Efficient Transcoder Adaptation for Fine-Tuned Models: Revealing Medical Reasoning Mechanisms in Large Language Models

Zhouxing Tan[1], Hanlin Xue[2], Yulong Wan[1], Ruochong Xiong[1], Xu Chu[2], Xiang Li[3], Junfei Liu[1]*

[1]National Engineering Research Center for Software Engineering, Peking University
[2]School of Software and Microelectronics, Peking University
[3]College of Electrical and Information Engineering, Northeast Agricultural University
{tzhx, liujunfei}@pku.edu.cn, {2301210487, 2401110755, 2501110753, chuxu}@stu.pku.edu.cn, A20230133@neau.edu.cn

Abstract

Large language models (LLMs) suffer from a lack of decision-making transparency, limiting their deployment in high-stakes domains such as healthcare. We propose a mechanistic interpretability framework that introduces two novel paradigms: Medical Fine-Tuning with Frozen Attention Layers (FTFA) and Posterior Adaptation Transcoders (PAT). FTFA freezes attention layers while fine-tuning only feed-forward network (FFN) parameters, enabling PAT to efficiently adapt pre-trained transcoders on the same data. This approach achieves over 1000× efficiency improvement compared to training transcoders from scratch. We theoretically justify this methodology and demonstrate its cost-effectiveness for cross-domain transfer. Transcoders are sparse autoencoders that replace MLP layers to provide interpretable feature representations. By substituting MLP layers of both base Gemma2-2b and its medical fine-tuned variant with per-layer transcoders, we enable feature-level attribution analysis. Through systematic pruning and node merging of resulting attribution graphs, we construct human-interpretable decision pathways. Our analysis reveals that LLMs employ two parallel mechanisms for medical diagnosis: pattern matching and multi-hop reasoning, with fine-tuned models demonstrating enhanced correct reasoning patterns. This work provides a practical framework for training transcoders on fine-tuned models at minimal cost, enabling broader application of mechanistic interpretability across domains and potentially guiding model training through transcoder-based analysis.

Code — https://github.com/ColinNeverLand/llm-medical-reasoning-mechanisms

Introduction

Large language models (LLMs) have demonstrated strong performance in medical diagnosis tasks, leading researchers to investigate their potential for clinical decision support (McDuff et al. 2025; Goh et al. 2024). However, the deployment of AI systems in healthcare faces unique interpretability requirements due to the high-stakes nature of medical decisions. Healthcare professionals need to understand model reasoning processes to make informed judgments about when to trust or override AI recommendations (Reese et al. 2024). While chain-of-thought prompting has been proposed as a method to understand model reasoning (Savage et al. 2024), recent studies indicate that models' verbalized reasoning often differs from their actual internal computations (Turpin et al. 2023; Arcuschin et al. 2025; Lindsey et al. 2025; Barez et al. 2025). This gap between external explanations and internal processes poses significant challenges for building trustworthy medical AI systems, highlighting the need for methods that can directly analyze the internal mechanisms of LLMs in medical contexts.

Recent advances in mechanistic interpretability have demonstrated remarkable capabilities in revealing models' internal mechanisms (Dunefsky, Chlenski, and Nanda 2024; Lindsey et al. 2025; Ameisen et al. 2025). Fig. 1 illustrates this potential using the Gemma2-2b base model with corresponding per-layer transcoders (Lieberum et al. 2024).

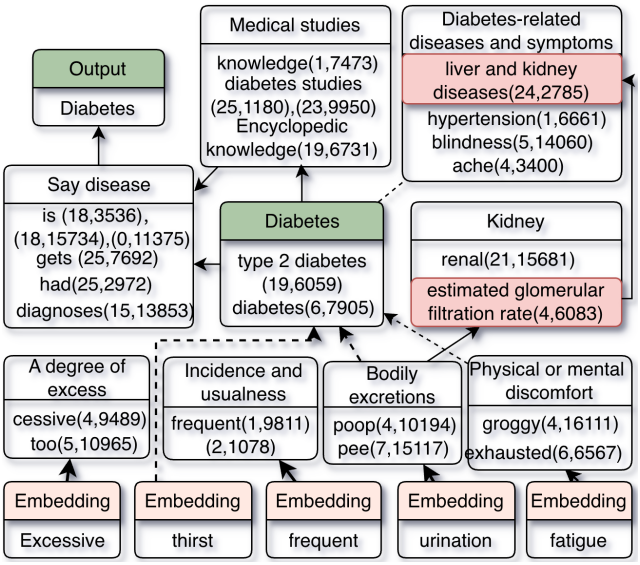


Figure 1: Attribution graph of the Gemma2-2b base model using Neuronpedia per-layer transcoders on a diabetes example. The model successfully predicts diabetes. The correct prediction primarily relies on pattern matching pathways (dashed lines) that directly associate symptoms with the disease.

*Corresponding author.

