TOWARD MONOSEMANTIC CLINICAL EXPLANATIONS FOR ALZHEIMER'S DIAGNOSIS VIA ATTRIBUTION AND MECHANISTIC INTERPRETABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Interpretability remains a central barrier to the safe deployment of large language models (LLMs) in high-stakes domains such as neurodegenerative disease diagnosis. In Alzheimer's disease (AD), early and explainable predictions are critical for clinical decision-making, yet attribution-based methods (e.g., saliency maps, SHAP) often suffer from inconsistency due to the polysemantic nature of LLM representations. Mechanistic interpretability promises to uncover more coherent features, but it is not directly aligned with individual model outputs, limiting its applicability in practice. To address these limitations, we propose a unified interpretability framework that integrates attributional and mechanistic perspectives via monosemantic feature extraction. First, we evaluate six common attribution techniques and further develop an explanation-optimization step that updates explanations to reduce inter-method variability and improve clarity. In the second stage, we train sparse autoencoders (SAEs) to transform LLM activations into a disentangled latent space in which each dimension corresponds to a coherent semantic concept. This monosemantic representation enables more structured and interpretable attribution analysis. We then compare feature attributions in this latent space with those from the original model, demonstrating improved robustness and semantic clarity. Evaluations on in-distribution (IID) and out-of-distribution (OOD) Alzheimer's cohorts across binary and three-class classification tasks confirm the effectiveness of our framework. By bridging attributional relevance and mechanistic clarity, our approach provides more trustworthy, consistent, and human-aligned explanations, and reveals clinically meaningful patterns in multimodal AD data. This work takes a step toward safer and more reliable integration of LLMs into cognitive health applications and clinical workflows.

1 INTRODUCTION

Explainable Artificial Intelligence (XAI) plays a crucial role in building trust in machine learning systems, especially in sensitive and high-stakes areas such as finance, climate, autonomous driving and healthcare (Doshi-Velez & Kim, 2017; Manifold et al., 2021). In medical settings, interpretability is essential for clinical integration and regulatory approval, especially in complex diseases such as Alzheimer's Disease (AD), where early and accurate detection can substantially alter treatment results (Jack et al., 2018).

Although machine learning has shown promise in AD diagnostics using multimodal clinical data (Bron et al., 2015), the application of large language models (LLMs) such as GPT-4 and LLaMA-2 in structured clinical settings remains limited (Brown et al., 2020; Touvron et al., 2023). A key obstacle is the *polysemanticity* of internal neural representations—individual neurons or features often encode multiple, semantically unrelated concepts (Olah et al., 2020b; Cunningham et al., 2023; Elhage et al., 2022a). This entanglement undermines the interpretability of standard attribution techniques such as gradients, perturbations, and integrated paths, which typically assume one-to-one correspondence between features and meanings. Moreover, existing attribution methods assign importance scores to input features (e.g., words or tokens), yet they fall short in addressing the polysemantic nature of internal representations. This limitation often leads to ambiguous or

misleading explanations—particularly problematic in clinical applications, where interpretability is critical (Samek et al., 2021; van der Velden et al., 2022; Quellec et al., 2021; Mamalakis et al., 2023). In contrast, mechanistic interpretability aims to uncover the internal structure of neural computation by identifying semantically coherent components within the model. Sparse Autoencoders have played a pivotal role in advancing our understanding of such representations in both language and vision domains (Gorton, 2024). The behavior of neural networks is often interpreted through the lens of *computational circuits*—specialized groups of neurons responsible for interpretable functions, such as edge detection (Olah et al., 2020a) or identity mapping (Olsson et al., 2022). However, these mechanistic tools typically lack attributional resolution, limiting their utility for explaining how specific inputs contribute to model predictions in real-world, decision-critical scenarios (Elhage et al., 2022a).

This reveals a critical gap in the current landscape: attributional techniques offer surface-level explanations that lack semantic clarity, while mechanistic methods offer structural insights without attributional grounding. To date, no unified framework successfully integrates both paradigms—particularly in the domain of LLM-based clinical inference.

To address this, we propose a monosemantic attribution framework that combines both attributional and mechanistic interpretability (see Figure 1). Our approach uses sparse autoencoders (SAEs) to map LLM activations into a *monosemantic feature space*, where each latent feature corresponds to a clear and disentangled concept. This transformation reduces complexity and enables attribution methods to assign more precise and meaningful scores. We evaluated six well-established attribution methods: Feature Ablation, Layer Activations, Layer Conduction, Layer Gradient SHAP (Lundberg & Lee, 2017a), Layer Integrated Gradients (Sundararajan et al., 2017b), and Layer Gradient × activation, both in the original LLM activation space and in the SAE-transformed feature space. Attribution scores are then refined using an *Explanation Optimizer*, which selects the most coherent and informative explanations based on alignment with model behavior and dataset-level consistency (Mamalakis et al., 2025). To facilitate interpretable visualization and meta-level assessment, we project the optimized attribution vectors into a 2D embedding space using UMAP (McInnes et al., 2018). We define a global *meta-rule*—attributional coherence across the first and second UMAP components—as a constraint to evaluate and regularize explanation quality. To this end, we impose a linearity constraint in the UMAP space to further enhance the interpretability of spatially structured explanation clusters.

Contributions and Novelty: We propose a unified interpretability framework that couples explainer optimisation with monosemantic feature extraction and an optional geometry-aware constraint. Concretely:

- **Transformer Explainer Optimiser (TEO).** We introduce a learning-based optimiser that refines the outputs of six common attribution methods, reducing inter-method variance and improving clarity without retraining the base model.
- **Monosemantic Bottleneck (SAE).** We train sparse autoencoders on LLM activations to obtain a disentangled latent space whose dimensions align with coherent semantic concepts, enabling structured, human-interpretable and mechanistic attribution.
- **Latent vs. Native Attributions.** We compare attributions computed in the monosemantic latent space with those in the original model space, showing improved robustness (lower RIS/ROS) and greater semantic coherence under our pipeline.
- **Tunable Sparsity–Stability Frontier.** Across IID (ADNI) and OOD (BrainLat) cohorts, and for binary and three-class tasks, TEO with monosemantic feature extractions (TEO-SAE) yields the most stable explanations, while a geometry-aware constraint (TEO-UMAP) reliably recovers higher sparsity with a modest stability cost.
- **Clinical Signal Discovery.** The framework surfaces clinically meaningful patterns in multimodal Alzheimer's data and produces class-specific, human-aligned attributions that are suitable for integration into clinical reasoning workflows.

Overall, the work unifies attributional refinement with monosemantic mechanistic structure, delivering explanations that are more stable, more coherent, and practically actionable across datasets and tasks. These improvements offer a pathway toward more trustworthy AI systems capable of providing actionable clinical insights, particularly in early-stage Alzheimer's disease detection (Super et al., 2023).

