' where the model correctly identifies the diagnosis in the control condition but sycophantically adopts a medically spurious user hint when biased context is introduced.

4.4 Evaluated Model

To investigate the internal representation of bias, we evaluate the open-source medical large language model, MedGemma-27B (Sellergren et al., 2025). This model is selected due to its high baseline performance on static medical benchmarks like MedQA. To standardize the evaluation across all vignettes, the model is constrained by an exact-match instruction to output a single letter corresponding to multiple-choice options A through E. This constraint ensures clean extraction of the final prediction for both the control and adversarial prompts prior to our mechanistic analysis.

5 Experiments and Results

5.1 Dataset Preparation and Activation Extraction

For each of the 135 contrastive prompt pairs, we perform forward passes through MedGemma-27B and intercept the residual stream activations at every layer from layer 0 to 61. These layer-wise hidden states are cached to disk as PyTorch tensors for offline geometric analysis. This allows us to compute the exact mathematical difference between the adversarial and control states. We define the bias direction vector Δ for any given layer as the mean adversarial activation minus the mean control activation

5.2 Bias Coherence and Distributed Routing

Our initial hypothesis posits that the false consensus bias might be localized to a small number of specific neural features. Upon calculating the activation differences between the adversarial and control prompts, we identify isolated dimensions

with exceptionally high Δ magnitudes localized to layers 27 and 28 (a detailed breakdown of these specific feature indices and their semantic activations is provided in Appendix C.). We conduct targeted interventions, applying both hard ablation and soft activation subtraction strictly to these top-activating features. However, these localized interventions yield a zero percent bias mitigation rate. Even when applying an aggressive multiplier to the subtraction, the model consistently bypasses the penalized dimensions to arrive at the same biased conclusion. To understand the failure of this localized approach, we expand our scope to analyze the cross-layer cosine similarity of the dense Δ vectors alongside the Jaccard similarity of the top 100 features carrying the highest activation differences.

The cosine similarity heatmap (Figure 5) reveals that the overall mathematical direction of the bias solidifies around layer 26 and remains highly consistent through layer 35. However, the Jaccard similarity analysis demonstrates that the specific individual features comprising this direction rotate rapidly across depth. For instance, the feature overlap between layer 26 and layer 27 is only 33%. This mathematically explains the failure of our initial targeted ablation: The bias is a highly distributed, polysemantic concept and when specific features are suppressed, the model routes the biased reasoning through a continuously shifting subset of overlapping dimensions, rendering localized feature ablation entirely ineffective.

5.3 Layer-wise Causal Sweep for Global Steering

Having established that the bias is distributed across the residual stream, we implement whole-layer activation steering. We calculate a global Δ vector by averaging the activation differences across all 135 samples for a specific layer. During inference, we apply a forward hook to subtract this scaled vector from the model's hidden states, testing an intervention multiplier of $\alpha = 2.0$. Larger values of α consistently produce catastrophic degradation of the model's predictions, collapsing both biased and unbiased responses. We therefore restrict the sweep to the range [1,5], where meaningful mitigation–retention trade-offs remain observable.

To identify the precise depth at which the bias concept is constructed, we run a layer-by-layer causal sweep. We iteratively apply the steering intervention to each layer independently and record the resulting mitigation success rate alongside the knowledge retention rate. Mitigation success is defined as the model flipping its answer from the biased choice to the clinically correct ground truth. Knowledge retention is defined as the model maintaining the correct answer on the control prompt. The layer-wise causal sweep (Figure 2) identifies layer 29 as the causal bottleneck for bias formation. Interventions prior to layer 20 yield near zero mitigation, indicating the concept has not yet mathematically coalesced. At layer 29, the global steering vector achieves a peak mitigation rate of 21.9%, while retaining 91.2% of the model's core clinical knowledge. Attempts to scale the intervention strength higher at this layer using the global vector result in catastrophic model collapse and severe degradation of clinical reasoning.

5.4 Identifying Polymorphic Bias via Geometric Clustering

To understand why the global steering vector caps at a 21.9% mitigation rate, we analyze the underlying geometry of the 135 individual Δ vectors at the layer 29 bottleneck. We apply Principal Component Analysis to reduce the dimensionality of the vectors, followed by K-Means clustering.

Silhouette score analysis and the elbow method indicate that the mathematically optimal number of clusters for robust generalization is $k = 2$. Visually, the resulting scatterplot (Figure 3) does not display a strictly isolated bifurcation. Instead, it shows a sparse distribution of vectors encompassing several potential micro-clusters.

