Lezak: Explaining Complex Neural Networks through a Neuropsychological Lens

Nima Mirnateghi[†], Sharjeel Tahir[†], Syed Mohammed Shamsul Islam[†‡§], Joseph M Barnby[†*], Syed Afaq Ali Shah[†],
[†] Centre for Artificial Intelligence and Machine Learning (CAIML), Edith Cowan University, Joondalup, Australia
[‡] Department of Computing and Information Systems, Daffodil International University, Dhaka, Bangladesh
[§] Department of Computer Science and Software Engineering, The University of Western Australia
[*] Institute for Psychiatry, Psychology, and Neuroscience, King's College London, UK
{n.mirnateghi, s.tahir, syed.islam, j.barnby, afaq.shah}@ecu.edu.au

***Abstract*—Large language models (LLMs) are now widely used in interactive dialogues, where behaviour depends on context and conversational dynamics. Despite their impressive success, they are often treated as black boxes, with limited understanding of the internal mechanisms that maintain a consistent behavioural stance across turns. Inspired from neuropsychological approaches to understanding the human brain, we develop Lezak as a general-purpose ablation-based algorithm to probe black-box dialogue models. Lezak, named after the famous neuropsychologist, uses sparse autoencoders (SAEs) to decompose model behaviour before performing target ablations while preserving conversational context. We apply Lezak to therapeutic dialogues as a proof-of-principle for our approach, where context depends on more than semantic correctness. We find that different aspects of therapeutic behaviour evolve across deeper layers. Results suggest that Lezak can be useful as a causal tool to identify mechanistic regions of interest within LLMs, with future directions to expand our case-study into a generalised tool.**

***Index Terms*—Mechanistic Interpretability, Cognitive AI, Therapeutic LLMs, Causal Interpretability**

I. Introduction

The past few years have seen rapid progress in large language models (LLMs), transforming them from task-oriented systems into interactive agents. As these systems move from benchmark settings to real-world use, their impact increasingly depends on behavioural properties expressed in interaction, including how they sustain tone, follow conversational norms, and adapt to user context over time. This creates a particular challenge for interpretability because such properties are rarely tied to a single component or representation [1]. Rather, language architectures are shaped by conversational history, and embedded across multiple layers. As a result, understanding model behaviour involves relating patterns of interaction to model's internal processes that span multiple levels of representation. A closely related distinction has long been central in neuroscience, where lesions are used as interventions to test which brain regions are necessary for particular behaviours [2]. This intervention process supports functional attribution by providing causal evidence about which components are necessary for specific behaviours [2].

Recent work in mechanistic interpretability has made substantial progress in connecting internal structure in LLMs to specific capabilities [3]. Studies have localized token-level behavior and identified internal circuits [3] using targeted interventions, such as activation patching and causal tracing [4]. These approaches are particularly effective when the target behavior is well defined, and tightly linked to particular inputs or outputs. Behavioural properties expressed in interaction, however, differ in important ways. They unfold across turns and are shaped by many small lexical and stylistic choices, which makes it challenging to map internal components to the model's output in context. This motivates interpretability approaches that connect internal structure to behavioural patterns across interaction instead of isolated outputs alone.

To address this gap, recent work emphasises targeted intervention to study internal functions in LLMs [5]–[7]. Intervention-based approaches go beyond treating internal activations as explanations by examining how ablating internal components causally changes model behaviour. This approach is well established in neuroscience, where lesions are used to test how disrupting specific structures affects behaviour and to support functional attribution [8]. This connection is also relevant in AI interpretability [9], where LLMs are often studied as models of language processing and interventions provide a direct way to test internal function.

Therapeutic dialogues provide a compelling setting for studying behaviour in interaction. Effective therapeutic responses depend not only on semantic relevance but also on subtle linguistic choices that convey tone, empathy, and appropriateness at all turns. Small changes in wording can alter how a response is perceived, even when its propositional content remains similar. This makes therapeutic conversation sensitive to disruptions that affect linguistic expression and subtext, rather than what is conveyed at a coarse semantic level. In this work, therapeutic dialogue is used to analyse model behaviour over a multi-turn context instead of single-step outputs.