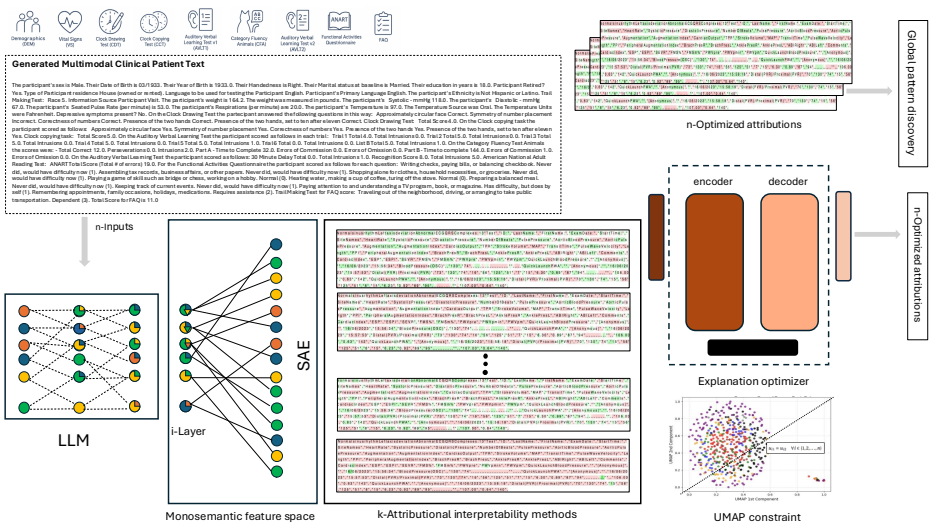

Figure 1: Proposed interpretability framework for LLM in Alzheimer's diagnosis. The model integrates k-attributional methods with a SAE to generate a monosemantic feature space. An explanation optimizer refines attribution outputs, enhancing clarity and reducing variability. Global explanation quality is visualized and assessed using UMAP and a linear meta-rule, supporting both individual prediction interpretability and cohort-level pattern discovery.

2 METHODS

2.1 ATTRIBUTIONAL THEORY AND METHODS

Attribution explainability methods follow the framework of additive feature attribution, where the explanation model $g(f,\mathbf{x})$ is represented as a linear function of simplified input features:

$$g(f,\mathbf{x}) = \phi_0 + \sum_{i=1}^{M} \phi_i x_i \tag{1}$$

Here, f is the predictive model, $\phi_i \in \mathbb{R}$ is the attribution (importance) assigned to feature x_i, and M is the number of simplified input features (here 512).

For this study, we employed six well-established attributional interpretability methods applied to large language models (LLMs), denoted as $K = 6$: *Feature Ablation*, *Layer Activations* (which capture the embedding activation space of a specific layer of interest within the LLM), *Layer Conduction*, *Layer Gradient-SHAP* (Lundberg & Lee, 2017b), *Layer Integrated Gradients* (Sundararajan et al., 2017a), and *Layer Gradient* × *Activation* (For analytical mathematic formulation see Appendix A.1.). To align these layer-wise interpretability methods with the additive feature attribution framework, we reinterpret the internal activations (i.e., latent units) of a network layer L as simplified input features. The objective is to estimate an attribution score ϕ_i for each unit, where $\phi_i \in \mathbb{R}$ quantifies the contribution of the corresponding neuron to the model's prediction.

All attribution methods were applied to the final (22nd) layer of the MODERN-BERT LLM—the model variant that achieved the highest classification accuracy in our evaluations (see Supplementary section B.4). These formulations allow us to ground various neural attribution techniques within a unified additive explanation model, facilitating their comparison and hybridization under shared theoretical assumptions.

2.2 ATTRIBUTIONAL EXPLANATION OPTIMIZER FRAMEWORK

Let $\mathscr{A} = \{A_1, A_2, \ldots, A_K\}$ denote the set of $K = 6$ attribution methods applied to the final layer L of the model f. Each method A_k generates an attribution vector $\boldsymbol{\phi}^{(k)} = [\phi_1^{(k)}, \phi_2^{(k)}, \ldots, \phi_M^{(k)}]$, where M

is the number of latent features (neurons) in layer L. The goal is to derive a unified attribution vector $\bar{\boldsymbol{\phi}}$ that captures the consensus explanation across methods. Each attribution vector $\boldsymbol{\phi}^{(k)}$ is evaluated using the Relative Input Stability (RIS), Relative Output Stability (ROS) Agarwal et al. (2022b), and Sparseness Chalasani et al. (2020) metrics (For analytical mathematic formulation see Appendix A.2.).

2.2.1 Aggregation of Attributions

The weighted average attribution vector $\bar{\boldsymbol{\phi}}$ serves as the target explanation for the optimization process and it is calculated as:

$$\bar{\boldsymbol{\phi}} = \sum_{k=1}^{K} w_k \cdot \boldsymbol{\phi}^{(k)} \tag{2}$$

2.2.2 Explanation Reconstruction via Encoder–Decoder Models

An encoder–decoder model is trained to generate a reconstructed explanation $\hat{\boldsymbol{\phi}}$ from the original input $\boldsymbol{x}$. Two architectures are considered the Diffusion UNet1D Ronneberger et al. (2015) (*Diffusion Explanation Optimizer*, DEO) and the x-transformer autoencoder Vaswani et al. (2017); Nguyen & Salazar (2019) (*Transformer Explanation Optimizer*, TEO). For the analytical mathematical formulation, see Appendix A.2.4.

2.2.3 The total cost function of the optimizer

As previously highlighted, the reconstruction of the optimal explanation and the associated cost function adhere to the same principles and architectural design outlined in Mamalakis et al. (2025). The cost function consists of three key components: sparseness, as defined in Chalasani et al. (2020); ROS and RIS scores Agarwal et al. (2022a); and similarity. The integration of these components ensures a robust and interpretable evaluation. The total cost function for training the reconstruction model is:

$$\begin{aligned}\mathscr{L}_{\text{total}}(\boldsymbol{\phi}^{(k)}, \hat{\boldsymbol{\phi}}) = &\lambda_1 \cdot \frac{1}{M_{\text{RIS}}(f, \hat{\boldsymbol{\phi}})} + \lambda_2 \cdot \frac{1}{M_{\text{ROS}}(f, \hat{\boldsymbol{\phi}})} \\ &+ \lambda_3 \cdot M_{\text{sparse}}(f, \hat{\boldsymbol{\phi}}) + \lambda_4 \cdot \mathscr{L}_{\text{similarity}}(\hat{\boldsymbol{\phi}}, \bar{\boldsymbol{\phi}})\end{aligned} \tag{3}$$

where:$\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are hyperparameters controlling the influence of each loss term. This formulation enables a principled and quantitative integration of multiple attribution methods, optimizing toward a robust and interpretable explanation.