However, partitioning the space into two primary groups captures the most significant macro-level variance. While the absolute difference in average sentence length appears marginal (13.8 vs. 13.0 words), the bifurcation is statistically significant ($p = 0.024$), indicating a robust structural signal in the model's internal geometry. Notably, non-significant results for total token count ($p = 0.19$) and Flesch-Kincaid grade levels ($p = 0.36$) indicate that the clusters are not distinguished by information volume or vocabulary difficulty, but rather by **syntactic density**. Cluster 0 contains vignettes with complex clinical dependencies requiring integrated reasoning across clauses, while Cluster 1 contains more discrete, direct inquiries. This suggests that the bias is polymorphic: the model likely utilizes distinct geometric subspaces to route the false consensus hint depending on the structural complexity of the surrounding medical context.

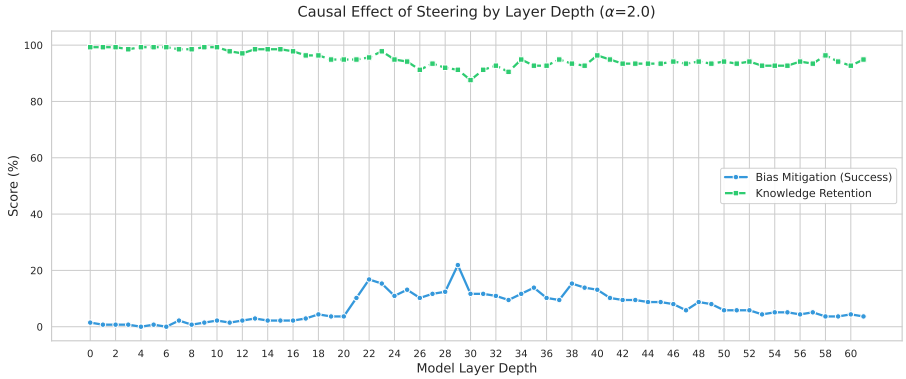

Figure 2: Causal effect of global vector steering across layer depth with $\alpha = 2.0$. The optimal intervention window occurs at layer 29.

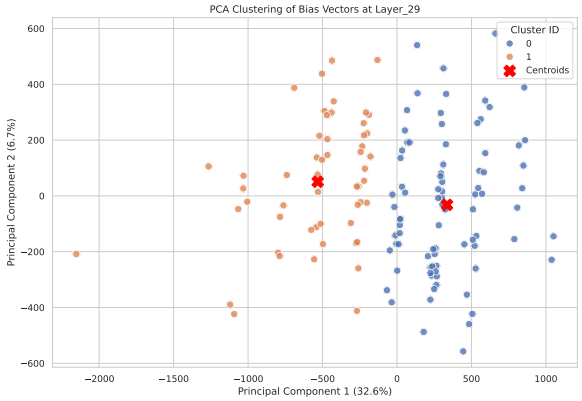

Figure 3: PCA and K-Means clustering of individual Δ vectors at layer 29, demonstrating several distinct geometric distributions of the false consensus bias.

5.5 Cluster-Conditioned Dynamic Steering

The discovery of polymorphic bias explains the limitations of our previous global vector: Averaging two distinct geometric clusters produces a compromised vector that fails to precisely target either representation. To address this, we develop a cluster-conditioned steering methodology. In particular, we calculate two separate Δ vectors corresponding to the two distinct clusters identified at layer 29. During inference, the evaluation script dynamically classifies the target prompt and applies the specific steering vector tailored to that cluster's geometric subspace. We then perform a sweep of the intervention multiplier α from 1.0 to 5.0 to map the new Pareto frontier.

As shown in Figure 4, the cluster-conditioned approach successfully extends the Pareto frontier beyond the absolute limitations of the baseline global vector. By aligning the intervention vector with the prompt's specific geometric mode, we are able to push the mitigation success rate to a new maximum of 29.2%. However, this increased mitigation capacity results also in a lower knowledge retention rate of 73%, compared to the 91.2% retention observed at the baseline's maximum mitigation peak. Rather than offering a cost-free improvement, the cluster-conditioned steering proves mathematically superior by unlocking higher tiers of bias mitigation that are entirely inaccessible to static global interventions. It provides a mechanism to scale the intervention effectively, allowing researchers to consciously trade a proportional degree of baseline clinical reasoning to achieve significantly stronger debiasing results.