Motivated by this, we propose Lezak, a framework for layer-wise, intervention-based analysis of behaviour in LLMs using sparse internal representations. Inspired by classical neuropsychological lesion studies, we name this framework Lezak, referencing the established tradition of neuropsychological assessment [10]. Lezak builds on recent advances in sparse autoencoders (SAEs) to identify interpretable features,

and uses targeted ablation of these features during generation of model response to examine their functional contribution. Unlike prior intervention studies that focus on single responses or narrow tasks, Lezak is designed to probe how internal mechanisms support behavioural patterns that unfold across extended interaction. In this work, therapeutic dialogues are used as a compelling setting, where small ablations can alter tone, coherence, and perceived appropriateness without necessarily changing semantic content. This setting allows a systematic examination of how different layers contribute to distinct aspects of behavioural expression under realistic conversational conditions.

The contributions of Lezak are as follows: (i) Lezak permits the analysis behaviour in LLMs across layers and interaction contexts, which is theoretically task and domain-agnostic. (ii) Lezak enables layer-wise causal analysis of therapeutic dialogue, demonstrating how sparse internal features contribute to behavioural properties beyond semantic correctness. (iii) We provide an explicit assesment of representational redundency, useful for making networks more efficient or testing how sparsity in networks can support more abstract features. We show this by conducting controlled feature ablations across multiple layers and evaluate the resulting behavioural changes over multi-turn context, and assess intervention effects using lexical, semantic, and confidence-based metrics to characterise changes in generated behaviour. The implementation of Lezak is available at: https://github.com/ai-voyage/Lezak.git .

II. Related Work

Large language models (LLMs) have achieved remarkable performance across a wide range of applications, from dialogue systems and assistive technologies to decision support and multimodal reasoning [11]. Despite these advances, the internal representational mechanisms of LLMs remain poorly understood. This has motivated increasing interest in interpretability methods toward analysing internal representations, such as intermediate activations, and their causal role in model behaviour [3].

A central challenge in this line of work is that neuron activations in LLMs are typically polysemantic. Individual neurons in large models often respond to several unrelated patterns, making it unreliable to treat single units as interpretable components [1]. As a result, recent work has moved away from neuron-level analysis toward the study of features. In this context, "features" are commonly treated as recurring activation patterns that reflect coherent internal functions extracted from dense representations [12]. This perspective recognises that meaningful structure in LLMs is often distributed across many dimensions.

Earlier approaches to analysing internal representations relied mainly on probing and attribution methods, which measure correlations between hidden states and linguistic properties [3]. Although these methods are useful for describing what information is present in a representation, they do not systematically establish whether particular components or sets of components are responsible for differences in model output. For this reason, recent research in mechanistic interpretability has focused on analysing internal computations directly, including how representations change across layers and how specific components contribute to downstream behaviour [13], [14].

SAEs have been proposed as a practical tool for addressing polysemanticity by decomposing dense residual stream activations into sparse, disentangled features [15]. By enforcing sparsity constraints, SAEs encourage representations in which individual latent units correspond more closely to monosemantic concepts, enabling finer-grained analysis of internal informational structure. Recent studies have demonstrated that SAEs can uncover meaningful abstractions across deep layers that are otherwise obscured in dense representations [3]. These advances position SAEs as a powerful mechanism for studying internal representations at scale, particularly when combined with layer-wise analysis.

In addition to representation analysis, several works have also explored causal interventions such as neuron ablation, feature suppression, and circuit-level manipulation to assess the functional importance of internal components [3]. Such intervention-based analyses provide stronger evidence of causal necessity than attribution or probing alone. However, existing studies [6], [7] have typically focused on narrow tasks, with limited analysis of how intervention effects vary across depth. Our approach applies systematic layer-wise causal interventions on sparse feature representations to analyse how internal components contribute to behaviour under context-dependent settings, such as therapeutic dialogue.

III. Lezak

Here we describe Lezak for identifying and causally testing internal representations, with a case-example of therapeutic dialogue generation. The framework combines residual-stream activation extraction, sparse representation learning, and targeted intervention during generation. By progressively ablating learned sparse features and measuring the resulting changes in model output, we aimed to distinguish correlative and causative features that maintain therapeutic quality (see Figure 1 for a summary).

