
HalluShield: A Mechanistic Approach to Hallucination Resistant Models

Tanzeel Shaikh
Gac Group

Naiyarah Hussain
CodersHQ, UAE

Samuel Nihoul
Harmonia.cloud

Sonal Joshi
Independent work

Lucia de la Torre
Columbia University

With
Apart Research and GoodFire AI

Abstract

Our project tackles the critical problem of hallucinations in large language models (LLMs) used in healthcare settings, where inaccurate information can have serious consequences. We developed a proof-of-concept system that classifies LLM-generated responses as either factual or hallucinated. Our approach leverages sparse autoencoders (GoodFire’s Ember) trained on neural activations from Meta Llama 3. These autoencoders identify monosemantic features that serve as strong indicators of hallucination patterns. By feeding these extracted features into tree-based classification models (XGBoost), we achieved an impressive F1 score of 89% on our test dataset. This machine learning approach offers several advantages over traditional methods and LLM as a judge. First, it can be specifically trained on in-domain datasets (eg: medical) for domain-specific hallucination detection. Second, the model is interpretable, showing which activation patterns correlate with hallucinations and acts as a post-processing layer applied to LLM output.

Keywords: Mechanistic Interpretability, Monosemanticity, sparsity, Hallucination, Sparse Autoencoder, Binary classification, Supervised learning, Sparsity, Activation strength

1. Problem Statement

Large Language Models (LLMs) are increasingly used in healthcare, yet their tendency to hallucinate poses significant safety risks that current approaches cannot adequately address. Prior work shows near-ubiquitous hallucinations in medical LLMs, with risks amplified by diagnostic overreliance (Clinicians’ Guide to LLMs, 2025). The University of Massachusetts Amherst and Mendel AI study (2024) found 327 medical event inconsistencies in just 50 GPT-4o summaries, highlighting critical areas like symptoms and medication instructions.

Current approaches face two main issues:

- **Symptom Treatment vs. Cause Elimination:** Retrieval-augmented generation (RAG) fails to address internal mechanisms generating false information.
- **Lack of Targeted Intervention:** Fine-tuning and prompt engineering lack visibility into how hallucinations form at the neural level.

We propose using Goodfire’s sparse autoencoder technology to perform targeted interventions at the neural level. This involves applying sparse autoencoder monosemanticity to isolate hallucination-prone features using the MedHallu dataset,

which contains 10,000 PubMedQA-derived medical Q&A pairs annotated for factual vs. hallucinated responses (MedHallu, 2024).

a. Background and Motivation

In an interview with a Johns Hopkins professor specializing in AI in medicine, he emphasized that medical hallucinations pose significant risks, such as incorrect responses to medical questions or incorrect emergency room summaries. Studies show 22% of Llama-3 medical summaries contained incorrect drug instructions (UMass/Mendel, 2024), and MedHallu (2025) highlights "hard" hallucinations that are 38% harder to detect without domain knowledge. Goodfire's sparse autoencoders, originally successful in genomic pattern detection, enable precise modification through the Ember API, addressing hallucination root causes while preserving legitimate medical knowledge. Unlike prompt engineering, Goodfire's tools allow surgical model edits without retraining (Lightspeed Ventures, 2024). Real-time feature monitoring further ensures clinical scalability and robustness.

b. Threat Model and Safety Implications

The proposed method addresses two key threats:

1. Medical hallucinations directly endanger patients through incorrect diagnoses or harmful treatment recommendations and suggestions.
2. Hallucinations in AI-based summarizing of patient reports during high-stakes medical scenarios like emergency room / intensive care unit reporting.

For our proposed work, we use the MedHallu dataset of medical Q&A pairs to evaluate and address hallucination-prone outputs in LLM responses (see Figure 1).