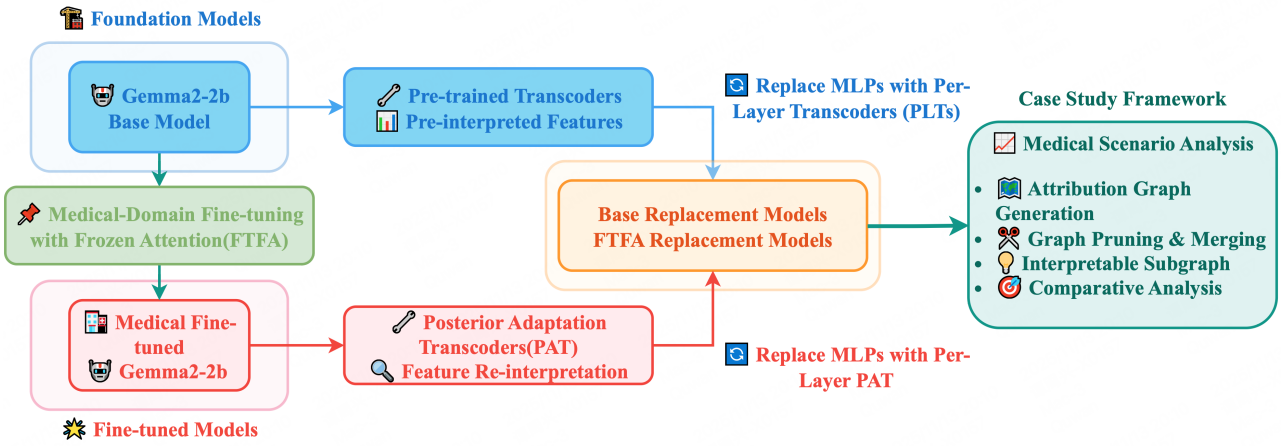


Figure 2: Our proposed FTFA and PAT enable efficient adaptation of transcoders for interpretable medical adaptation.

When presented with symptom descriptions, the model correctly outputs *diabetes* as the diagnosis. The resulting attribution graph reveals the model's complex internal reasoning process: dashed lines indicate pattern matching pathways where the model directly associates symptoms with diagnoses, while solid paths show multi-hop reasoning chains that connect symptoms through intermediate medical concepts. Notably, the analysis uncovers both correct reasoning pathways (shown in green) and erroneous reasoning attempts (shown in red), providing unprecedented insight into how the model arrives at its medical conclusions.

Despite these promising results, existing mechanistic interpretability methods face practical limitations when applied to domain-specific fine-tuned models. Training transcoders from scratch for each fine-tuned model is computationally expensive, while directly applying base model transcoders to fine-tuned models may not capture the nuanced changes introduced by domain adaptation (Hanna et al. 2025). This creates a significant gap between the need for interpretable fine-tuned models and the practical constraints of computational resources.

We address this challenge by introducing Medical Fine-Tuning with Frozen Attention Layers (FTFA) and Posterior Adaptation Transcoders (PAT), two paradigms that enable cost-effective interpretability analysis of fine-tuned models. As shown in Fig. 2, FTFA freezes attention layers while fine-tuning only feed-forward network (FFN) parameters, enabling PAT to efficiently adapt corresponding transcoders on the same data. We provide theoretical justification for this approach, demonstrating that it offers a computationally efficient alternative to training transcoders from scratch (Dunefsky, Chlenski, and Nanda 2024; Ameisen et al. 2025) while achieving over 1000× efficiency improvement and maintaining the ability to capture domain-specific adaptations.

Using this framework, we conduct a comprehensive analysis of medical reasoning in both base and medical fine-tuned Gemma2-2b models. By substituting MLP layers with interpretable transcoders and applying attribution graph analysis, we uncover the internal mechanisms underlying medical diagnosis. Our findings reveal two parallel reasoning pathways and show how medical fine-tuning enhances correct reasoning patterns while improving overall diagnostic performance.

Our main contributions are: (1) We introduce FTFA and PAT, achieving over 1000× efficiency improvement in transcoders training for fine-tuned models with theoretical justification. (2) We discover dual reasoning mechanisms in base models and reveal how fine-tuning transforms these mechanisms alongside parameter shifts and feature distribution changes. (3) We provide a practical framework enabling cross-domain mechanistic interpretability analysis.

This work bridges the gap between mechanistic interpretability and practical fine-tuned model analysis, offering a scalable approach for understanding and improving LLM behavior across domains.

Related Work

Large Language Models in Medical Applications

In medical domains, LLMs have shown promise across various tasks (Thirunavukarasu et al. 2023; Webster 2023), including medical question answering (Chen et al. 2025, 2024), clinical note analysis (Huang et al. 2024), and diagnostic assistance (Yang et al. 2024; Yin et al. 2025). Models such as Med-PaLM (Tu et al. 2024) and ChatDoctor (Li et al. 2023) have achieved competitive performance on medical licensing examinations and clinical reasoning benchmarks, suggesting their potential for real-world medical applications.

However, medical decisions require not only high accuracy but also explainable reasoning processes that clinicians can validate and trust (Savage et al. 2024; Liang, Li, and Jiang 2025). Studies have highlighted several concerns including hallucination in medical contexts (Masanneck, Meuth, and Pawlitzki 2025), inconsistent performance across different medical specialties, and sensitivity to prompt variations that could lead to different diagnostic conclusions (Wang et al. 2024).

These challenges highlight that current evaluation approaches focusing solely on accuracy are insufficient for

medical applications. Healthcare practitioners need to understand not just what the model predicts, but how it arrives at its conclusions. The gap between model performance and interpretability remains a critical barrier to adoption in medical applications, motivating the development of methods that can reveal and validate the internal reasoning processes underlying medical AI decisions.

Mechanistic Interpretability of Neural Networks

Mechanistic interpretability aims to understand neural network computations by decomposing model behavior into human-interpretable components and describing how these components interact to produce outputs. Early successes in computer vision demonstrated that neurons in convolutional networks often correspond to interpretable visual concepts (Cammarata et al. 2020). However, this neuron-centric approach faces significant limitations when applied to language models, where individual neurons frequently exhibit polysemanticity (Elhage et al. 2022). This phenomenon is known as superposition, where models represent more concepts than neurons by distributing concept representations across multiple neurons in a complex, overlapping manner.

Recent advances in sparse coding have offered promising solutions to superposition. Sparse autoencoders (SAEs) have emerged as a powerful tool for decomposing neural activations into sparse, interpretable features (Cunningham et al. 2023). These methods learn to represent model activations as linear combinations of a larger number of sparsely active components, many of which correspond to human-interpretable concepts. Transcoders further enable analysis of feature attribution relationships, allowing researchers to construct attribution graphs that reveal internal model mechanisms (Dunefsky, Chlenski, and Nanda 2024).