A. Activation Extraction (Therapeutic Agent)

Let $\mathcal{M}$ denote the agent fine-tuned on a corpus of therapeutic dialogues $\mathcal{D} = \{(u_i, a_i)\}_{i=1}^{N}$, where u_i represents client utterances and a_i denotes therapist responses. The model consists of $L = 32$ transformer layers, each containing multi-head self-attention and feed-forward (MLP) sublayers with residual connections. We extract internal representations from the post-MLP residual stream at layer $\ell \in L$. For a sequence of tokens $X = \{x_1, \ldots, x_t\}$, a forward pass through the model yields hidden states $\mathbf{H}^{(\ell)} \in \mathbb{R}^{T \times d_{\text{model}}}$ at layer ℓ. Tokens belonging to the assistant (agent) responses are retained by identifying the assistant boundary in each sequence. The resulting activation matrix $\mathbf{A}^{(\ell)} \in \mathbb{R}^{N_\ell \times d_{\text{model}}}$ contains post-MLP residual activations for N_ℓ assistant tokens at layer ℓ. In addition, a mapping from each activation row

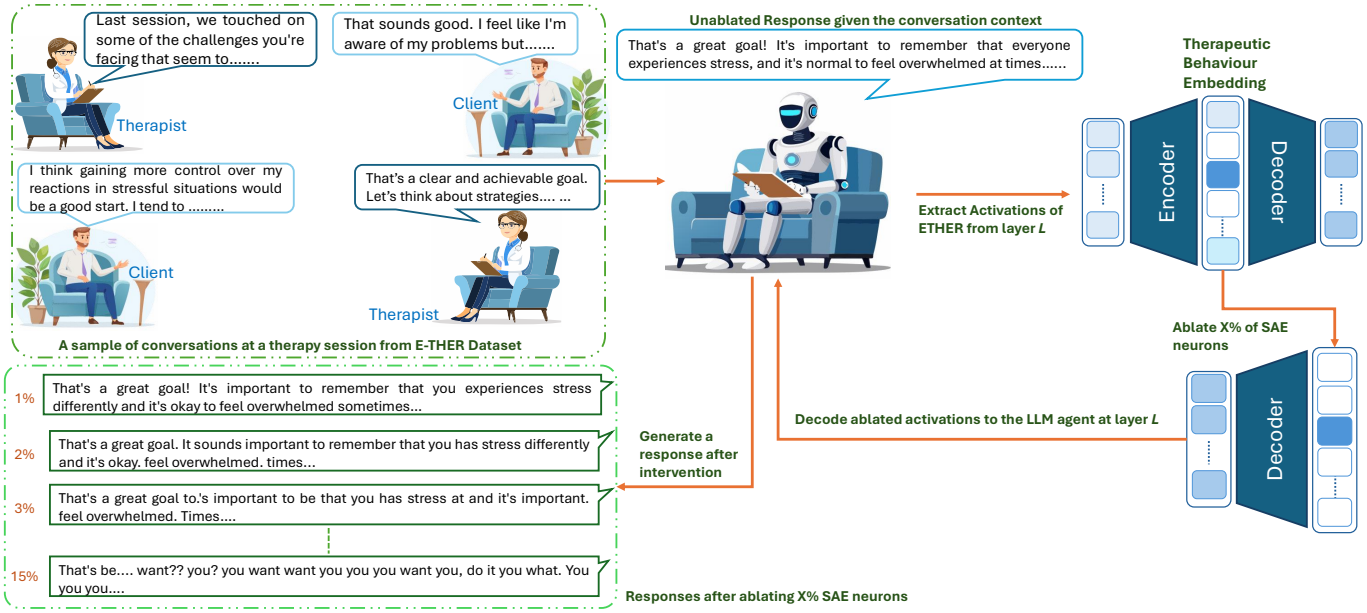


Fig. 1. Flow diagram of the proposed Lezak framework. Post-MLP residual stream activations are extracted from a fine-tuned therapeutic language model and encode them using Topk-SAE to obtain interpretable sparse features. Causal interventions are performed by ablating subsets of these features during generation, while preserving full multi-turn conversational context. The resulting responses are compared against human therapist references and the model's unablated baseline to analyze depth-dependent effects on therapeutic behavior.

to its original token index is recorded, enabling subsequent alignment between internal features and generated response text.