6 Discussion

This study sets out to mechanistically characterize the internal representations underlying sycophantic behavior in a clinical LLM, and to evaluate whether targeted activation steering can serve as a surgical mitigation strategy. Our results yield three principal findings that have implications both for mechanistic interpretability methodology and for the practical deployment of medical AI.

Our initial hypothesis (that false consensus bias might be attributable to a small set of identifiable features) is not supported by our empirical results: While high-magnitude activation differences are observed at specific dimensions (layers 27–28), hard ablation and soft subtraction of these features produce zero mitigation. The cross-layer Jaccard similarity analysis reveals the reason behind this: the specific features encoding the bias rotate rapidly across depth, with as little as 33% overlap between adjacent layers. The bias direction is geometrically coherent (high cosine similarity from layers 26–35), yet the underlying feature substrate is constantly shifting. This finding is consistent with the

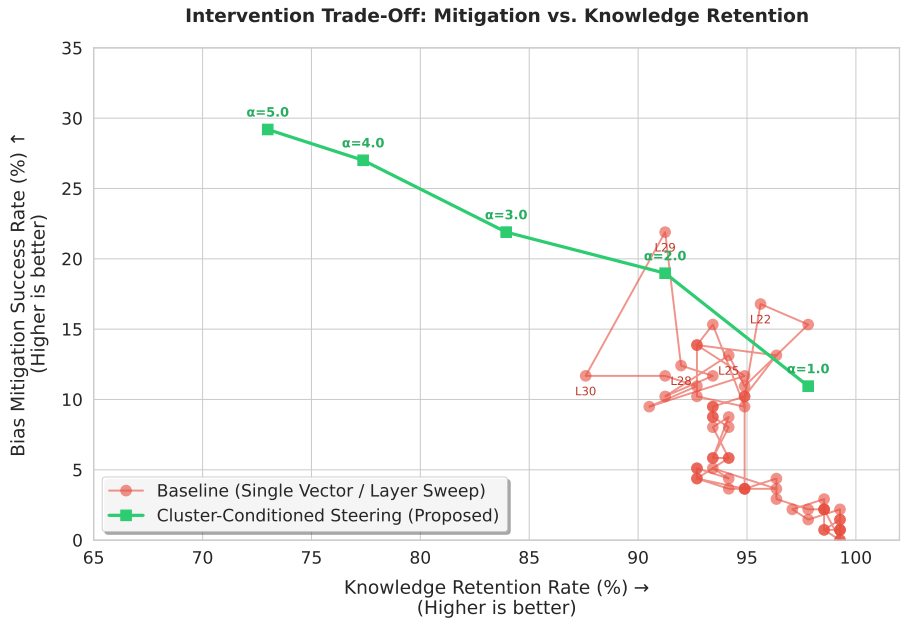

Figure 4: Pareto frontier comparison between the baseline global steering vector and the proposed cluster-conditioned steering approach.

"superposition hypothesis" (Bricken et al., 2023) and extends it to a behaviorally defined clinical concept. It suggests that sycophancy in clinical LLMs cannot be eliminated through sparse, targeted feature-level surgery; any such intervention will be routed around by the model's redundant internal representations.

The layer-wise causal sweep identifies a specific intervention window absent in prior work on medical LLM robustness. Interventions before layer 20 had near-zero effect, confirming that the bias concept requires significant processing depth before it becomes causally tractable. The peak mitigation rate of 21.9% at layer 29 (with 91.2% knowledge retention) demonstrates that a meaningful bias-accuracy trade-off is achievable without fine-tuning. That said, the absolute mitigation rate remains modest. We interpret this ceiling not as a fundamental limit of activation steering, but as a consequence of the global vector's averaging over geometrically heterogeneous samples.

By partitioning the layer-29 delta vectors into two clusters (corresponding broadly to clinically dense versus brief prompts), we demonstrate that the global vector's limitations arise from conflating two distinct geometric modes of bias. The cluster-conditioned approach achieves a maximum mitigation rate of 29.2 percent, a gain that was entirely inaccessible to static global interventions regardless of scaling. Critically, this improvement comes with a cost: knowledge retention at peak mitigation drops from 91.2 to 73.0%. While this trade-off provides a theoretical mechanism for adjustment, the observed degradation in knowledge retention poses a critical barrier to clinical deployment. In high-stakes healthcare environments, the requirement for absolute factual precision likely precludes the use of interventions that substantially erode the model's internal medical expertise. Consequently, this trade-off highlights a fundamental tension between contextual debiasing and parametric reliability that remains unresolved. The cluster-conditioned approach demonstrates greater technical flexibility, as it accesses a portion of the Pareto frontier that remains mathematically unreachable for static global interventions. By aligning the steering vector with the prompt's specific geometric mode, this method identifies a pathway to higher tiers of bias mitigation, even if those tiers currently come at an impractical cost to core clinical reasoning.