However, most existing work has focused on base models, with limited exploration of how fine-tuning affects these internal mechanisms. The application of mechanistic interpretability to fine-tuned models presents unique challenges. Training transcoders from scratch for each fine-tuned variant is computationally prohibitive, while applying base model features may miss important changes introduced by domain adaptation. This gap motivates the development of efficient methods for adapting interpretability tools to fine-tuned models.

Training Transcoders on Fine-Tuned Models at Minimal Cost

In large-scale pretrained language models, transcoders are commonly used to replace the MLP submodules within Transformer layers, leveraging sparse structures to enhance interpretability and computational efficiency. Taking the Gemma2-2B model as an example, each hidden layer has a size of 2K. To capture the polysemy of neurons more effectively, transcoders typically expand the representation space to 16K dimensions, enabling improved sparse expressiveness and semantic disentanglement. One transcoder is trained to replace the original MLP layer using the following objective:

$$L_{\text{TC}} = \|\text{MLP}(x) - \text{TC}(x)\|_2^2 + \lambda_1 \|\theta_{\text{TC}}(x)\|_1 \qquad (1)$$

Dataset	Count	Task Type
MEDQA	10178	Multiple-Choice
Symptom2Disease	853	Question Answering
DDXPlus	13310	Question Answering

Table 1: Summary of medical datasets used for FTFA.

where $\text{TC}(x)$ denotes the output of the transcoder given input x, and $\theta_{\text{TC}}(x)$ represents the sparse activation coefficients. The first term ensures output consistency with the original MLP, while the second term promotes sparsity.

However, training transcoders is extremely costly, involving backpropagation through large sparse encoders and requiring compute resources comparable to pretraining. When researchers attempt to analyze fine-tuned models using transcoders, they face the prohibitive cost of retraining from scratch. To address this challenge, we propose a novel adaptation strategy that enables efficient transcoders training for fine-tuned models at minimal cost. We provide systematic analysis of the necessary conditions and rigorous theoretical justification for this approach, laying a solid foundation for low-cost interpretability studies of LLMs.

Medical Fine-Tuning with Frozen Attention Layers

We adopt a Frozen Attention Fine-tuning strategy (FTFA) to adapt the Gemma2-2B language model for medical tasks. In large language models, attention modules typically handle information selection and relational modeling, while MLP layers focus on semantic transformation and decision-making. To isolate the contribution of MLP submodules and simplify the sources of representational change during fine-tuning, we freeze all attention components within Transformer layers and update only the MLP parameters.

This design serves two purposes: (1) it preserves the model's original attention behavior, enabling controlled analysis of MLP replacements such as transcoders; and (2) it provides the theoretical basis for the low-cost posterior transcoders training introduced later.

This setup aligns naturally with our transcoders framework. Specifically, we initialize each transcoders TC_1 with the parameters of its corresponding fine-tuned MLP module MLP_1, and connect it to the input/output interface between adjacent attention layers. The transcoders thus replaces the MLP computation while maintaining the original attention flow. Training is then performed on dataset D using the following objective:

$$\mathcal{L}(\theta_{\text{MLP}}) = E_{(x,y)\sim\mathcal{D}}[\ell(f(x; \theta^0_{\text{Attn}}, \theta_{\text{MLP}}), y)] \qquad (2)$$

where $f(x; \theta^0_{Attn}, \theta_{MLP})$ denotes the model with frozen attention parameters θ^0_{Attn} and trainable MLP parameters θ_{MLP}. The loss function $\ell(\cdot, \cdot)$ is a standard supervised learning objective, such as cross-entropy. The training dataset $\mathcal{D} = (x_i, y_i)_{i=1}^N$ consists of user inputs and corresponding targets , as illustrated in Tab. 1.

This design simplifies the optimization process and ensures that representational changes arise solely from MLP modules, which is essential for faithful transcoder analysis.

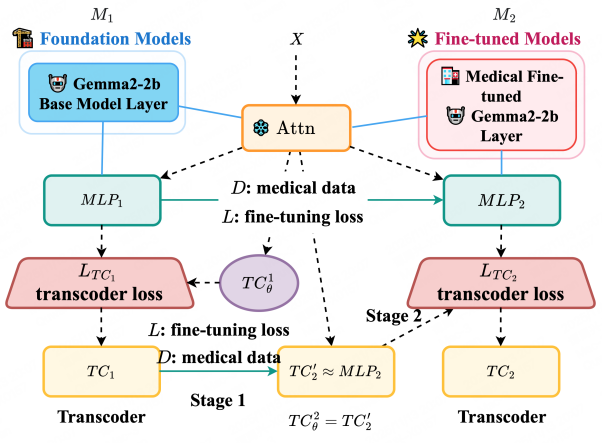


Figure 3: A minimal-cost method for training transcoders for fine-tuned models.

Since methods like transcoders and Sparse Autoencoders (SAEs) primarily target the interpretability of MLP components rather than attention, freezing attention layers is both natural and necessary.

Training Transcoders via PAT

As illustrated in Fig. 3, let TC_1 and TC_2 denote the transcoders corresponding to models M_1 and M_2, respectively. Suppose M_2 is obtained by fine-tuning only the MLP module of M_1 on dataset D, while keeping the Attention module frozen. Our key observation is that TC_2 can be approximated by TC_2', whose parameters are denoted as θ', and which is obtained by further fine-tuning TC_1 on the same dataset D (as shown in Tab. 1): $\theta' = \arg\min_{\theta_{TC}} E_{x\sim D}[L((f_{\theta_{TC}}(x), y)]$, with $\theta_{TC}^{(0)} = \theta_{TC_1}$ By the design of transcoder, which is used to approximate the output of an MLP module, it suffices to demonstrate:

$$\lim_{m\to\infty} \mathrm{MLP}_2(x) = \mathrm{TC}_2'(x) \qquad \text{a.e.} \tag{3}$$