B. Sparse Feature Discovery Agent

We employ sparse dictionary learning [5] to uncover interpretable and causally testable internal features associated with therapeutic behaviour. Specifically, we train a Top-K SAE on residual stream activations extracted from the therapeutic agent. The objective is to learn an overcomplete, sparse representation in which each activation is reconstructed using a small, fixed number of latent features.

Let $\mathcal{A}^{(\ell)} = \{\mathbf{x}_j\}_{j=1}^{N_\ell}$ denote the set of post-MLP residual activation vectors corresponding to assistant tokens at layer $\ell \in \{1, \ldots, L\}$, where L denotes the total number of layers in $\mathcal{M}$. Each activation $\mathbf{x} \in \mathbb{R}^{d_{\text{model}}}$ represents a single token-level residual stream state. Here, d_{model} denotes the dimensionality of the model's residual stream. Prior to training, activations are normalised feature-wise using statistics computed over $\mathcal{A}^{(\ell)}$:

$$\tilde{\mathbf{x}} = \frac{\mathbf{x} - \boldsymbol{\mu}}{\boldsymbol{\sigma}}, \tag{1}$$

where $\boldsymbol{\mu}, \boldsymbol{\sigma} \in \mathbb{R}^{d_{\text{model}}}$ are the empirical mean and standard deviation. The encoder maps the normalised activation into an overcomplete latent space:

$$\mathbf{h} = \mathbf{W}_{\text{enc}}\tilde{\mathbf{x}} + \mathbf{b}_{\text{enc}}, \tag{2}$$

where $\mathbf{W}_{\text{enc}} \in \mathbb{R}^{d_{\text{hidden}} \times d_{\text{model}}}$ and $d_{\text{hidden}} = \alpha \cdot d_{\text{model}}$ with expansion factor $\alpha > 1$. To enforce structured sparsity, we apply a hard Top-K operator to the encoder activations. Let $\mathcal{I}_K$ denote the indices of the K largest-magnitude elements of $\mathbf{h}$:

$$\mathcal{I}_K = \text{TopK}(|\mathbf{h}|, K). \tag{3}$$

The sparse latent vector $\mathbf{z} \in \mathbb{R}^{d_{\text{hidden}}}$ is defined as:

$$z_i = \begin{cases} h_i, & i \in \mathcal{I}_K \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

which enforces an exact ℓ_0 sparsity constraint, ensuring that each token activation is represented using a fixed number of active features. The decoder reconstructs the original activation from the sparse latent code:

$$\hat{\mathbf{x}} = \mathbf{W}_{\text{dec}}\mathbf{z} + \mathbf{b}_{\text{dec}}, \tag{5}$$

where $\mathbf{W}_{\text{dec}} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{hidden}}}$. The Top-$K$ SAE is trained to minimise mean squared reconstruction error:

$$\mathcal{L}_{\text{SAE}} = \|\hat{\mathbf{x}} - \tilde{\mathbf{x}}\|_2^2. \tag{6}$$

We train separate SAEs independently for each layer $\ell \in L$.

C. Ablating Neurons

To establish the causal role of discovered sparse features in model behaviour, we conduct systematic intervention experiments through feature ablation. The procedure operates on full multi-turn conversations, where dialogue context is accumulated across turns to preserve conversational state. This ensures that interventions are evaluated under realistic, context-dependent generation conditions. For a layer ℓ, we record the layer-ℓ residual stream states associated with the generated assistant tokens. Let $\mathbf{R}^{(\ell)} = [\mathbf{r}_1^{(\ell)}, \ldots, \mathbf{r}_G^{(\ell)}]^\top \in \mathbb{R}^{G \times d_{\text{model}}}$ denote the sequence of residual vectors collected

for the G generated tokens, where G is the number of tokens generated in the assistant continuation. Using the trained Top-K SAE for layer ℓ, we encode $\mathbf{R}^{(\ell)}$ to obtain sparse latent activations:

$$\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_G]^\top \in \mathbb{R}^{G \times d_{\text{hidden}}}, \tag{7}$$

where each $\mathbf{z}_g$ contains at most K non-zero entries. We refer to each latent dimension of the SAE as a feature.