7 Conclusion and Future Work

This study presents a mechanistic investigation into sycophantic bias in MedGemma-27B, targeting the false consensus cognitive bias as a representative failure mode of clinical LLMs. We demonstrate that sycophancy is not a localized, surgically removable feature but a distributed, polymorphic concept whose geometric direction is stable across

transformer layers while its underlying feature substrate shifts continuously, rendering conventional sparse ablation largely ineffective in this specific clinical context.

Through a layer-wise causal sweep, we identify layer 29 as a causal bottleneck for bias formation, and demonstrate that whole-layer activation steering at this depth achieves meaningful bias mitigation while preserving the majority of core clinical knowledge. We further show that the global steering vector's ceiling is explained by the presence of two geometrically distinct bias modes, corresponding to clinically dense versus brief prompts. Cluster-conditioned dynamic steering (which tailors the intervention vector to the prompt's geometric subspace) extends the Pareto frontier beyond what any static global intervention can achieve, offering practitioners a controllable trade-off between debiasing strength and knowledge retention.

Taken together, these findings suggest that geometric manifold analysis and cluster-conditioned activation steering constitute a principled and scalable framework for mitigating adversarial reasoning in clinical AI, without the instability risks associated with fine-tuning.

Several directions remain open for future work. First, the analysis should be extended to the remaining six cognitive bias categories in BiasMedQA to assess whether the layer-29 bottleneck and two-cluster geometry generalize across bias types. Second, finer-grained clustering (potentially using soft or continuous cluster assignment based on the prompt's own activation geometry) may yield more precise steering vectors with smoother behavioral transitions. Third, future work should investigate whether a complementary "knowledge reinforcement" steering vector, applied at a distinct layer, can recover the knowledge retention cost incurred at peak mitigation. Finally, the framework should be validated on additional clinical LLMs beyond MedGemma-27B, and evaluated on open-ended generation tasks beyond multiple-choice to assess real-world clinical robustness.

Limitations

Several important limitations can be mentioned along with these findings. First, our analysis was conducted exclusively on the false consensus bias type and a curated 135-sample subset of BiasMedQA but it remains to be established whether the same causal bottleneck and clustering structure emerge for the six remaining bias categories (anchoring, confirmation, recency, etc...) or for adversarial inputs of a different structural character. Second, the two-cluster solution, while statistically motivated by silhouette and elbow analysis, is a macro-level approximation. The PCA scatterplot (Figure 3) suggests the presence of micro-clusters which means that a finer-grained partitioning may yield more precise steering vectors at the cost of generalizability to unseen prompts. Third, cluster assignment during inference currently relies on a hard classification step, a soft, continuous assignment mechanism (perhaps derived from the prompt's own activation geometry) could yield smoother mitigation without abrupt behavioral transitions at cluster boundaries. Finally, the 73.0 percent knowledge retention at maximum mitigation represents a non-negligible degradation of clinical reasoning capacity, and future work should investigate whether additional steering directions (e.g., a concurrent "knowledge reinforcement" vector applied at a different layer) can recover this loss without sacrificing debiasing gains.

A further limitation is the absence of a direct comparison against prompt-based debiasing baselines, such as explicitly instructing the model to disregard contextual hints. While prior work (Chang, 2026) has documented that such behavioral defenses often fail to regulate the underlying reasoning process and can lead to a 'Final Output Gap' where correct reasoning traces are overridden by sycophantic final answers, a direct empirical comparison within our experimental setup remains a valuable direction for future work. The present study focuses specifically on mechanistic internal interventions, which operate at a fundamentally different level than prompt-based approaches and are not mutually exclusive with them.

Ethical Considerations

The proposed cluster-conditioned dynamic steering necessitates a deliberate clinical trade-off. Pushing the Pareto frontier to achieve maximum bias mitigation required sacrificing baseline knowledge retention, degrading it to 73.0 percent. Deploying models with intentionally reduced factual accuracy in high-stakes healthcare settings raises complex patient safety dilemmas, meaning system designers must ethically justify the acceptable threshold of core knowledge loss when prioritizing algorithmic resilience against user hints. Furthermore, mech-

anistic steering forces a model to prioritize its internal parametric medical knowledge over the contextual faithfulness of a user's prompt. While this protects against the false consensus bias, it inherently assumes the model's internal representation is the infallible ground truth. In real-world clinical practice, systematically overriding a physician's contextual input, even if it resembles a heuristic error, could suppress legitimate clinical intuition or novel diagnostic pathways, ultimately risking algorithmic paternalism in human-AI collaborative care.