2.3 UMAP projection and linear constraint

Given a dataset $\hat{\boldsymbol{\Phi}} = \{\hat{\boldsymbol{\phi}}_1, \hat{\boldsymbol{\phi}}_2, \ldots, \hat{\boldsymbol{\phi}}_n\} \subset \mathbb{R}^D$, UMAP McInnes et al. (2018) aims to find a low-dimensional embedding $U = \{u_1, u_2, \ldots, u_n\} \subset \mathbb{R}^d$ where typically $d = 2$ or $d = 3$, such that the local topological structure of the data in $\hat{\boldsymbol{\Phi}}$ is preserved in U (further mathematical formulation please see Appendix A.4). Let $u_i = (u_{i1}, u_{i2}, \ldots, u_{id})$ represent the embedding of the i-th data point in the d-dimensional space (see Appendix A.4). The constraint that the first and second components of the embedding are equal can be written as:

$$u_{i1} = u_{i2} \quad \forall i \in \{1, 2, \ldots, n\} \tag{4}$$

In other words, the first component u_{i1} and the second component u_{i2} of each embedding vector u_i must be equal. This can be written as a linear equality constraint, $u_{i1} - u_{i2} = 0 \quad \forall i \in \{1, 2, \ldots, n\}$. This constraint ensures that for each data point i, the first and second components of the corresponding embedding vector u_i are equal. In the $\mathscr{L}_{\text{total}}(\boldsymbol{\phi}^{(k)}, \hat{\boldsymbol{\phi}})$, of eq. 3, we can add an extra penalty term to the loss function to enforce this constraint. The penalty term would be:

$$\lambda_5 \sum_{i=1}^{n} (u_{i1} - u_{i2})^2 \tag{5}$$

where λ_5 is a regularization parameter that controls the strength of the penalty. This penalty term enforces the condition that the first and second components of each embedding point of the reconstructed explanation from the optimizer ($\hat{\boldsymbol{\phi}}$) are equal, but it allows flexibility depending on the value of λ_5.

2.4 THE SAE APPROACH AND ARCHITECTURES

The mathematical formulation situates SAE architectures within the theoretical framework of superposition and semantic disentanglement (for an analytical mathematical formulation, see Appendix A.5). By expressing hidden states as sparse linear combinations of interpretable features, SAEs bridge the gap between low-level activations and human-understandable concepts. Let $\mathbf{x} \in \mathbb{R}^d$ denote a layer's neuron activation vector in a pretrained model. A Sparse Autoencoder learns a sparse feature representation $\mathbf{a} \in \mathbb{R}^F$ such that:

$$\hat{\mathbf{x}} = W\mathbf{a} + \mathbf{b}, \tag{6}$$

where $W \in \mathbb{R}^{d \times F}$ is the decoder (dictionary) matrix and $\mathbf{b} \in \mathbb{R}^d$ is a learned bias term. Each column $W_{\cdot,i}$ represents the direction of feature i in neuron space, and a_i is its activation. This linear mapping enables complex activations to be expressed as combinations of more interpretable features. If $F > d$, then the feature space is overcomplete, and W cannot be full-rank. This leads to superposition, where multiple features overlap in the same subspace, and individual neurons encode multiple unrelated concepts Elhage et al. (2022b). If W is invertible and aligned to a basis, each neuron corresponds to a single feature. The representation is monosemantic and disentangled Olah et al. (2020b). When W has overlapping columns, neurons can respond to multiple features, yielding polysemantic behavior. That is, for some j, $x_j = \sum_i W_{j,i} a_i$ involves multiple nonzero terms Bills et al. (2023). Variants of SAEs like TopK, JumpReLU, and Gated-SAEs offer increasingly precise control over the mapping between low-level activations and human-understandable concepts, enabling fine-grained analysis and intervention (analytical mathematical formulation, see Appendix A.6.).

2.5 ATTRIBUTION FROM SPARSE FEATURE SPACE TO INPUT TOKENS

Let $\mathbf{x}_{\text{input}} \in \mathbb{R}^{d_{\text{input}}}$ denote the input embedding vector (e.g., LLM token embeddings), $\mathbf{x} = f(\mathbf{x}_{\text{input}}) \in \mathbb{R}^d$ the hidden layer activation of the LLM, $\mathbf{a} = \text{Encoder}(\mathbf{x}) \in \mathbb{R}^F$ the SAE sparse feature vector, and $\hat{\mathbf{x}} = W\mathbf{a} + \mathbf{b}$ the reconstructed activation from the SAE decoder. Now suppose we have a sparse attribution vector ψ_i over features $\mathbf{a}$, i.e., $\psi \in \mathbb{R}^F$, where each ψ_i reflects the importance of SAE feature a_i. We aim to assign importance Φ_k to each input token dimension $x_{\text{input},k}$.

We propagate the feature attributions backward through the encoder to the input. Using the chain rule:

$$\Phi_k = \sum_{i=1}^{F} \psi_i \cdot \frac{\partial a_i}{\partial x_{\text{input},k}} = \sum_{i=1}^{F} \psi_i \cdot \frac{\partial a_i}{\partial \mathbf{x}} \cdot \frac{\partial \mathbf{x}}{\partial x_{\text{input},k}} \tag{7}$$

where $\frac{\partial a_i}{\partial \mathbf{x}}$ is the encoder Jacobian (SAE layer), and $\frac{\partial \mathbf{x}}{\partial x_{\text{input},k}}$ is the LLM gradient from input token to hidden layer. This gives us a scalar attribution $\Phi_k \in \mathbb{R}$ for each token/input embedding dimension k. This represents how much each input token contributes to the sparse SAE features identified as important. In doing so, we assess the contribution of input features based on the monosemantic behavior of the trained network's internal mechanisms. Building on our study, we apply the six previously discussed attribution methods at two levels: from the SAE feature space to the encoder layer, and from the encoder layer to the input embedding space. This dual-level attribution analysis enables us to investigate how interpretable sparse features relate to model internals and ultimately shape the input-level representations. Attribution methods (e.g., Integrated Gradients, SHAP) can directly estimate:

$$\boldsymbol{\phi}^{\text{input}} = \text{AttributionMethod}(f, \mathbf{x}_{\text{input}}, \boldsymbol{\phi}^{\text{SAE}}) \tag{8}$$

where $\boldsymbol{\phi}^{\text{SAE}}$ denotes the monosemantic feature space of the SAE network. Thus, the dual-level approach allows us to connect semantically meaningful sparse features to the raw input representation space.