To intervene, we define an ablation set of feature indices $\mathcal{S} \subseteq \{1, \ldots, d_{\text{hidden}}\}$ and suppress these features by setting:

$$(\mathbf{z}'_g)_k = \begin{cases} 0, & k \in \mathcal{S}, \\ (\mathbf{z}_g)_k, & \text{otherwise}, \end{cases} \tag{8}$$
$$\text{for all } k \in \{1, \ldots, d_{\text{hidden}}\},\ g \in \{1, \ldots, G\}.$$

The ablated latent activations are decoded back into residual space to produce modified residual vectors $\mathbf{R}^{(\ell)\prime} \in \mathbb{R}^{G \times d_{\text{model}}}$. When generating responses with the intervention applied, we replace the layer-ℓ residual stream states for generated tokens with $\mathbf{R}^{(\ell)\prime}$, yielding an intervened response. By increasing $|\mathcal{S}|$ according to activation magnitude (percentile-based selection), we obtain a controlled family of outputs under progressively stronger feature ablations. This enables causal assessment of how sparse features influence therapeutic behaviour.

D. Evaluation Metrics

To quantify the behavioural effect of Lezak interventions, we generate an ablated response $y^{(p)}$ at ablation level p for each dialogue turn and compare it to the human reference $y^\star$. We report lexical and semantic similarity using BLEU-2 [16] and BERTScore F1 [17], and report perplexity to quantify deviation from the unablated baseline. Scores are averaged across turns for each p and analysed as a function of intervention strength; BLEU-2 captures stable local phrasing in open-ended dialogue, while BERTScore F1 provides a balanced semantic similarity measure. Together, these metrics provide a controlled proxy for therapeutic behaviour change under intervention and enable identification of causal neurons driving the observed shifts.

IV. Experimental Results

This section evaluates the causal contribution of sparse internal features to therapeutic dialogue generation using the Lezak framework. Inspired by lesion and ablation studies in neuroscience, Lezak systematically ablates neurons identified via Top-K SAE neurons and measures the resulting changes in model behaviour. By analysing degradation patterns across depth, we assess *where* and *how* internal representations support therapeutic competence. Following [14], we group layers into early (Layer 7), mid (Layers 15 and 23), and late (Layer 31) stages.

A. Dataset

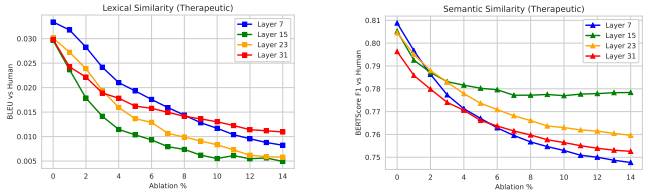
In our experiments, we used the E-THER (Empathic THERapy conversations) [18], which is a corpus of therapist–client interactions drawn from 18 sessions conducted by certified clinical psychologists with clients. The dataset comprises 789 dialogue pairs and 1,578 utterances. Our experiments are conducted on the text component of the dataset.

B. Experimental Setup

We use Llama3.1-8B [19] fine-tuned on E-THER [18] corpus as a baseline for our therapeutic agent. Following the experimental protocol of [18], we set the therapeutic agent's training hyperparameters as follows: learning rate $3e^{-2.5}$, batch size 16, 2 training epochs, and LoRA settings with rank 8 and $\alpha = 64$. We conducted a series of sweep experiments to select the best Sparse Feature Discovery Agent for each target layer. Specifically, we varied the learning rate over $1 \times 10^{-4}, 3 \times 10^{-4}$ and the expansion factor over $4, 8$, while keeping the batch size fixed at 128 and setting the sparsity coefficient to $\lambda = 1 \times 10^{-3}$. Each configuration was trained for up to 100 epochs with early stopping. We selected the SAE for each layer by maximizing cosine similarity to prioritize directional alignment between original and reconstructed activations while reducing sensitivity introduced by sparsity regularization. Based on these empirical results, we chose a learning rate of 1×10^{-4} across all layers. We selected an expansion factor of 4 for layers 7 and 31, achieving the highest cosine similarities of 0.9941 and 0.9947, respectively. For layers 15 and 23, we selected an expansion factor of 8, achieving cosine similarities of 0.9968 and 0.9966. We performed all our experiments using an NVIDIA RTX A6000 GPU.