where $m = \min\{m_{MLP_2}, m_{TC_2'}\}$ denotes the hidden dimension. We now prove Equation (3). According to the Neural Tangent Kernel (NTK) theory (Jacot, Gabriel, and Hongler 2018), the output trajectory $f_\theta(x)$ of a sufficiently wide neural network during training can be described by:

$$f_\theta^{(t)}(x) = f_\theta^{(0)}(x) - \eta \sum_{s=0}^{t-1} K_\theta(x, X)(f_\theta^{(s)}(X) - Y) \tag{4}$$

here, $K_\theta(x, x') := \nabla_\theta f(x)^T \nabla_\theta f(x')$ denotes the NTK, and $f^{(s)}(X)$ is the prediction on the training data at step s. When the network hidden dimension $m \to \infty$, the NTK converges to a deterministic kernel: $\lim_{m\to\infty} K_\theta^{(m)}(x, x') = K^\infty(x, x')$. For networks with identical architecture and activation functions , the limiting NTK is the same regardless of minor initialization differences. Both the MLP and the transcoder share the same architecture $f(x) = W_1 ReLU(W_2 x + b_2) + b_1$ and the same activation function ReLU. Given this, and that the fine-tuning objective of the transcoder is to match the output of the updated MLP Equation (1), and their initial outputs are almost everywhere equal. Therefore, Equation (4) provides the justification for Equation (3).

Note that in practice, both MLP and TC are finite-width networks, and the sparsity constraint $||\theta_{TC}||_1$ imposes additional optimization difficulty.Therefore, further refinement of TC via supervised fine-tuning on D_{sample}, based on the original transcoder training loss Equation (1), constitutes the stage 2 optimization, where D_{sample} is a sampled subset from the larger dataset required for transcoder training.

Feature Re-interpretation

To understand the semantic meaning of transcoder features, we employ a comprehensive interpretation framework that combines activation pattern analysis with direct logit attribution. Our interpretation process begins by analyzing how each feature directly influences the model's output predictions through the unembedding layer using Direct Logit Attribution (DLA) (Lin 2023). Complementing the output analysis, we examine the feature's activation patterns across diverse textual inputs from multiple datasets. For improved readability, we merge semantically similar features into super-nodes, ultimately constructing attribution graphs that reveal the model's reasoning pathways. Our entire pipeline maintains consistency with Neuronpedia's methodology, with detailed descriptions and examples provided in our extended version.

Experiments

This section aims to answer the following research questions through quantitative experiments and qualitative case studies.

Q1: Is FTFA effective for medical diagnosis tasks?
Q2: How does FTFA reshape internal representations?
Q3: How efficient and accurate is PAT versus baselines?
Q4: How does FTFA change internal mechanisms?

Performance of FTFA(Q1)

Model	Accuracy	BERTScore
Gemma-2-2b-FTFA	52.27 (**+ 1.48**)	56.36 (**+ 5.15**)
Gemma-2-2b-SFT	50.06 (**- 0.73**)	54.18 (**+ 2.97**)
Gemma-2-2b	50.79	51.21
Llama-3.2-3B	60.5	60.19
Llama-3.1-8B	65.55	61.36
Qwen2.5-7B-Instruct	**68.72**	**63.51**
II-Medical-8B	66.45	60.26
MedFound-7B	63.29	60.15

Table 2: Performance comparison across finetuned models, base models, and clinical LLMs on the medical diagnosis benchmark. Bold values indicate highest scores per metric.

To verify the effectiveness of FTFA, we evaluate its downstream performance on a medical benchmark (Perets et al.

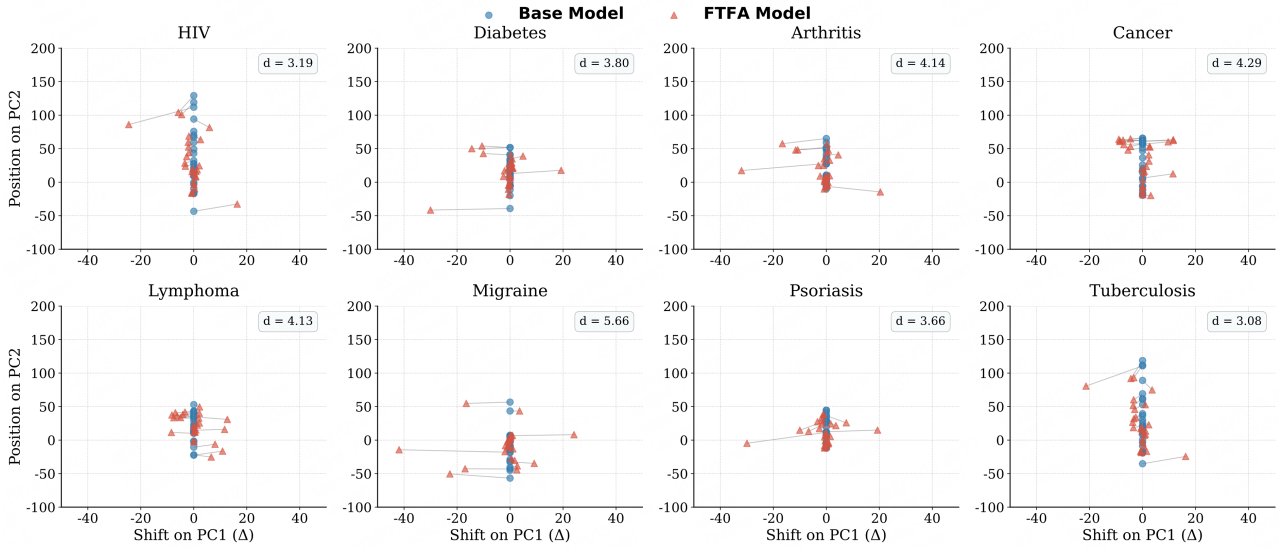


Figure 4: PCA projection of hidden activations before and after FTFA across eight diseases. Blue circles and red triangles denote base and FTFA model activations, respectively. Gray lines connect the same input, showing the shift direction. d indicates the average absolute shift along PC1.