2.6 LLM NETWORKS AND HYPERPARAMETER TUNING

We evaluate encoder-based LLMs (BERT, RoBERTa, DistilBERT, ALBERT, BioBERT, ModernBERT) on ADNI (IID) and BRAINLAT (OOD) under a unified protocol spanning full fine-tuning, zero-shot,

few-shot with temperature control, and parameter-efficient LoRA. *ModernBERT* outperformed all other networks on ADNI: in the binary task it achieved the highest F1 (75.89%), AUC-PR (86.41%), ROC-AUC (83.95%), and Accuracy (72.37%), and it remained strongest in the three-class setting (F1 68.80%, AUC-PR 78.48%, ROC-AUC 78.67%, Accuracy 65.05%). For OOD model adaptation, *ModernBERT* zero-shot yielded 55% Accuracy, few-shot/LoRA provided modest gains (62%), while full fine-tuning peaked at 84% Accuracy but lies outside our scope. Accordingly, we use *ModernBERT* fine-tuned on ADNI for IID and zero-shot on BRAINLAT for OOD, and all interpretability analyses are conducted on the 22nd layer of *ModernBERT*. We conducted extensive hyperparameter tuning for all components. The explanation optimizer performed best at a learning rate of 2e^{-4}, with the optimal weight configuration $(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (0.1, 0.3, 0.1, 0.5)$. UMAP constraints were most effective at the 4× batch size level (4×64). Among SAE variants, *TopK* achieved the best results with a 32× feature depth. All models were trained using the AdamW optimizer with early stopping and standard evaluation metrics. Further hyperparameter tuning and evaluation details are provided in Supplementary section B.5 and B.6.

2.7 DATASET AND CODE AVAILABILITY

The dataset used in this study originates from the ADNI cohort Mueller et al. (2005), represented as text generated from multiple modalities, serving as the in-distribution (IID) dataset. For the out-of-distribution (OOD) cohort, we used text generated from multimodal sources (MRI and clinical files) in the Latin American Brain Health Institute (BrainLat) dataset, a multi-site initiative providing neuroimaging, cognitive, and clinical data across several Latin American countries Prado et al. (2023). Additional demographic and preprocessing details are provided in Supplementary Sections B.1–B.4. The code is implemented in Python using PyTorch and runs on an NVIDIA cluster in one A100-SXM-80GB GPU. It leverages `SAE_LENs` Bloom et al. (2024) for SAE training, `quantus` Hedström et al. (2023) for evaluation, and `captum` for attribution analysis. The codebase `https://anonymous.4open.science/r/TEO_SAE-6D24/README.md`.

3 RESULTS

3.1 ABLATION AND EVALUATION SCORES WITH/WITHOUT UMAP CONSTRAINT AND SAE LAYER.

In this study, sparseness is defined such that higher values correspond to more selective and concentrated attributions across input features—that is, greater sparseness. However, sparseness alone is insufficient to assess explanation quality, as it does not account for robustness or stability. Therefore, the most effective explanation method is one that simultaneously achieves high sparseness and low RIS and ROS values.

Across IID, OOD datasets and for both binary and three-class classification tasks, Table 1 shows a consistent stability–sparsity frontier governed by the proposed optimizers and a monosemantic bottleneck (SAE). In the binary IID case, SAE substantially improves stability for feature-learning explainers, most notably: Layer Conductance and especially TEO, with large drops in RIS/ROS for both Alzheimer's and Control, whereas Activation–SAE increases RIS/ROS relative to its no-SAE variant and is therefore less robust. In the binary OOD case, this pattern persists and strengthens: TEO–SAE achieves the lowest RIS/ROS overall (strong cross-dataset stability), while TEO–UMAP recovers higher sparsity (> 0.40) at a modest stability cost versus TEO–SAE, offering a tunable sparsity–stability trade-off. In the three-class IID setting, Feature Ablation is the sparsity leader across Control/LMCI/MCI (~ 0.52–0.53) with moderate, steady yet still high RIS/ROS values; Layer Conductance–SAE markedly reduces RIS/ROS for LMCI/MCI; and TEO–SAE again delivers the most stable attributions across all classes, albeit with reduced sparsity compared with no-SAE variants. The same rank ordering holds OOD. Across all blocks, gradient-formulaic methods (Grad-SHAP, Guided Backprop, Integrated Gradients) show near-invariant RIS/ROS (~ 5.6/ ~ 16.93) regardless of SAE, class, or domain, indicating that SAE chiefly benefits learned-attribution methods. Further analyses are provided in Supplementary §B.7, Tables 2–5.

3.2 INDIVIDUAL-LEVEL AND COHORT-LEVEL EXPLANATIONS AND PATTERNS

Figure 2 shows qualitative local attributions for the the LMCI class of the three-class classification task on IID and OOD using our proposed optimiser TEO, TEO-SAE, and TEO-UMAP (SAE). Tokens

Table 1: Unified attribution summary split vertically by task and setting. Values are mean±std per class. Binary columns use (A/C) = Alzheimer/Control; Three-class columns use (L/M/C) = LMCI/MCI/Control.