C. Lezak Interpretability

Figure 2 presents the behavioural effect of interventions against human therapist references under progressive feature ablation. Across early, mid, and late layers, increasing ablation leads to a consistent decline in both BLEU and BERTScore. This pattern indicates that SAE features are causally linked to therapeutic response generation. Notably, BLEU degrades substantially faster than BERTScore across layers. This suggests that the ablated sparse features mainly support lexical and stylistic realisation, while semantic alignment remains comparatively more robust. Tables I and Table II quantify the behavioral impairment and reveal the following three key patterns.



(a) Lexical Similarity (b) Semantic

Fig. 2. Layer-wise effects of sparse feature ablation on LLM-human similarity across layers. For selected layers (early, mid, late), we report a) changes in lexical similarity (BLEU), b) semantic similarity (BERTScore) as increasing top-k features are ablated. We show that sparse features are more critical for surface-level therapeutic expression than for preserving semantic intent.

TABLE I

EFFECT OF ABLATION ON THERAPEUTIC QUALITY AGAINST HUMAN THERAPIST REFERENCES. VALUES ARE THE PERCENTAGE CHANGE BETWEEN THE UNABLATED CONDITION AND THE HIGHEST ABLATION LAYER. ARROWS INDICATE THE DIRECTION OF DEGRADATION.

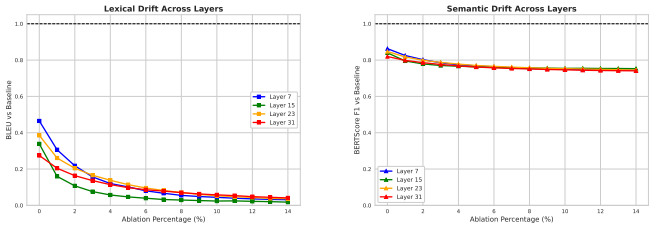
Layer	Δ BLEU ↓	Δ BERT ↓	Δ PPL
7	−75.4%	−7.6%	−73.1%
15	−83.4%	−3.3%	+641.8%
23	−80.8%	−5.6%	+245.0%
31	−63.3%	−5.5%	−25.5%

TABLE II

BEHAVIOURAL CHANGE RELATIVE TO THE UNABLATED OUTPUT UNDER ABLATION. VALUES ARE THE PERCENTAGE CHANGE IN BLEU AND BERTSCORE SIMILARITY TO THE MODEL'S UNABLATED OUTPUT AT THE HIGHEST ABLATION LAYER.

Layer	Δ BLEU ↓	Δ BERT ↓
7	−93.3%	−13.7%
15	−94.8%	−10.4%
23	−90.5%	−12.0%
31	−85.1%	−9.6%

First, Table I reports that lexical similarity to human therapist responses (BLEU) exhibits severe degradation across all layers, with endpoint reductions of approximately 63% to 83% at the highest ablation level relative to the unablated condition. This suggests that the ablated sparse features are critical in the surface-level realisation of therapeutic language. **Second**, semantic similarity (BERTScore) declines more moderately (≈ 3–8% across layers). This asymmetric degradation suggests that semantic intent of the model is partially preserved even when the therapeutic style is disrupted. **Third**, model confidence (perplexity ratio) responds non-uniformly across depth. Mid layers (Layers 15 and 23) show dramatic perplexity increases (up to +642%), whereas early (Layer 7) and late (Layer 31) layers exhibit smaller or even decreasing perplexity despite substantial quality loss. This depth-dependent dissociation suggests that model confidence is not uniformly aligned with therapeutic competence. These results indicate that the SAE neurons are causally necessary for therapeutic behaviour, but their functional contribution varies by layer. In the following sections, we analyze how Lezak enables layer-wise causal attribution.



(a) Lexical Similarity (b) Semantic Similarity

Fig. 3. Comparison of ablated outputs against the unablated model baseline: a) Lexical similarity (BLEU) and b) Semantic similarity (BERTScore) are computed relative to the model's original (non-ablated) output. Across all layers, ablation causes substantial divergence from the model's original output, with stronger degradation in lexical form than semantic content. This demonstrates that Lezak interventions induce systematic behavioural shifts beyond reduced alignment with human references.