2025). This experiment aims to demonstrate that FTFA preserves the model's reasoning ability and avoids performance collapse. Moreover, it establishes a controlled foundation for interpreting representational changes via transcoders by ensuring that all adaptations are confined to the MLPs.

Surprisingly, FTFA not only preserves but improves model performance, outperforming full-parameter fine-tuning (SFT) in both Accuracy and BERTScore (Tab. 2). FTFA improves accuracy by 1.48 points, while SFT slightly degrades it. For BERTScore, FTFA yields a +5.15 gain, surpassing the +2.97 from SFT. These results indicate that general-domain attention weights already capture useful priors for medical reasoning, and that tuning only MLP layers enables effective adaptation without disrupting core reasoning behavior.

Internal Representation Changes(Q2)

To examine how FTFA affects internal representations, we perform PCA shift analysis to capture global changes in the hidden activation space and use PAT to analyze the distribution of features.

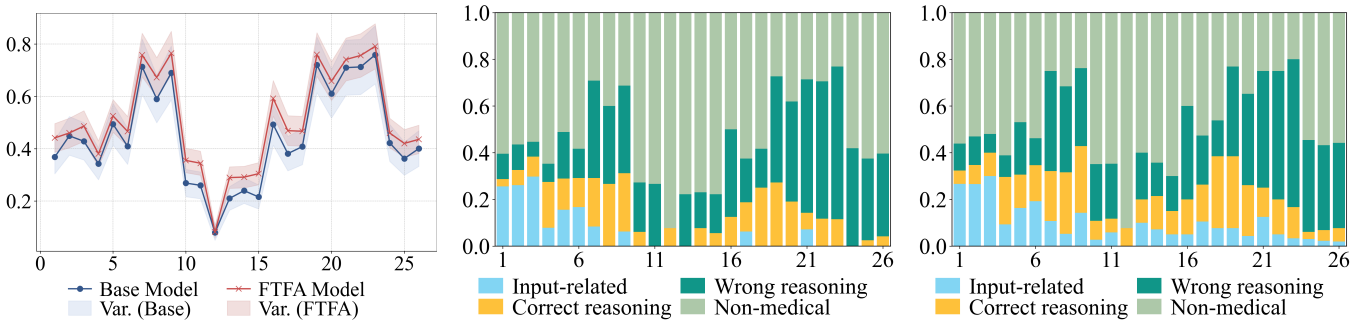
PCA shift analysis To measure latent shift, we apply PCA (n=2) to hidden states from each layer of M_1 and M_2, and compute layer-wise coordinates $(m_{i,1}, m_{i,2})$ (Huan et al. 2025; Xu et al. 2025). The PC_1 shift is defined as $\Delta m_{i,1} = m_{i,1}^{(2)} - m_{i,1}^{(1)}$, while the PC_2 $m_{i,2}^{(*)}$ serves as an auxiliary indicator of distributional change. We then calculate the global shift as $d = \left\| \mathbf{z}^{(2)} - \mathbf{z}^{(1)} \right\|_2$, where $\mathbf{z}^{(*)}$ is the mean of PCA coordinates across all layers.

As shown in Fig. 4, we observe that common diseases like diabetes and cancer show relatively small shifts, suggesting limited adaptation due to sufficient pretraining exposure. In contrast, less represented conditions such as HIV, Migraine, and Psoriasis exhibit larger shifts, indicating that FTFA leads to substantial reconfiguration of the model's internal parameters to accommodate new reasoning patterns. These results reveal that FTFA effects are disease-dependent, reinforcing the need for task-aware adaptation in medical LLMs.

Feature Distribution Analysis To examine how medical FTFA reshapes internal representations, we analyze the activation of medical-related features across all Transformer layers. We categorize features into four types: Input-related, which are associated with input token embeddings and typically appear in lower layers; Correct reasoning, found along correct inference paths in attribution graphs; Wrong reasoning, which contribute to incorrect reasoning chains; and Non-medical, referring to features unrelated to medical semantics. For large-scale identification, input-related and non-medical features are detected automatically using a trained classifier on 1,000 medical diagnostic queries. In contrast, correct and wrong reasoning features are manually annotated via attribution graph analysis on a curated set of 10 representative samples.

As shown in Fig. 5a, medical feature activation follows a U-shaped pattern: early (1–8) and late layers (20–25) show strong activation, while middle layers (10–15) exhibit reduced, more variable activation indicative of general-purpose processing. FTFA increases medical feature activation across the board, with notable gains in Layers 7–9, 16, and 19–23, suggesting enhanced alignment with medical semantics.

To assess reasoning quality, we further categorize features into input-related, correct reasoning, and wrong reasoning. As shown in Fig. 5b and Fig. 5c, the FTFA model activates more input-related and correct reasoning features, especially in lower and middle layers, while reducing non-medical and wrong reasoning features. This indicates that



(a) Layer-wise medical feature proportions for Base (blue) and FTFA (red) models.

(b) Distribution of categorized features via PLTs in the base model.

(c) Distribution of categorized features via PAT in the FTFA model.

Figure 5: Analysis of internal feature dynamics via Posterior Adaptation Transcoders (PLTs).

Figure 6: Comparison of Top-1 error rates between PAT and other per-layer transcoders.

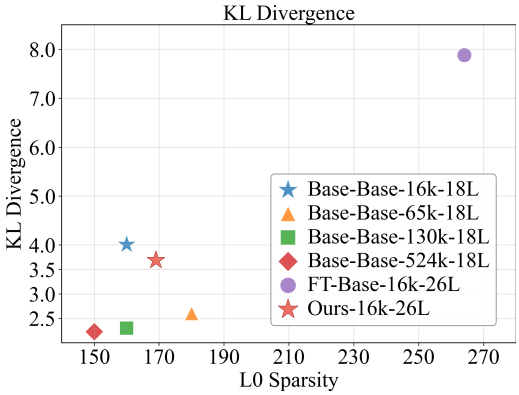


Figure 7: Comparison of KL divergence between PAT and other per-layer transcoders.