Task & Setting	Method (row = variant)	Sparseness	RIS	ROS
Binary — IID (ADNI)	Columns: (A/C)			
	Activation	0.3164/0.2562 ± 0.0076/0.0176	14.3023/14.2365 ± 0.3686/0.3381	25.5485/25.5487 ± 0.4828/0.3286
	Activation-SAE	0.2966/0.2520 ± 0.0071/0.0047	21.3084/19.3275 ± 0.3110/0.9323	32.6174/30.6394 ± 0.3079/0.9334
	Layer Conductance	0.3966/0.3745 ± 0.0261/0.0071	12.3985/5.6502 ± 2.6406/0.0391	23.1466/16.9615 ± 1.6865/0.0318
	Layer Conductance-SAE	0.3915/0.2480 ± 0.0076/0.0079	5.6285/5.6141 ± 0.0236/0.0184	16.9471/16.9301 ± 0.0103/0.0052
	Feature Ablation	0.5236/0.5256 ± 0.0098/0.0110	23.1523/22.5611 ± 0.8198/0.2884	33.9088/33.9076 ± 0.1613/0.3747
	Feature Ablation-SAE	0.5235/0.5265 ± 0.0104/0.0088	23.5609/23.6221 ± 0.1033/0.0933	34.9298/34.9726 ± 0.0745/0.1033
	Gradient SHAP	0.3192/0.4333 ± 0.0043/0.0030	5.6231/5.6325 ± 0.0237/0.0235	16.9357/16.9461 ± 0.0018/0.0022
	Gradient SHAP-SAE	0.0820/0.1339 ± 0.0155/0.0099	5.6218/5.6196 ± 0.0227/0.0190	16.9345/16.9348 ± 1.5e−5/6.0e−6
	Gradient Activation	0.3277/0.2500 ± 0.0384/0.0230	5.6149/5.6170 ± 0.0193/0.0221	16.9303/16.9347 ± 0.0034/0.0
	Gradient Activation-SAE	0.2035/0.1668 ± 0.0117/0.0072	5.6252/5.6173 ± 0.0213/0.0221	16.9343/16.9347 ± 6.8e−5/4.0e−5
	Integrated Gradient	0.2983/0.4304 ± 0.0080/0.0066	5.6206/5.6278 ± 0.0180/0.0190	16.9326/16.9434 ± 0.0015/0.0024
	Integrated Gradient-SAE	0.1212/0.0644 ± 0.0058/0.0059	5.6224/5.6214 ± 0.0178/0.0169	16.9345/16.9346 ± 1.2e−5/8.3e−6
	DEO	0.3383/0.3377 ± 0.0033/0.0017	9.2839/9.3131 ± 0.0800/0.1427	20.6342/20.6159 ± 0.0866/0.2026
	DEO-SAE	0.3374/0.3140 ± 0.0029/0.0010	9.2790/9.1750 ± 0.0646/0.1088	20.6150/20.5150 ± 0.0880/0.1299
	TEO	0.4220/0.4199 ± 0.0003/0.0005	5.0520/5.0688 ± 0.0192/0.0184	16.3529/16.3777 ± 0.0056/0.0011
	TEO-SAE	0.2672/0.2682 ± 0.0010/0.0007	1.6227/0.9964 ± 0.1708/0.2639	12.9250/12.2983 ± 0.1703/0.2613
	TEO-UMAP *(SAE)*	0.3989/0.4057 ± 0.0004/0.0003	5.4394/5.4709 ± 0.0332/0.1746	16.3037/16.2102 ± 0.0033/0.0079
Binary — OOD (BrainLat)	Columns: (A/C)			
	Activation-SAE	0.1533/0.3965 ± 0.0103/0.0303	19.1625/18.2412 ± 0.3642/0.5392	31.2827/29.0406 ± 1.5414/0.4731
	Layer Conductance-SAE	0.2392/0.2543 ± 0.0298/0.0210	6.1621/6.2149 ± 0.1495/0.2076	16.9438/16.9403 ± 0.0071/0.0050
	Feature Ablation-SAE	0.5288/0.5285 ± 0.0070/0.0044	23.5834/24.1474 ± 0.0645/0.1160	34.6531/34.9613 ± 0.2526/0.2205
	Gradient SHAP-SAE	0.1201/0.0571 ± 0.0144/0.0271	6.0440/6.0303 ± 0.0396/0.0471	16.9347/16.9348 ± 5.8e−5/5.7e−6
	Gradient Activation-SAE	0.1140/0.0630 ± 0.0177/0.0069	6.0328/6.0339 ± 0.0277/0.0398	16.9347/16.9348 ± 3.6e−6/3.7e−5
	Integrated Gradient-SAE	0.0643/0.0143 ± 0.0052/0.0003	6.0579/6.0276 ± 0.0456/0.0339	16.9348/16.9349 ± 7.8e−6/1.1e−5
	TEO-SAE	0.2691/0.2725 ± 0.0016/0.0004	0.6835/0.4734 ± 0.6676/0.2801	11.5236/11.2130 ± 0.6591/0.5150
	TEO-UMAP *(SAE)*	0.3989/0.4043 ± 0.0005/0.0029	5.4394/5.4282 ± 0.0383/0.1944	16.3037/16.1577 ± 0.0039/0.1054
Three-class — IID (ADNI)	Columns: (L/M/C)			
	Activation	0.2715/0.2626/0.3030 ± 0.0384/0.0379/0.0377	15.0786/16.6568/14.4042 ± 1.9754/2.8208/0.1660	26.3951/27.9606/25.7217 ± 1.9727/2.8233/0.171
	Activation-SAE	0.2644/0.3091/0.3450 ± 0.0630/0.0603/0.0095	18.4231/19.4227/18.9968 ± 2.3518/3.4745/4.4840	29.7333/30.7419/30.3026 ± 4.8472/6.1045/0.2636
	Layer Conductance	0.3623/0.3053/0.2315 ± 0.0064/0.0076/0.0096	13.1429/6.6083/5.6260 ± 0.3255/2.2363/0.0209	24.5060/17.9191/16.9440 ± 0.4120/2.2581/0.0045
	Layer Conductance-SAE	0.2464/0.2930/0.3315 ± 0.0628/0.0578/0.0062	5.6236/5.6291/5.6222 ± 0.9099/1.1307/0.0149	16.9338/16.9384/16.9390 ± 2.7457/3.4067/0.0105
	Feature Ablation	0.5226/0.5222/0.5239 ± 0.0097/0.0094/0.0067	22.2447/23.4984/23.3250 ± 0.1629/0.4587/0.4109	33.6064/34.8737/34.6653 ± 0.1615/0.4407/0.4580
	Feature Ablation-SAE	0.5268/0.5257/0.5261 ± 0.0840/0.1048/0.0120	21.9794/23.0006/23.0766 ± 3.6802/4.5402/0.1403	33.3071/34.3136/34.4179 ± 5.4997/6.8050/0.1463
	Gradient SHAP	0.1292/0.0891/0.2310 ± 0.0326/0.0131/0.0206	5.6152/5.6292/5.6189 ± 0.0144/0.0187/0.0139	16.9255/16.9392/16.9358 ± 0.0014/0.0021/0.0021
	Gradient SHAP-SAE	0.3011/0.2881/0.1844 ± 0.0721/0.0618/0.0148	5.6217/5.6186/5.6219 ± 0.9462/1.1670/0.0253	16.9348/16.9348/16.9348 ± 2.8625/3.5241/1.4e−5
	Guided Backprop	0.3839/0.2917/0.2697 ± 0.0177/0.0200/0.0061	5.6272/5.6269/5.6290 ± 0.0212/0.0193/0.0225	16.9339/16.9340/16.9347 ± 0.0008/0.0006/0
	Guided Backprop-SAE	0.4310/0.2579/0.2296 ± 0.1156/0.1095/0.0036	5.6297/5.6172/5.6210 ± 0.9478/1.1684/0.0194	16.9347/16.9347/16.9348 ± 2.8625/3.5241/1.22e−5
	Integrated Gradient	0.1084/0.1102/0.0451 ± 0.0262/0.0157/0.0071	5.6094/5.6283/5.6207 ± 0.0178/0.0209/0.0215	16.9276/16.9358/16.9331 ± 0.0006/0.0020/0.0011
	Integrated Gradient-SAE	0.3889/0.2660/0.2639 ± 0.0841/0.0905/0.0042	5.6282/5.6203/5.6209 ± 0.9476/1.1685/0.0209	16.9343/16.9346/16.9348 ± 2.8625/3.5240/1.22e−5
	TEO	0.4131/0.3909/0.3918 ± 0.0003/0.0047/0.0008	5.0938/4.8283/4.8080 ± 0.0188/0.0377/0.0184	16.4043/16.1354/16.1172 ± 0.0024/0.0324/0.0090
	TEO-SAE	0.2860/0.2838/0.2682 ± 0.0374/0.0523/0.0649	2.2642/2.1617/1.5468 ± 0.4877/0.4547/0.1171	13.5646/13.4676/12.8570 ± 2.2745/2.7641/0.1179
	TEO-UMAP *(SAE)*	0.4161/0.4172/0.3973 ± 0.0870/0.2372/0.0749	5.1017/5.1116/5.1086 ± 0.1697/0.1072/0.2083	16.4031/16.4088/16.4123 ± 3.8616/0.4924/6.8439
Three-class — OOD (BrainLat)	Columns: (L/M/C)			
	Activation-SAE	0.1907/0.1400/0.4505 ± 0.0016/0.0124/0.0375	18.6406/18.0787/19.0940 ± 0.9117/0.0313/0.1796	29.5628/28.8583/29.9240 ± 0.8978/0.0452/0.2538
	Layer Conductance-SAE	0.1857/0.2006/0.3252 ± 0.0073/0.0119/0.0145	6.0546/6.2684/6.2120 ± 0.0477/0.0401/0.2450	16.9637/17.0146/16.9582 ± 0.0071/0.0206/0.0173
	Feature Ablation-SAE	0.5262/0.5293/0.5281 ± 0.0057/0.0112/0.0058	23.5693/23.5916/22.6406 ± 0.0577/0.1118/0.0331	34.5169/34.3720/33.4853 ± 0.0737/0.0804/0.2338
	Gradient SHAP-SAE	0.0637/0.1137/0.1951 ± 0.0265/0.0420/0.0264	6.0315/6.1274/6.1227 ± 0.0298/0.1207/0.1619	16.9349/16.9347/16.9346 ± 7.6e−5/8.9e−5/4.3e−5
	Gradient Activation-SAE	0.1836/0.4303/0.1772 ± 0.0016/0.0022/0.0112	6.0269/6.1450/6.1234 ± 0.0302/0.1210/0.1912	16.9348/16.9345/16.9346 ± 2.6e−6/2.8e−5/1.5e−5
	Integrated Gradient-SAE	0.0072/0.0361/0.0671 ± 0.0009/0.0049/0.0121	6.0356/6.1478/6.1225 ± 0.0189/0.0924/0.1502	16.9348/16.9346/16.9346 ± 1.1e−6/8.9e−6/1.3e−5
	TEO-SAE	0.3716/0.4224/0.4162 ± 0.0009/0.0002/0.0029	4.9396/5.5421/5.7520 ± 0.0148/0.0611/0.3645	15.8121/16.2773/16.3792 ± 0.0099/0.0010/0.0034
	TEO-UMAP *(SAE)*	0.4239/0.4246/0.4238 ± 5.0e−5/2.4e−4/5.6e−4	5.4583/5.5576/5.5525 ± 0.0297/0.0954/0.1862	16.3726/16.3572/16.3661 ± 0.0017/0.0035/0.0073