Question string · lengths	Knowledge sequence · lengths	Ground Truth string · lengths	Difficulty Level string · classes	Hallucinated Answer string · lengths	Category of Hallucination string · classes
					
Are group 2 innate lymphoid cells (ILC2s ...	["Chronic rhinosinusitis (CRS) is a heterogeneous...	As ILC2s are elevated in patients with...	easy	Group 2 innate lymphoid cells...	Misinterpretation of #Question#
Does vagus nerve contribute to the...	["Phosphatidylethanolamin...	Neuronal signals via the hepatic vagus...	medium	Capsaicin-induced disruption of hepati...	Misinterpretation of #Question#
Does psammaplin A induce Sirtuin 1-dependent...	["Psammaplin A (PsA) is a natural product...	PsA significantly inhibited MCF-7/adx...	medium	Psammaplin A induces Sirtuin 1-independen...	Misinterpretation of #Question#
Is methylation of the FGFR2 gene associated...	["This study examined links between DNA...	We identified a novel biologically plausibl...	hard	Methylation of the FGFR2 gene is...	Misinterpretation of #Question#
Do tumor-infiltrating immune cell profiles an...	["Tumor microenvironment immunity is associated...	Breast cancer immune cell subpopulation...	hard	The levels of CD20-positive B cells in...	Incomplete Information

Figure 1 – MedHallu Dataset snippet

Our approach offers several safety benefits over previous methods:

1. Prevention over Detection: We modify underlying neural features to prevent hallucinations, unlike just detection-focused methods.
2. Targeted Intervention: Our Goodfire-based approach disrupts hallucination mechanisms at their source, unlike retrieval-augmented generation.
3. Mechanistic Transparency: This transparency builds trust with medical professionals and regulators (Ennab et al., 2022), enabling validation of AI-driven diagnoses and compliance with regulatory requirements (Ennab et al., 2024). The safety

implications extend beyond clinical applications to other high-stakes domains requiring factual reliability.

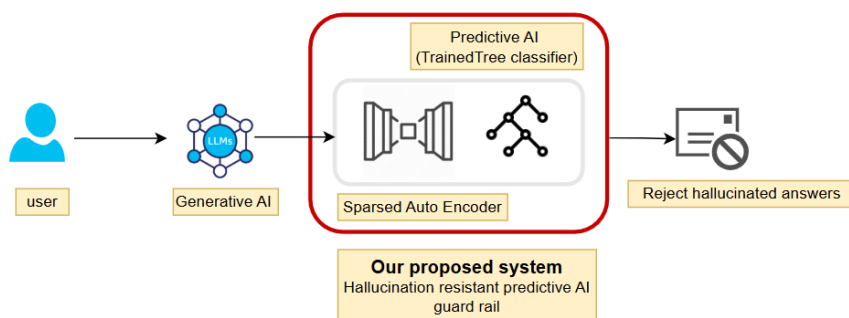
It is important to note that the safety implications extend beyond immediate clinical applications, as our framework could be applied to other high-stakes domains where factual reliability is essential.

2. Methods

a. Approach

- A user query is processed by an LLM to generate a response.
- The response is passed through a **sparse autoencoder**, which extracts the activation strengths of computed features(selected in the training phase of the classifier).
- These activation strengths serve as input features for a **pre-trained tree classifier**, which is trained on them.
- The tree classifier then determines whether the response is **hallucinated or not**.
- This approach relies on training the tree classifier using the **neural activations** obtained from the sparse autoencoder.

Figure 2 - System Architecture



b. Implementation

1. **Dataset Preparation:** We split the Med-Hallu dataset into:

- **Non-hallucinated:** (question, knowledge, ground truth).
- **Hallucinated:** (question, knowledge, hallucinated answer).

2. **Feature Extraction:** Use Contrast API (Ember) to compute distinguishing features between hallucinated and non-hallucinated responses.

3. **Feature Ranking:** Select the top K features from the sparse autoencoder and rank them using Goodfire ReRank based on the below queries:

- **Hallucination query:** "medical misinformation and factual errors in healthcare answers"

- **Non-hallucination query:** "accurate and factual medical information with proper evidence"
- Activation Strengths:** Pass ranked features to Inspect API to compute activation strengths of the above-computed features for the dataset.
 - Labeled Dataset:** Construct a dataset with activation strengths and hallucination labels for supervised learning.
 - Model Training:** Train classifiers (Decision Tree, Random Forest, XGBoost) and select the best model-feature combination.