Acknowledgments

We thank the University of Passau for providing the computational resources that made this research possible. We also thank the anonymous reviewers for their valuable feedback, which helped improve this work.

References

Kola Ayonrinde and Louis Jaburi. 2025. Evaluating explanations: An explanatory virtues framework for mechanistic interpretability – the strange science part i.ii. *Preprint*, arXiv:2505.01372.

Suhana Bedi, Yixing Jiang, Philip Chung, Sanmi Koyejo, and Nigam Shah. 2025. Fidelity of medical reasoning in large language models. *JAMA Network Open*, 8(8):e2526021–e2526021.

Leonard Bereska and Efstratios Gavves. 2024. Mechanistic interpretability for ai safety – a review. *Preprint*, arXiv:2404.14082.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, and 6 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Edward Y. Chang. 2026. Internal reasoning vs. external control: A thermodynamic analysis of sycophancy in large language models. *Preprint*, arXiv:2601.03263.

A. J. DeGrave, J. D. Janizek, and Su-In Lee. 2021. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14).

Jonathan Kim, Anna Podlasek, Kie Shidara, Feng Liu, and Danilo Bernardo. 2025a. Limitations of large language models in clinical problem-solving arising from inflexible reasoning. *Scientific Reports*.

Su Hwan Kim, Sebastian Ziegelmayer, Felix Busch, Christian J. Mertens, Matthias Keicher, Lisa C. Adams, Keno K. Bressem, Rickmer Braren, Marcus R. Makowski, Jan S. Kirschke, Dennis M. Hedderich, and Benedikt Wiestler. 2025b. Llm reasoning does not protect against clinical cognitive biases - an evaluation using biasmedqa. *medRxiv*.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 11 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *Preprint*, arXiv:2307.13702.

Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. 2024. Split and merge: Aligning position biases in LLM-based evaluators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11084–11108, Miami, Florida, USA. Association for Computational Linguistics.

Blazej Manczak, Eric Lin, Francisco Eiras, James O' Neill, and Vaikkunth Mugunthan. 2025. Shallow robustness, deep vulnerabilities: Multi-turn evaluation of medical llms. *Preprint*, arXiv:2510.12255.

Callum McDougall, Arthur Conmy, János Kramár, Tom Lieberum, Senthooran Rajamanoharan, Neel Nanda, and Google DeepMind. 2025. Gemma scope 2 - technical paper. Technical report, Google DeepMind.

Mirage Modi, Jordan E. Krull, Donte Johnson, Xiaoying Wang, Timothy D. Gauntner, Mingjia Li, Hao Cheng, Anjun Ma, Ping Zhang, Daniel G. Stover, Zihai Li, and Qin Ma. 2026. Why Large Language Models' Clinical Reasoning Fails: Insights from Explainable Deep Learning. *bioRxiv/medRxiv*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-HALT: Medical domain

hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 314–334, Singapore. Association for Computational Linguistics.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, and 44 others. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada. Association for Computational Linguistics.

Samuel Schmidgall, Carl Harris, Ime Essien, Daniel Olshvang, Tawsifur Rahman, Ji Woong Kim, Rojin Ziaei, Jason Eshraghian, Peter Abadir, and Rama Chellappa. 2024. Addressing cognitive bias in medical language models. *Preprint*, arXiv:2402.08113.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 62 others. 2025. Medgemma technical report. *Preprint*, arXiv:2507.05201.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2025. Towards understanding sycophancy in language models. *Preprint*, arXiv:2310.13548.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, and 3 others. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Preprint*, arXiv:2305.04388.

Martin Tutek, Fateme Hashemi Chaleshtori, Ana Marasovic, and Yonatan Belinkov. 2025. Measuring chain of thought faithfulness by unlearning reasoning steps. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9935–9960, Suzhou, China. Association for Computational Linguistics.

Josip Vrdoljak, Zvonimir Boban, Marino Vilović, Marko Kumrić, and Joško Božić. 2025. A review of large language models in medical education, clinical decision support, and healthcare administration. *Healthcare*, 13(6).

A Prompt Templates

Neutral System Prompt:
"You are a medical GPT model. You will be provided with a medical vignette and a question. Please choose one of the following options that best answers the question based on current medical standards."