are colour-coded (green = positive relevance; red = negative). Visually, TEO-SAE produces the tightest, least noisy explanations—fewer spurious highlights and clearer token groupings—consistent with its lowest RIS/ROS in Table 1. Adding the UMAP constraint restores higher sparseness while preserving much of TEO-SAE stability: TEO-UMAP (SAE) yields compact, well-separated patterns that remain clinically interpretable across IID and OOD. Across classes and datasets, higher Sparseness corresponds to less diffuse maps with balanced positive/negative highlights, whereas low Sparseness with high RIS/ROS manifests as saturated red/green patches and unstable saliency (see Supplementary Figures 7–16). Among the six classical attribution methods (Activation, Layer Conductance, Feature Ablation, Gradient SHAP, Gradient Activation, Integrated Gradient), Feature Ablation attains the highest Sparseness but exhibits poorer stability (elevated RIS/ROS), a gap that worsens with SAE due to decoder-driven "decompression" (Supplementary Figures 7–8). Layer Conductance shows the opposite trade-off: SAE reduces Sparseness but improves stability (lower RIS/ROS), with similar stability gains observed for Gradient Activation, Integrated Gradient, and Gradient SHAP (Supplementary Figures 9–10). Overall, none of these classical methods match the proposed framework as TEO-SAE is consistently most stable, and TEO-UMAP (SAE) offers a tunable sparsity–stability compromise that generalises from ADNI to BrainLat (for extended analyses see Supplementary §B.9).

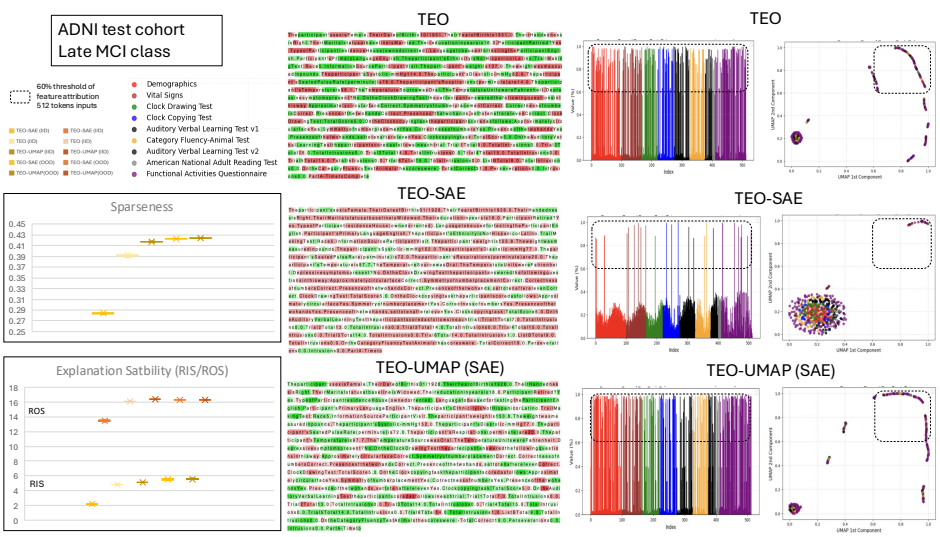

Figure 2: Stability–sparsity frontier for explanation optimizers on ADNI (Late MCI). Scatter shows TEO, TEO–SAE, and TEO–UMAP (SAE) on ADNI (IID) and BrainLat (OOD). Metrics: Sparseness (higher is better) vs. RIS/ROS stability (lower is better). Thresholding uses 60% feature attribution over 512 tokens. TEO–SAE yields the most stable explanations, TEO–UMAP recovers higher sparsity with modest stability cost, and TEO lies between.

Moving from TEO to TEO–SAE produces tighter, more homogeneous low-to-high attribution and the lowest RIS/ROS (highest stability), but also a marked reduction in sparseness, evident as broader token spread in the 2D manifold (see Figure 2). In some cases, this stabilisation concentrates signals so strongly that few features exceed the significance threshold (square box in the 2D scatter plot where $UMAP\text{-}PC1/PC2 \geq 0.6$), and not all subgroups are represented (Figure 2, Supplementary Figure 32). Imposing a linear UMAP constraint (TEO–UMAP) mitigates this effect by restoring sparsity in significant attributions while retaining stability, yielding compact, clinically interpretable maps with more uniform subgroup coverage (Figure 2; Supplementary Figures 33–35). The behaviour of the proposed framework (TEO, TEO–SAE, TEO–UMAP) shows that higher sparseness corresponds to less diffuse, more balanced highlights, whereas lower sparseness with higher RIS/ROS results in saturated red/green patches. This mirrors the patterns observed across the six classical methods (Activation, Layer Conduction, Feature Ablation, Gradient SHAP, Gradient Activation, Integrated Gradient; Supplementary §B.10). With SAE, feature-learning explainers such as Layer Conduction generally gain stability (lower RIS/ROS) at some sparsity cost, while Feature Ablation maintains high sparsity but remains unstable. None, however, match the stability–sparsity trade-off achieved by TEO–SAE and TEO–UMAP (box plots in Figure 2; Table 1). Supplementary §B.10 (Figures 22–31) provides more details about cohort-level attributions.