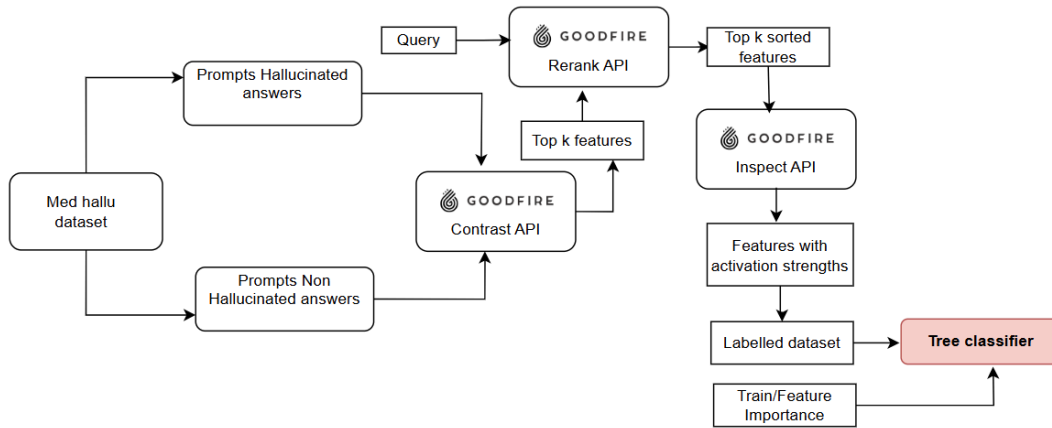


Figure 3 – Training process for Tree classifier

Sample	Class	Discussion of new diagnostic techniques and medical innovations	Periods separating distinct observations in clinical/technical reports	Assistant making potentially incorrect claims about its identity or capabilities
H1	Hallucinated (-1)	0.000	0.000	2.000
H2	Hallucinated (-1)	2.000	1.000	2.000
H3	Hallucinated (-1)	0.000	0.000	2.000
H4	Hallucinated (-1)	0.000	1.000	0.000
H5	Hallucinated (-1)	3.000	1.000	2.000
NH1	Non-Hallucinated (1)	1.000	2.000	1.000

Figure 4 – Training data for Supervised learning

3. Results

a. Analysis and Findings

We evaluated three classifiers—Decision Tree, Random Forest, and XGBoost—for binary classification using consistent hyperparameters. Notably, all models performed best with the same top three feature sets, as illustrated by the XGBoost results (Figure 5):

Feature Count	Features	F1 Score	Accuracy	Precision	Recall	Balanced Accuracy	True Positives	False Positives	True Negatives	False Negatives
0	1 Assistant making potentially incorrect claims about its identity or capabilities	0.814	0.820	0.868	0.820	0.820	50	18	32	0
1	2 Periods separating distinct observations in clinical/technical reports, Assistant making potentially incorrect claims about its identity or capabilities	0.870	0.870	0.874	0.870	0.870	46	9	41	4
2	3 Evidence-based implementation and evaluation strategies, Periods separating distinct observations in clinical/technical reports, Assistant making potentially incorrect claims about its identity or capabilities	0.890	0.890	0.894	0.890	0.890	47	8	42	3

Figure 5 – XGBoost results with 3 different feature combinations

The XGBoost model achieved its best F1 score with three features, a finding further supported by the feature importance analysis in Random Forest, which also identified similar key features.

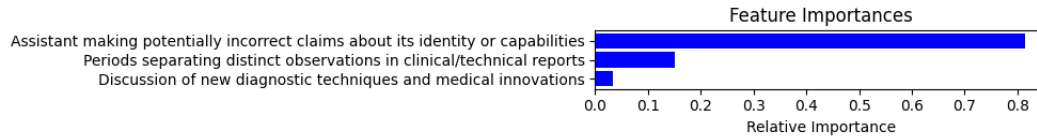


Figure 6 – Feature importance by Random classifier

Similar features were also observed in the decision tree model, where the Gini impurity values for three leaf nodes were approximately 0, 0.17, 0.27, 0.38.