Figure 3 quantifies model drift by measuring similarity between each ablated output and its corresponding 0% ablation output (i.e., the model's own baseline response for the same prompt). Across all depths, similarity to the model's own baseline response decreases sharply: BLEU drops by ≈85–95% from its 0% ablation, while BERTScore drops more moderately (≈10–14%). This demonstrates that Lezak lesions do not merely reduce match to a human reference; they induce a strong and systematic behavioural shift relative to the model's own unablated output therapeutic responses. The same asymmetry observed in human-reference evaluation reappears internally: lexical form is highly sensitive to sparse-feature removal, whereas semantic similarity is more resilient. In neuroscience terms, this resembles lesion effects where high-level intent can remain partially intact while the realisational form of behaviour deteriorates [2]. The following layer-wise analysis draws on both human-reference degradation (Table I) and behavioural change relative to the model's unablated output (Table II) to attribute functional roles across depth.

D. Lexical Fragility

Layer 7 exhibits strong degradation in lexical similarity to the human reference in Table I (BLEU -75%), alongside the largest semantic drop among the tested layers (BERTScore -7.6%). Table I also reports a substantial decrease in perplexity ratio (-73%), meaning the model becomes more confident despite degraded therapeutic alignment. In Table II, the same layer exhibits a large reduction in BLEU relative to the model's unablated output. This pattern suggests that early-layer sparse features support lexical and stylistic realisation of therapeutic phrasing patterns. Moreover, removing these features produces a substantial behavioural change without a corresponding increase in predictive uncertainty, consistent with a lesion-style dissociation between competence and confidence.

E. Therapeutic Coherence and Control

Layers 15 and 23 show the most pronounced confidence disruption in Table I with perplexity increasing sharply (+642% and +245%). In contrast, semantic similarity to the human reference declines only modestly (BERTScore −3.3% and −5.6%), while lexical similarity collapses (BLEU ≈−83% and −81%). This pattern suggests mid-layer sparse features are critical for stabilising the generation dynamics that support coherent therapeutic responses. Under ablation, the sharp rise in perplexity reflects loss of generative control, while preserved semantic similarity suggests that therapeutic intent remains but is expressed incoherently.

F. Therapeutic Tone

Layer 31 shows the smallest degradation in lexical similarity to the human reference in Table I (BLEU −63%), while still exhibiting a large reduction in BLEU relative to the model's

unablated output in Table II (-85%). The perplexity ratio decreases moderately (-25.5%), indicating that confidence is not a reliable proxy for therapeutic quality. This pattern suggests that late-layer sparse features primarily contribute to output-level shaping, such as wording choices and therapeutic tone, rather than upstream generation stability. Ablation of this layer alters the expressed form of therapeutic responses without producing the sharp uncertainty increases as observed in mid layers.

Lezak yields a depth-dependent lesion profile with clear dissociation: (i) lexical form is consistently fragile across depth, (ii) semantic similarity is comparatively resilient, and (iii) confidence disruptions concentrate in mid layers, while early and late layers can show reduced perplexity despite degraded therapeutic quality. These findings support the Lezak interpretation that TopK-SAE features are not only correlational but causally necessary sparse mechanisms whose functional contribution varies by depth. This mirrors neuroscience-style ablation studies, where specialisation is inferred from behavioural deficits and dissociation [2].

V. Conclusion

We introduced Lezak, a layer-wise intervention framework for interpreting LLM behaviour. Lezak ablates sparse internal features during multi-turn generation to test their causal role. When applied to therapeutic dialogue, Lezak uncovers a depth-dependent lesion profile with clear dissociation between and within layers. We find that the expressed therapeutic form, including wording and style, is fragile under ablation, while semantic similarity is comparatively more resilient. Furthermore, mid-layer lesions produce the strongest disruptions to generative control, whereas early and late layers can exhibit reduced perplexity despite degraded therapeutic alignment. Overall, the impairment profile suggests sparse features play causally necessary roles in shaping therapeutic behaviour.

The presented results establish Lezak in one therapeutic configuration, and broader generalisation remains to be tested. Future work should include evaluating additional LLM architectures and datasets. It can also assess alternative SAE variants, and extending interventions to other transformer components such as attention pathways.