3.3 Are monosemantic representation–based attribution methods statistically distinct from standard attribution techniques?

A statistical evaluation of interpretability metrics (sparseness, RIS, and ROS) across methods with and without the SAE layer was computed. In the binary ADNI task, *paired t-tests* with FDR correction showed that adding an SAE bottleneck significantly reduced Complexity ($p < 10^{-10}$) and RIS ($p < 10^{-4}$) in both groups, while ROS changes were small and inconsistent (marginal for Control, non-significant for Alzheimer's). The strongest SAE effects appeared in attribution metrics, with Gradient SHAP ($p < 10^{-45}$), Layer Conduction ($p = 3.2 \times 10^{-7}$), Integrated Gradients ($p < 10^{-55}$), and the TEO ($p < 10^{-95}$) all showing decisive reductions, confirming robust stability gains under SAE. In the three-class ADNI task, *paired t-tests and Wilcoxon signed-rank tests* (BH-FDR) indicated that the MCI group showed the clearest improvement: ROS decreased strongly ($t(17) = -10.12,\ p = 1.30 \times 10^{-8}$; $W = 0,\ p = 8.0 \times 10^{-6},\ q = 2.3 \times 10^{-5}$), RIS showed a smaller reduction detected non-parametrically ($W = 19,\ p = 0.00117$), and Complexity increased modestly by Wilcoxon ($W = 17,\ p = 7.9 \times 10^{-4}$) while paired t-tests were non-significant. Control and LMCI

had incomplete pairs, preventing matched testing with correction. Overall, SAE reliably improves attribution stability (lower RIS/ROS scores) and reduces Complexity in binary tasks, with the three-class MCI group showing the most consistent ROS gains. Full analytical results and additional details are provided in Supplementary §B.8.

3.4 The clinical impact and outcome in the diagnosis of Alzheimer

This study shows that the TEO-SAE and TEO-UMAP provide the most reliable identification of informative sources across this nine multimodal subgroups. Using a significance threshold of 0.6 on UMAP principal components PC1/PC2 (Figure 2), we observe in the binary task of the IID test cohort (ADNI) that, for Control, TEO-SAE is dominated by FAQ, whereas TEO-UMAP emphasises DEM, AVLT2, and FAQ; for Alzheimer's, TEO prioritises FAQ, AVLT1, and CFA, while TEO-UMAP highlights ANART, FAQ, and DEM. In the three-class task, for Control the main contributors are AVLT1, CDT, and ANART under TEO, and AVLT1, CDT, and CFA under TEO-UMAP; for MCI, TEO favours CCT, AVLT2, and FAQ, whereas TEO-UMAP favours AVLT2, ANART, and CFA; and for LMCI, TEO elevates AVLT1, FAQ, and CDT, while TEO-UMAP elevates FAQ, ANART, and AVLT2. These patterns are summarised in Supplementary Table 6 (Section B.11) and the acronyms follow the description in Sections B.1–B.3.

3.5 Limitation and Future Work

While our framework demonstrates substantial improvements in attribution clarity and robustness, some limitations remain. A generalized outcome about clinical LLMs is not feasible at the level of this study, as the analysis was restricted to the neurodegenerative domain, limiting generalisability to other areas such as oncology. Methodologically, we evaluated only one type of monosemantic bottleneck (SAE) and a linear UMAP constraint, leaving alternative architectures and constraint families (e.g., neuro-symbolic or meta-learning) unexplored. Constraining the manifold space of explanations with explicit guidance from clinical experts could further improve explanation quality and enhance pattern discovery within the proposed framework. In addition, while stability–sparsity assessment focused on RIS/ROS and sparsity indices as important first-level evaluation metrics, additional measures such as uncertainty quantification and fairness auditing should be incorporated in future work. Future work will strengthen more these results by validate prospectively, extend to additional centres/modalities and other clinical domains (e.g., oncology), explore alternative constraints, and incorporate uncertainty and fairness auditing.

4 Conclusion

We introduce a unified interpretability framework that combines explainer optimisation with a monosemantic bottleneck (TEO–SAE) and an optional geometry-aware constraint (TEO–UMAP). Across IID (ADNI) and OOD (BrainLat), and in both binary and three-class settings, TEO–SAE consistently achieves the lowest RIS/ROS (highest stability), while TEO–UMAP reliably recovers higher sparsity at a modest stability cost—establishing a tunable sparsity–stability frontier that generalises across tasks and distribution shift. Gradient-based baselines change little with SAE, whereas SAE substantially benefits feature-learning explainers (like Layer Conduction), and none of the classical techniques surpass our optimisers, underscoring the value of learning monosemantic features within an explainer-optimisation pipeline. Clinically, stable contributors concentrate on functional status (FAQ) and memory (AVLT1/AVLT2), with visuospatial performance (CDT) features recurring in Control/LMCI. TEO–SAE emphasises neuropsychological performance signals, whereas TEO–UMAP exposes complementary demographic/language markers, yielding class-specific, clinically interpretable profiles. A simple UMAP PC1/PC2 0.6 rule produces actionable cohort-level attribution maps that can prioritise assessments, reduce testing burden, inform trial enrichment, and support personalised monitoring. Critically, while increasing feature dimensionality can erode attribution quality in standard methods, transformer-based optimization explainers remain resilient when guided by geometric and structural constraints. Taken together, our results, both theoretical and empirical, indicate that integrating monosemantic encoding with geometry-aware explanation can enhance robust, human-aligned interpretability in neuroscience-focused AI.

Reproducibility Statement

For reproducibility during review, anonymised source codes, results and datasets are included in the supplementary material and in the subsection 2.7 'Dataset and code availability' of the main manuscript.

Ethics Statement

The paper discusses several potential positive societal impacts, particularly emphasizing its relevance to clinical applications such as the early diagnosis and treatment planning of Alzheimer's Disease. By proposing a unified interpretability framework that combines attributional and mechanistic techniques, the authors aim to enhance the trustworthiness, consistency, and human alignment of large language model (LLM) outputs. This improved interpretability is presented as a means to support safer and more effective integration of LLMs into cognitive health and clinical decision-making, with the potential to uncover clinically meaningful patterns and ultimately improve patient outcomes. However, the paper does not explicitly address possible negative societal impacts of the work. It does not discuss risks such as the misinterpretation of model ex-677 planations, over-reliance on machine-generated insights in high-stakes medical contexts, or the potential for the framework to inadvertently reinforce biases embedded in training data. Societal impacts can be better established through future work, in which we plan to incorporate clinician-in-the-loop evaluation and patients.