FTFA improves both the quantity and quality of medical feature activation.

Accuracy and Efficiency of PAT(Q3)

Fig. 6 and Fig. 7 present a comprehensive comparison across different transcoder configurations. All baseline *base-base model* results are from Anthropic (Ameisen et al. 2025), who trained an 18-layer model with corresponding per-layer transcoders ranging from 16k to 524k features per layer to construct replacement models. *FT-Base-16k-26L* represents fine-tuned Gemma2-2b directly paired with pre-trained base model transcoders, while *Ours-16k-26L* represents our FTFA fine-tuned Gemma2-2b with PAT.

The experimental results reveal several critical insights. First, directly applying pre-trained transcoders from the base model to construct replacement models for fine-tuned models (purple circles) yields nearly zero Top-1 accuracy and poor KL divergence. Such an approach would introduce significant bias in attribution analysis, failing to capture the true internal mechanisms of fine-tuned models. This finding underscores the necessity of our proposed paradigm.

Second, our PAT method (red stars) achieves comparable or even superior performance to from-scratch training approach (blue stars) while dramatically reducing computational requirements. Both methods use equivalent model architectures and transcoder widths (16,384 features per layer), yet where traditional transcoders training requires approximately 1B training tokens and 210 H100 hours, our approach accomplishes similar results with only 1M tokens and 0.2 H100 hours. This represents over 1000× efficiency improvement in both data and computational resources. Our method demonstrates that efficient adaptation can match the performance of costly retraining from scratch.

Furthermore, our theoretical analysis indicates that the effectiveness of our method scales favorably with model width, suggesting even greater advantages for larger models. The promising results achieved with relatively narrow transcoders validate the efficacy of our approach and demonstrate its potential to establish a new efficient paradigm for mechanistic interpretability that can be transferred across domains.

Reasoning Pathways: A Case Study(Q4)

We analyze the internal reasoning dynamics of Gemma2-2B before and after FTFA by constructing attribution graphs using our framework. The transcoders decode hidden representations into semantic features and infers weighted edges between them, where thickness reflects contribution strength.

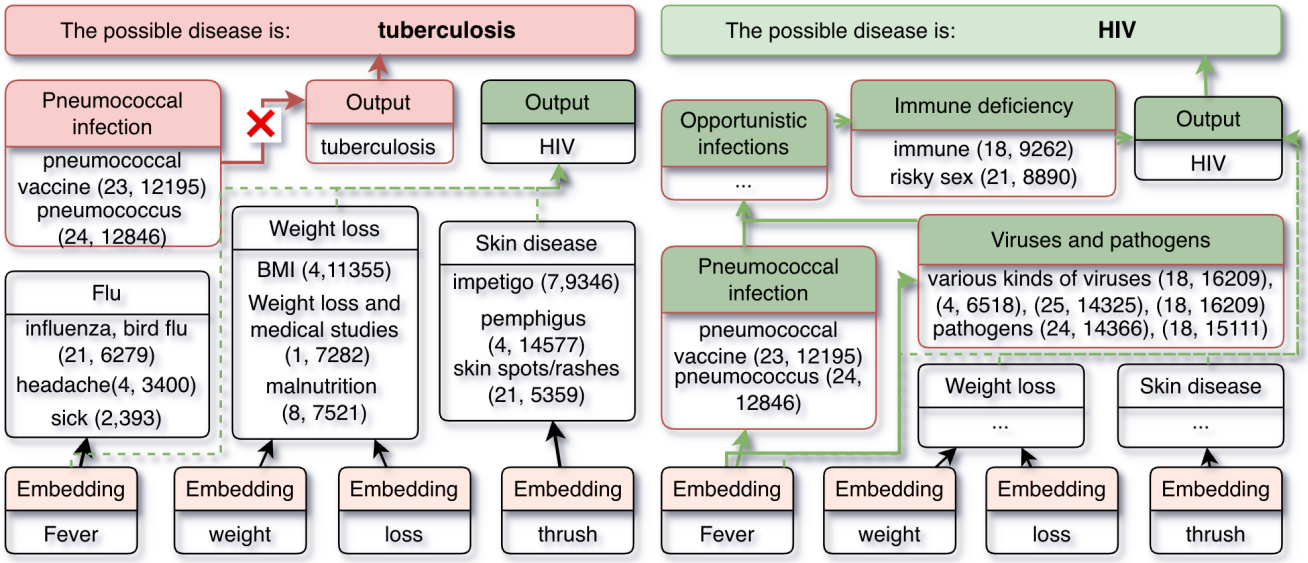


Figure 8: Attribution graphs of Gemma2-2B before (left) and after (right) FTFA on medical diagnostic data.

Solid edges represent causal reasoning, dashed edges indicate pattern matching, with red and green denoting incorrect and correct reasoning paths, respectively.

The base model (Fig. 8, left panel) exhibits two distinct reasoning mechanisms: **pattern matching** through shallow co-occurrence features (*fever*, *weight loss*, *thrush*) directly linked to HIV via dashed lines, and **multi-hop reasoning** following chains like *fever* $\rightarrow$ *flu* $\rightarrow$ *pneumococcal infection* $\rightarrow$ *tuberculosis*. Although this causal chain appears coherent, it leads to incorrect predictions since tuberculosis is caused by *Mycobacterium tuberculosis*, not pneumococcus, making this a false reasoning route (highlighted in red). Despite having correct pattern-matching paths available, the model assigns greater weight to flawed multi-hop reasoning.