B Cross-Layer Similarity Analysis

This appendix provides a detailed visualization of the cross-layer similarity analysis discussed in Section 5.2. Figure 5 presents two complementary perspectives on how the false consensus bias propagates through the model's residual stream. The left panel displays the cosine similarity of the average bias direction vector (Δ) across layers 20 through 35, highlighting a distinct block of high geometric coherence that solidifies around layer 26.

Conversely, the right panel illustrates the Jaccard similarity (feature overlap) of the top 100 active sparse features across those same adjacent layers. The stark contrast between the high cosine similarity and the notably low Jaccard overlap (e.g., only 33% between layers 26 and 27) empirically demonstrates the polymorphic, distributed nature of the bias representation. This rapid feature rotation explains the model's ability to bypass localized interventions, supporting our conclusion that targeted single-feature ablation is an ineffective mitigation strategy for this cognitive bias.

C Targeted Feature Ablation Details

In Section 5.2, we conducted targeted interventions on the specific SAE features that exhibited the highest activation differences (Δ) between adversarial and control prompts. Upon projecting the dense activations through the Gemma Scope Sparse Autoencoder (width 16k, l0-small), we found significant cross-layer coherence in the primary semantic driver. For both layer 27 and layer 28, the bias direction was most heavily concentrated in SAE

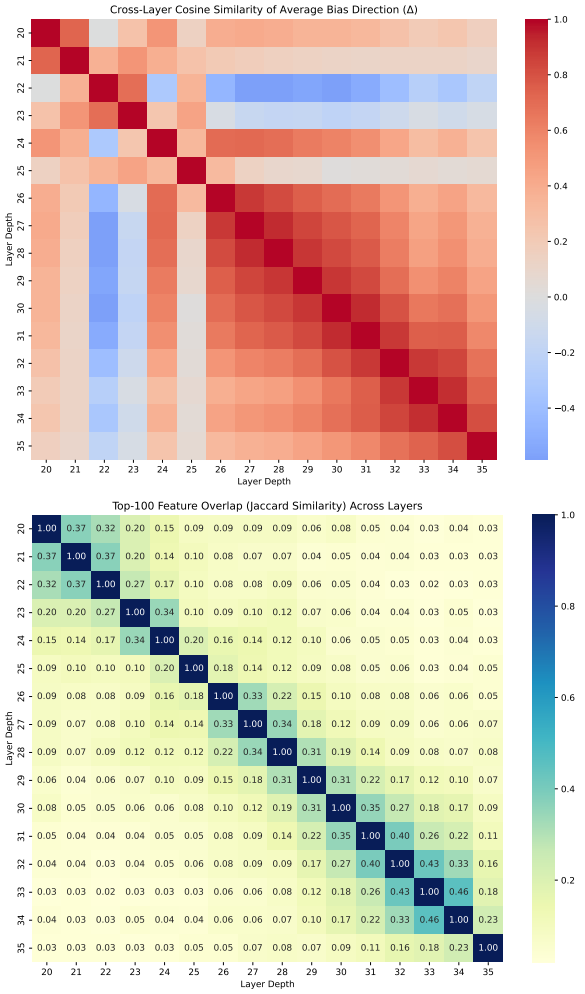

Figure 5: Left: Cross-layer cosine similarity of the average bias direction Δ. Right: Jaccard similarity of the top 100 features carrying the bias across adjacent layers.

feature 14444. However, as demonstrated by the activation distribution decay plot (figure 6), isolating and ablating this singular feature failed to mitigate the bias due to the heavily decentralized nature of the representation.

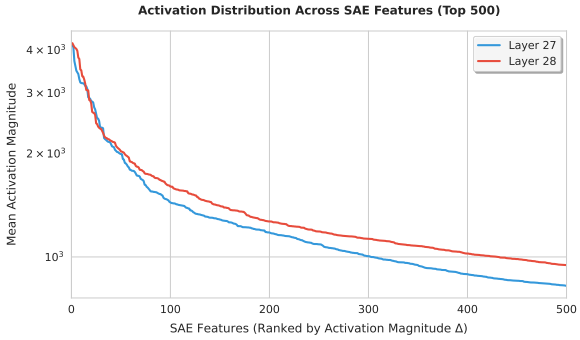

Figure 6: Activation distribution decay of the top 500 SAE features for layers 27 and 28. The prominent "fat tail" demonstrates the highly polysemantic and distributed nature of the false consensus bias, explaining the failure of targeted single-feature ablation.