Acknowledgments

Data collection and sharing for the Alzheimer's Disease Neuroimaging Initiative (ADNI) is funded by the National Institute on Aging (National Institutes of Health Grant U19AG024904). The grantee organization is the àNorthern California Institute for Research and Education. Past funding was obtained from: the National Institute of Biomedical Imaging and Bioengineering, the Canadian Institutes of Health Research, and private sector contributions through the Foundation for the National Institutes of Health (FNIH) including generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; BristolMyers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research Development, LLC.; Johnson Johnson Pharmaceutical Research Development LLC.; Lumosity; Lundbeck; Merck Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.

References

Chirag Agarwal, Nari Johnson, Martin Pawelczyk, Satyapriya Krishna, Eshika Saxena, Marinka Zitnik, and Himabindu Lakkaraju. Rethinking stability for attribution-based explanations, 2022a. URL https://arxiv.org/abs/2203.06877.

Chirag Agarwal, Nari Johnson, Martin Pawelczyk, Satyapriya Krishna, Eshika Saxena, Marinka Zitnik, and Himabindu Lakkaraju. Rethinking stability for attribution-based explanations, 2022b. URL https://arxiv.org/abs/2203.06877. arXiv preprint arXiv:2203.06877.

Peter Bills, Jyothi Guntupalli, et al. Language models represent space and time. *Nature Neuroscience*, 26(5):707–717, 2023.

Joseph Bloom, Curt Tigges, Anthony Duong, and David Chanin. Saelens. https://github.com/jbloomAus/SAELens, 2024.

Esther E Bron et al. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural mri: The caddementia challenge. *NeuroImage*, 111:562–579, 2015.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models

are few-shot learners. In *Advances in neural information processing systems*, volume 33, pp. 1877–1901, 2020.

Prasad Chalasani, Jiefeng Chen, Amrita Roy Chowdhury, Xi Wu, and Somesh Jha. Concise explanations of neural networks using adversarial training. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1383–1391, Virtual Event, online, July 13–18 2020. PMLR. Originally released as arXiv:1810.06583 (2018).

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017. URL https://arxiv.org/abs/1702.08608.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition, 2022a. URL https://arxiv.org/abs/2209.10652.

Nelson Elhage, Neel Nanda, et al. A mechanistic interpretability analysis of superposition in neural networks. *Transformer Circuits Thread*, 2022b. URL https://transformer-circuits.pub/2022/superposition/.

Liv Gorton. The missing curve detectors of inceptionv1: Applying sparse autoencoders to inceptionv1 early vision. *arXiv preprint arXiv:2406.03662*, 2024. URL https://arxiv.org/abs/2406.03662.

Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina Marina M.-C. Höhne. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023. URL http://jmlr.org/papers/v24/22-0142.html.

Clifford R Jack et al. Nia-aa research framework: Toward a biological definition of alzheimer's disease. *Alzheimer's & Dementia*, 14(4):535–562, 2018.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017a.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874, 2017b. URL http://arxiv.org/abs/1705.07874.

Michail Mamalakis, Krit Dwivedi, Michael Sharkey, Samer Alabed, David Kiely, and Andrew J. Swift. A transparent artificial intelligence framework to assess lung disease in pulmonary hypertension. *Scientific Reports*, 13(1):3812, 2023. doi: 10.1038/s41598-023-30503-4. URL https://doi.org/10.1038/s41598-023-30503-4.

Michail Mamalakis, Antonios Mamalakis, Ingrid Agartz, Lynn Egeland Mørch-Johnsen, Graham K. Murray, John Suckling, and Pietro Lio. Solving the enigma: Enhancing faithfulness and comprehensibility in explanations of deep networks. *AI Open*, 6:70–81, 2025. ISSN 2666-6510. doi: 10.1016/j.aiopen.2025.02.001. URL http://dx.doi.org/10.1016/j.aiopen.2025.02.001.

Tom Manifold, Fei Jiang, et al. Trustworthy ai: A computational framework to guide clinical and regulatory policy. *Nature Machine Intelligence*, 3(8):667–677, 2021.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Susanne G. Mueller, Michael W. Weiner, Leon J. Thal, Ronald C. Petersen, Clifford Jack, William Jagust, John Q. Trojanowski, Arthur W. Toga, and Laurel Beckett. The alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics of North America*, 15(4):869–877, 2005. ISSN 1052-5149. doi: https://doi.org/10.1016/j.nic.2005.09.008. URL https://www.sciencedirect.com/science/article/pii/S1052514905001024. Alzheimer's Disease: 100 Years of Progress.

Toan Q. Nguyen and Julian Salazar. Transformers without tears: Improving the normalization of self-attention. 2019. doi: 10.5281/zenodo.3525484.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024.001, 2020a. doi: 10.23915/distill.00024.001. URL https://distill.pub/2020/circuits/zoom-in.

Chris Olah, Arvind Satyanarayan, Ludwig Schubert Wusser, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020b. doi: 10.23915/distill.00024.001.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/in-context-learning/index.html.

Pavel Prado, Vicente Medel, Agustín Sainz-Ballesteros, Hernando Santamaría-García, Sebastián Moguilner, Jhony Mejía, Raúl González-Gómez, Andrea Slachevsky, María Isabel Behrens, David Aguillón, Francisco Lopera, Mario A. Parra, Diana Matallana, Marcelo Adrián Maito, Adolfo M. García, Nilton Custodio, Alberto Ávila Funes, Stefanie Piña-Escudero, Agustina Birba, Sol Fittipaldi, Agustina Legaz, and Agustín Ibáñez. The brainlat project: a multimodal neuroimaging dataset of neurodegeneration from underrepresented backgrounds. *Scientific Data*, 10:889, 2023. doi: 10.1038/s41597-023-02806-8. URL https://www.nature.com/articles/s41597-023-02806-8.

Gwenolé Quellec, Hassan Al Hajj, Mathieu Lamard, Pierre-Henri Conze, Pascale Massin, and Béatrice Cochener. Explain: Explanatory artificial intelligence for diabetic retinopathy diagnosis. *Medical Image Analysis*, 72:102118, 2021. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2021.102118. URL https://www.sciencedirect.com/science/article/pii/S136184152100164X.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28.

Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J. Anders, and Klaus-Robert Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278, 2021. doi: 10.1109/JPROC.2021.3060483.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017a.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *Proceedings of the 34th International Conference on Machine Learning*, 2017b.

Jason Super et al. Explaining polysemantic neurons in language models via sparse autoencoder feature attribution. *arXiv preprint arXiv:2305.11867*, 2023.

Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models. https://ai.meta.com/llama/, 2023.

Bas H.M. van der Velden, Hugo J. Kuijf, Kenneth G.A. Gilhuijs, and Max A. Viergever. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, 79:102470, 2022. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2022.102470. URL https://www.sciencedirect.com/science/article/pii/S1361841522001177.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.