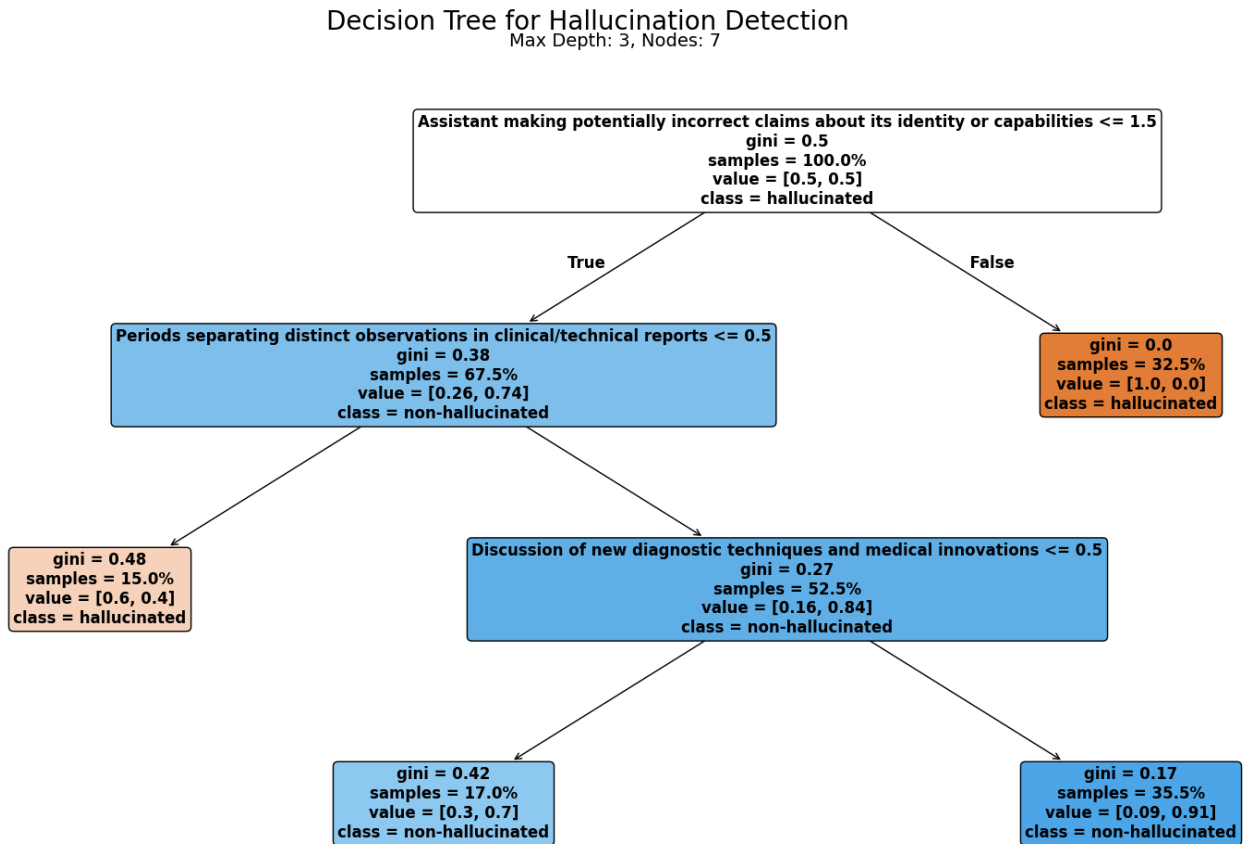


Figure 7 – Decision tree visualization showing gini impurity

b. Impact Assessment

The proposed mechanistic interpretability approach to hallucination mitigation demonstrates significant promise for AI safety in healthcare. Through our initial results, we show that targeted interventions on neural features are correlated with hallucinations. The solution directly addresses the critical challenge of medical hallucinations by:

1. **Pinpointing specific neural features** associated with incorrect identity claims in the Meta Llama 3 8B model
2. **Creating a predictive safety guardrail** that can effectively filter hallucinated outputs
3. Providing a mechanism to steer models away from hallucinations at their source. This approach is particularly valuable for healthcare applications where hallucinations can lead to potentially life-threatening misinformation. The practical implications of our work for healthcare AI safety are substantial:
 - a. Healthcare providers can deploy LLMs with greater confidence that harmful hallucinations will be detected and filtered
 - b. Regulatory bodies gain a more transparent and mechanistically interpretable approach to evaluate AI safety in medical applications
 - c. The method provides a foundation for developing similar safety guardrails in other high-stakes domains

c. Future work

Due to time constraints, we could not evaluate model performance across different hallucination levels (easy, medium, hard). Future work could involve an ablation study using mechanistic interpretability methods. Additionally, we were limited to the Meta-Llama-3.1-8B-Instruct model due to Goodfire API constraints. Expanding this work to other models and tasks like report summarization would be valuable. Notably, general LLMs have been shown to outperform medically fine-tuned models in medical hallucination tasks, offering further avenues for exploration. We can also improve performance further by refining the prompts and queries used.

4. Discussion and Conclusion

By focusing on the mechanistic origins of hallucinations rather than just their symptoms, our approach represents a meaningful step toward safer AI deployment in medicine, where the consequences of misinformation can directly impact patient outcomes and trust in healthcare AI systems. By identifying specific features like incorrect identity claims (importance: 0.8142) and implementing a predictive guardrail with an F1 score of 0.825, we have created a framework that enables safer AI deployment in medicine where misinformation can directly impact patient outcomes. It is important to note these results are with the small data and model we had and can be improved further by scaling. This approach opens new possibilities for interpretable AI safety across high-stakes domains, moving beyond traditional detection methods toward prevention at the source through targeted neural interventions.

5. References

- Clinicians' Guide to LLMs. (2025). A General Framework for Integration into Clinical Practice.
- University of Massachusetts Amherst & Mendel AI. (2024). Analysis of Medical Event Inconsistencies in LLM-Generated Summaries.
- UMass/Mendel. (2024). Evaluation of Drug Instruction Accuracy in Llama-3 Medical Summaries.
- Goodfire. (2024). Sparse Autoencoder Technology for Neural Circuit Editing.
- Lightspeed Ventures. (2024). IDE-like Tools for Large Language Model Editing.
- MedHallu Dataset. (2024). Fine-grained Hallucination Analysis for Clinical LLM Applications. Retrieved from <https://medhallu.github.io/>
- Pandit, S., Rao, A., & Gupta, M. (2024). MedHallu: A benchmark for hallucination detection in medical AI systems. PubMedQA. Retrieved from <https://pubmed.ncbi.nlm.nih.gov/39517994/>
- Ennab, M., & Mcheick, H. (2022). Designing an interpretability-based model to explain the artificial intelligence algorithms in healthcare. *Diagnostics (Basel)*, 12(7), 1557. doi: 10.3390/diagnostics12071557
- Ennab, M., & Mcheick, H. (2024). Enhancing interpretability and accuracy of AI models in healthcare: A comprehensive review on challenges and future directions. *Frontiers in Robotics and AI*, 11, 1444763.
- Templeton, A., Conerly, T., Marcus, J., & Clark, J. (2024). Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. Retrieved from <https://transformer-circuits.pub/2024/scaling-monosemanticity/>
- VentureBeat. (2024). Real-time Feature Monitoring in Medical AI Systems.

6. Appendix

The demo at <https://v0-ai-health-dashboard.vercel.app/> provides an interactive visualization of our system, showcasing how our approach integrates with real-time data to detect and mitigate hallucinations in medical AI outputs.

AI Health Performance Dashboard

Track your AI model's performance and reduce hallucinations in healthcare applications