After medical FTFA (Fig. 8, right panel), the model demonstrates markedly improved causal reasoning. It consolidates viral and pathogenic features into a *Viruses and Pathogens* supernode, triggering a clinically coherent pathway: *Viruses and Pathogens* $\rightarrow$ *Opportunistic Infections* $\rightarrow$ *Immune Deficiency* $\rightarrow$ *HIV*. This reflects accurate clinical knowledge where HIV compromises immunity, increasing susceptibility to opportunistic infections. The FTFA model constructs medically sound causal pathways and correctly predicts HIV with highest confidence, demonstrating the model's transformation from superficial pattern matching toward genuine causal understanding.

Limitations

Case Study Dependency. Due to the difficulty of analyzing global weights (Ameisen et al. 2025), our mechanistic insights rely primarily on case studies. While we present carefully selected representative cases and provide additional examples on our extended version, this approach limits the scalability of our analysis method.

Model Scale and Generalizability. We have not extensively validated whether the reasoning mechanisms discovered in Gemma2-2b generalize to larger models or different architectures. However, recent exciting research (Lindsey et al. 2025) strongly suggests that mechanisms found in smaller models persist in larger ones, providing optimism for broader applicability.

Theoretical Constraints. Our NTK-based theoretical framework assumes infinite-width networks (Jacot, Gabriel, and Hongler 2018), while practical implementations involve finite-width networks with sparsity constraints that may deviate from theoretical predictions.

Conclusion

This work addresses the critical challenge of mechanistic interpretability for fine-tuned language models by introducing FTFA and PAT, two paradigms that achieve over 1000× efficiency improvement compared to training transcoders from scratch. Our theoretical analysis provides rigorous justification for this approach.

Through comprehensive analysis, we reveal that LLMs employ dual reasoning mechanisms (pattern matching and multi-hop reasoning), with FTFA enhancing causal reasoning quality. Our attribution graphs provide unprecedented visibility into how domain adaptation reshapes internal reasoning processes, not just outputs.

By making mechanistic interpretability accessible for fine-tuned models, this framework enables broader understanding of model behavior across domains. This capability is crucial for high-stakes applications where transparency and reliability are paramount. Our work may help establish mechanistic interpretability as an integral component of the model development lifecycle, potentially bridging the gap between interpretability and practical AI deployment.

Ethics Statement

This research contributes to the field of mechanistic interpretability, with a specific focus on enhancing transparency and trustworthiness in Large Language Models (LLMs) applied to the medical domain. By utilizing transcoders to construct attribution graphs, our framework allows for the granular inspection of reasoning pathways—distinguishing between superficial pattern matching and genuine causal reasoning (as demonstrated in our HIV case study). This capability is ethically significant as it provides a means to identify and mitigate "Hallucinations" or plausible-sounding but medically incorrect logic (e.g., the incorrect tuberculosis causation chain) before models are deployed in high-stakes environments.

However, we acknowledge the risks associated with applying LLMs in healthcare. While our FTFA method improves the coherence of medical reasoning, the models analyzed (such as Gemma2-2B) are research prototypes and are not designed for clinical use. The "improved" reasoning pathways should not be interpreted as a substitute for professional medical diagnosis or advice. Furthermore, the test cases used in our experiments, including the reasoning pathways analyzed, were constructed based on general medical knowledge and do not contain any Personally Identifiable Information (PII) or private patient data. We believe our work supports the development of safer, more verifiable AI systems by moving beyond black-box predictions to auditable reasoning processes.

Acknowledgements

This work was supported by Guangzhou Quwan Network Technology Co., Ltd. under the cooperative project "AI Research Project on Emotional Companion Large Language Models".

References

Ameisen, E.; Lindsey, J.; Pearce, A.; Gurnee, W.; Turner, N. L.; Chen, B.; Citro, C.; Abrahams, D.; Carter, S.; Hosmer, B.; Marcus, J.; Sklar, M.; Templeton, A.; Bricken, T.; McDougall, C.; Cunningham, H.; Henighan, T.; Jermyn, A.; Jones, A.; Persic, A.; Qi, Z.; Thompson, T. B.; Zimmerman, S.; Rivoire, K.; Conerly, T.; Olah, C.; and Batson, J. 2025. Circuit Tracing: Revealing Computational Graphs in Language Models. https://transformer-circuits.pub/2025/attribution-graphs/methods.html. Published by Anthropic.

Arcuschin, I.; Janiak, J.; Krzyzanowski, R.; Rajamanoharan, S.; Nanda, N.; and Conmy, A. 2025. Chain-of-thought reasoning in the wild is not always faithful. *arXiv preprint arXiv:2503.08679*.

Barez, F.; Wu, T.-Y.; Arcuschin, I.; Lan, M.; Wang, V.; Siegel, N.; Collignon, N.; Neo, C.; Lee, I.; Paren, A.; et al. 2025. Chain-of-thought is not explainability. *Preprint, alphaXiv*, v2.

Cammarata, N.; Carter, S.; Goh, G.; Olah, C.; Petrov, M.; Schubert, L.; Voss, C.; Egan, B.; and Lim, S. K. 2020. Thread: Circuits. *Distill*. Https://distill.pub/2020/circuits.

Chen, H.; Fang, Z.; Singla, Y.; and Dredze, M. 2025. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 3563–3599.

Chen, X.; Zhang, W.; Xu, P.; Zhao, Z.; Zheng, Y.; Shi, D.; and He, M. 2024. FFA-GPT: an automated pipeline for fundus fluorescein angiography interpretation and question-answer. *npj Digital Medicine*, 7(1): 111.

Cunningham, H.; Ewart, A.; Riggs, L.; Huben, R.; and Sharkey, L. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

Dunefsky, J.; Chlenski, P.; and Nanda, N. 2024. Transcoders find interpretable llm feature circuits. *Advances in Neural Information Processing Systems*, 37: 24375–24410.

Elhage, N.; Hume, T.; Olsson, C.; Schiefer, N.; Henighan, T.; Kravec, S.; Hatfield-Dodds, Z.; Lasenby, R.; Drain, D.; Chen, C.; Grosse, R.; McCandlish, S.; Kaplan, J.; Amodei, D.; Wattenberg, M.; and Olah, C. 2022. Toy Models of Superposition. https://transformer-circuits.pub/2022/toy_model/index.html. Published by Anthropic and collaborators.