Acknowledgments

This research was supported by Edith Cowan University. The authors would like to thank Professor David Suter for his insightful guidance throughout this research. The authors acknowledge the use of generative AI tools, such as ChatGPT by OpenAI for minor language editing and grammar refinement.

References

[1] N. Saphra and S. Wiegreffe, "Mechanistic?" in *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, Y. Belinkov, N. Kim, J. Jumelet, H. Mohebbi, A. Mueller, and H. Chen, Eds. Miami, Florida, US: Association for Computational Linguistics, Nov. 2024, pp. 480–498. [Online]. Available: https://aclanthology.org/2024.blackboxnlp-1.30/

[2] H.-O. Karnath, C. Sperber, and C. Rorden, "Reprint of: Mapping human brain lesions and their functional consequences," *NeuroImage*, vol. 190, pp. 4–13, 2019, mapping diseased brains. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S105381191930045X

[3] L. Bereska and S. Gavves, "Mechanistic interpretability for AI safety - a review," *Transactions on Machine Learning Research*, 2024, survey Certification, Expert Certification. [Online]. Available: https://openreview.net/forum?id=ePUVetPKu6

[4] A. Makelov, G. Lange, A. Geiger, and N. Nanda, "Is this the subspace you are looking for? an interpretability illusion for subspace activation patching," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=Ebt7JgMHv1

[5] L. Gao, T. D. la Tour, H. Tillman, G. Goh, R. Troll, A. Radford, I. Sutskever, J. Leike, and J. Wu, "Scaling and evaluating sparse autoencoders," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=tcsZt9ZNKD

[6] C. Lu, Y.-C. Tang, and A. Cavallaro, "Minimal neuron ablation triggers catastrophic collapse in the language core of large vision-language models," *arXiv preprint arXiv:2512.00918*, 2025.

[7] Z. Qin, K. Lyu, Q. Yu, Y. Sun, and Z. Fan, "The achilles' heel of llms: How altering a handful of neurons can cripple language abilities," *arXiv preprint arXiv:2510.10238*, 2025.

[8] C. Rorden, H.-O. Karnath, and L. Bonilha, "Improving lesion-symptom mapping," *J. Cognitive Neuroscience*, vol. 19, no. 7, p. 1081–1088, Jul. 2007. [Online]. Available: https://doi.org/10.1162/jocn.2007.19.7.1081

[9] B. Kim, J. Hewitt, N. Nanda, N. Fiedel, and O. Tafjord, "Because we have llms, we can and should pursue agentic interpretability," *arXiv preprint arXiv:2506.12152*, 2025.

[10] M. D. Lezak, *Neuropsychological assessment*. Oxford university press, 2004.

[11] F. Liu, H. Zhou, B. Gu, X. Zou, J. Huang, J. Wu, Y. Li, S. S. Chen, Y. Hua, P. Zhou *et al.*, "Application of large language models in medicine," *Nature Reviews Bioengineering*, pp. 1–20, 2025.

[12] W. Gurnee, N. Nanda, M. Pauly, K. Harvey, D. Troitskii, and D. Bertsimas, "Finding neurons in a haystack: Case studies with sparse probing," *Transactions on Machine Learning Research*, 2023. [Online]. Available: https://openreview.net/forum?id=JYs1R9IMJr

[13] Z. He, W. Shu, X. Ge, L. Chen, J. Wang, Y. Zhou, F. Liu, Q. Guo, X. Huang, Z. Wu, Y.-G. Jiang, and X. Qiu, "Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders," *CoRR*, vol. abs/2410.20526, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2410.20526

[14] W. Shi, S. Li, T. Liang, M. Wan, G. Ma, X. Wang, and X. He, "Route sparse autoencoder to interpret large language models," *CoRR*, vol. abs/2503.08200, March 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2503.08200

[15] R. Huben, H. Cunningham, L. R. Smith, A. Ewart, and L. Sharkey, "Sparse autoencoders find highly interpretable features in language models," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=F76bwRSLeK

[16] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[17] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[18] S. Tahir, J. Johnson, J. Abu-Khalaf, and S. A. Shah, "E-ther: A pct-grounded dataset for benchmarking empathic ai," *arXiv preprint arXiv:2509.02100*, 2025.

[19] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.