Goh, E.; Gallo, R.; Hom, J.; Strong, E.; Weng, Y.; Kerman, H.; Cool, J. A.; Kanjee, Z.; Parsons, A. S.; Ahuja, N.; et al. 2024. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA network open*, 7(10): e2440969–e2440969.

Hanna, M.; Piotrowski, M.; Lindsey, J.; and Ameisen, E. 2025. circuit-tracer. https://github.com/safety-research/circuit-tracer. The first two authors contributed equally and are listed alphabetically.

Huan, M.; Li, Y.; Zheng, T.; Xu, X.; Kim, S.; Du, M.; Poovendran, R.; Neubig, G.; and Yue, X. 2025. Does Math Reasoning Improve General LLM Capabilities? Understanding Transferability of LLM Reasoning. *arXiv preprint arXiv:2507.00432*.

Huang, J.; Yang, D. M.; Rong, R.; Nezafati, K.; Treager, C.; Chi, Z.; Wang, S.; Cheng, X.; Guo, Y.; Klesse, L. J.; et al. 2024. A critical assessment of using ChatGPT for extracting structured data from clinical notes. *NPJ digital medicine*, 7(1): 106.

Jacot, A.; Gabriel, F.; and Hongler, C. 2018. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31.

Li, Y.; Li, Z.; Zhang, K.; Dan, R.; Jiang, S.; and Zhang, Y. 2023. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. *Cureus*, 15(6).

Liang, X.; Li, Z.; and Jiang, H. 2025. Enhancing Healthcare Recommendations: A Privacy-Protective and Interpretable Cross-Domain Framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 12139–12146.

Lieberum, T.; Rajamanoharan, S.; Conmy, A.; Smith, L.; Sonnerat, N.; Varma, V.; Kramár, J.; Dragan, A.; Shah, R.;

and Nanda, N. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *arXiv preprint arXiv:2408.05147*.

Lin, J. 2023. Neuronpedia: Interactive Reference and Tooling for Analyzing Neural Networks. Software available from neuronpedia.org.

Lindsey, J.; Gurnee, W.; Ameisen, E.; Chen, B.; Pearce, A.; Turner, N. L.; Citro, C.; Abrahams, D.; Carter, S.; Hosmer, B.; Marcus, J.; Sklar, M.; Templeton, A.; Bricken, T.; McDougall, C.; Cunningham, H.; Henighan, T.; Jermyn, A.; Jones, A.; Persic, A.; Qi, Z.; Thompson, T. B.; Zimmerman, S.; Rivoire, K.; Conerly, T.; Olah, C.; and Batson, J. 2025. On the Biology of a Large Language Model. https://transformer-circuits.pub/2025/attribution-graphs/biology.html. Published by Anthropic.

Masanneck, L.; Meuth, S. G.; and Pawlitzki, M. 2025. Evaluating base and retrieval augmented LLMs with document or online support for evidence based neurology. *npj Digital Medicine*, 8(1): 137.

McDuff, D.; Schaekermann, M.; Tu, T.; Palepu, A.; Wang, A.; Garrison, J.; Singhal, K.; Sharma, Y.; Azizi, S.; Kulkarni, K.; et al. 2025. Towards accurate differential diagnosis with large language models. *Nature*, 1–7.

Perets, O.; Shoham, O. B.; Grinberg, N.; and Rappoport, N. 2025. CUPCase: Clinically Uncommon Patient Cases and Diagnoses Dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 28293–28301.

Reese, J. T.; Danis, D.; Caufield, J. H.; Groza, T.; Casiraghi, E.; Valentini, G.; Mungall, C. J.; and Robinson, P. N. 2024. On the limitations of large language models in clinical diagnosis. *MedRxiv*, 2023–07.

Savage, T.; Nayak, A.; Gallo, R.; Rangan, E.; and Chen, J. H. 2024. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digital Medicine*, 7(1): 20.

Thirunavukarasu, A. J.; Ting, D. S. J.; Elangovan, K.; Gutierrez, L.; Tan, T. F.; and Ting, D. S. W. 2023. Large language models in medicine. *Nature medicine*, 29(8): 1930–1940.

Tu, T.; Azizi, S.; Driess, D.; Schaekermann, M.; Amin, M.; Chang, P.-C.; Carroll, A.; Lau, C.; Tanno, R.; Ktena, I.; et al. 2024. Towards generalist biomedical AI. *Nejm Ai*, 1(3): AIoa2300138.

Turpin, M.; Michael, J.; Perez, E.; and Bowman, S. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36: 74952–74965.

Wang, L.; Chen, X.; Deng, X.; Wen, H.; You, M.; Liu, W.; Li, Q.; and Li, J. 2024. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *NPJ digital medicine*, 7(1): 41.

Webster, P. 2023. Six ways large language models are changing healthcare. *nature medicine*, 29(12): 2969–2971.

Xu, X.; Yue, X.; Liu, Y.; Ye, Q.; Hu, H.; and Du, M. 2025. Unlearning Isn't Deletion: Investigating Reversibility of Machine Unlearning in LLMs. *arXiv preprint arXiv:2505.16831*.

Yang, B.; Jiang, S.; Xu, L.; Liu, K.; Li, H.; Xing, G.; Chen, H.; Jiang, X.; and Yan, Z. 2024. Drhouse: An llm-empowered diagnostic reasoning system through harnessing outcomes from sensor data and expert knowledge. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(4): 1–29.

Yin, C.; Li, F.; Zhang, S.; Wang, Z.; Shao, J.; Li, P.; Chen, J.; and Jiang, X. 2025. Mdd-5k: A new diagnostic conversation dataset for mental disorders synthesized via neuro-symbolic llm agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 25715–25723.