Avaliação do uso de Explicabilidade de LLMs e Grafos de Ativação (NRAGs) em estudos bibliométricos

Matheus Pereira[1], **Demian Pacheco**[1], **André Santanchè**[2], **Luiz Gomes-Jr**[1]

[1]Departamento de Informática – UTFPR – Curitiba, PR – Brasil

[2]Instituto de Computação – UNICAMP – Campinas, SP – Brasil

matheusp.2005@alunos.utfpr.edu.br, demianet@gmail.com

luizcelso@gmail.com, santanch@unicamp.br

Abstract. *Despite their popularity, LLMs are black-box models that pose challenges for interpretability and alignment. Explainability in AI (XAI) focuses on shedding light on these types of models, enabling researchers to detect patterns that aid in understanding their behavior. Besides revealing information about the models, these patterns can also reveal information about the data being fed to the algorithms. In this paper, we assess how the Explainability of LLMs could be applied in the field of bibliometry. The goal is to use LLM inference patterns to represent differences across research areas (using published papers in these areas). We use the LLM-MRI tool, which produces graphs (NRAGs) that represent neural activation patterns during LLM inference. We compute complex network metrics from the graphs to assess three different research fields in Computer Science.*

Resumo. *Apesar de sua popularidade, os Grandes Modelos de Linguagem (LLMs) são modelos de "caixa-preta"que apresentam desafios para a interpretabilidade e o alinhamento. A Explicabilidade em IA (XAI) foca em compreender esse tipo de modelo, permitindo que pesquisadores detectem padrões que auxiliem na compreensão de seu comportamento. Além de revelar informações sobre os modelos, esses padrões também podem revelar informações sobre os dados fornecidos aos algoritmos. Neste artigo, avaliamos como a Explicabilidade de LLMs poderia ser aplicada no campo da bibliometria. O objetivo é utilizar padrões de inferência de LLM para representar diferenças entre áreas de pesquisa (utilizando artigos publicados nessas áreas). Utilizamos a ferramenta LLM-MRI, que gera grafos (NRAGs) que representam padrões de ativação neural durante a inferência do LLM. Computamos métricas de redes complexas a partir dos grafos para avaliar três diferentes campos de pesquisa em Ciência da Computação.*

1. Introdução

Os modelos de linguagem de grande escala (LLMs) tem uso crescente na sociedade e na pesquisa científica, sendo uma ferramenta poderosa para análise e criação de texto, além de assistentes virtuais, tradutores e ferramentas de auxílio para tomadas de decisão [Zhao et al. 2024]. Para aumentar a segurança e permitir a auditoria e controle desses sistemas, a explicabilidade de LLMs surgiu como uma área de importância fundamental

[Sharkey et al. 2025]. Paralelamente, a bibliometria é um campo consolidado que mapeia a estrutura e a dinâmica da produção científica [Passas 2024]. O uso de LLMs para análises bibliométricas se apresenta como uma área promissora de estudo, ainda pouco explorada.

A complexidade estrutural da arquitetura transformer usada nos LLMs torna a explicabilidade um desafio científico ainda não resolvido. Os LLMs operam com um espaço de alta dimensionalidade, com informações distribuídas em padrões de ativação em múltiplos neurônios, com processamento semântico complexo e mediado por múltiplas camadas não-lineares. Combinado com o problema da superposição de conceitos, no qual os mesmos neurônios são usados para armazenamento de diferentes informações, isso faz com que esses modelos sejam "caixas-pretas" em termos de interpretabilidade [Samek et al. 2017].

Para obter melhor visualização do funcionamento interno dos LLMs, várias técnicas já foram propostas e implementadas na literatura. As técnicas de interpretabilidade local, como LIME [Ribeiro et al. 2016] e SHAP [Lundberg and Lee 2017]explicam predições individuais, identificando quais palavras e features mais contribuem para um resultado dado uma entrada em especifico. Já as técnicas de interpretabilidade global, de aplicação mais generalista, permitem mapear padrões de funcionamento global dos modelos de linguagem, através de métodos como o (*Sparse Auto Encoders*) [Cunningham et al. 2023]. A Anthropic tem avançado significativamente nessa direção, aplicando SAEs para mapear "circuitos" computacionais [Bricken et al. 2023] e identificar "vetores de persona" que controlam aspectos comportamentais dos modelos [Chen et al. 2025].

Nete artigo, foi escolhida a técnica de geração de NRAG (*Neural Region Activation Graphs*) como abordagem para explicabilidade e visualização dos padrões presentes nas diferentes áreas da literatura cientifica. Os NRAGs reduzem de dimensionalidade dos modelos, agrupando neurônios em regiões, para então conectar as regiões co-ativadas usando grafos [Horta et al. 2021], e gerando artefatos na forma de mapas 2d das ativações, permitindo a visualização. A biblioteca LLM-MRI, desenvolvida por [Costa et al. 2024], implementa essa abordagem de forma acessível, menos custosa que SAEs e técnicas mais avançadas como *circuit tracing*, permitindo comparar diferentes domínios e visualizar como eles ativam regiões diferentes do modelo.

Para validar a abordagem, obtivemos um corpus de referência de diferentes áreas de produção científica, contendo 500 resumos (*abstracts*) de cada uma das seguintes áreas: Inteligência Artificial, Visão Computacional e Bibliotecas Digitais. O restante do artigo está organizado da seguinte maneira: a Seção 2 descreve fundamentos e trabalhos relacionados a pesquisa. A Seção 3 descreve a metodologia utilizada, com o uso do módulo LLM-MRI. A Seção 4 apresenta os resultados e métricas extraídas e sua interpretação. Os NRAGs gerados a partir do corpus foram processados aplicando a Análise de Variância ANOVA, quantificando estatisticamente a diferença estrutural entre as categorias. A Seção 5 avalia os resultados obtidos e as contribuições do trabalho como um todo.

2. Fundamentos e trabalhos relacionados

2.1. Bibliometria

A bibliometria é um método quantitativo de estudo da literatura científica. As técnicas de análise bibliométrica consistem na coleta e tratamento dos dados da literatura, seguida da aplicação de ferramentas estatísticas, de forma a obter métricas e a mapear padrões e tendências na produção científica [Passas 2024, Sousa et al. 2024]. Os resultados da análise bibliométrica possibilitam compreender como o conhecimento e a informação são estruturados e disseminados dentro de uma área de estudo, também auxiliando a identificar lacunas e temas emergentes [Celestino et al. 2024].

2.2. LLMs

Os modelos de linguagem são sistemas computacionais treinados para prever os elementos subsequentes em um texto de linguagem natural [Jurafsky and Martin 2026]. O estudo de modelos de linguagem se consolidou na década de 1980, com modelos estatísticos probabilísticos com representação discreta. Com o advento dos word embeddings [Bengio et al. 2003], a área possibilitou mapear a semântica de itens lexicais na forma de vetores, permitindo representação continua. Posteriormente, a arquitetura *Transformer*, introduzida em 2017 [Vaswani et al. 2017], introduziu o processamento paralelo e mecanismos de auto-atenção, viabilizando o surgimento dos LLMs (Grandes Modelos de Linguagem). Podendo ser pré-treinados com volumes massivos de dados, os LLMs têm aplicação crescente em tradução, sumarização, análise e geração de texto, assim como interfaces conversacionais.

2.3. Explicabilidade em Modelos de Linguagem

A crescente integração de algoritmos de Aprendizado de Máquina (*Machine Learning*) em processos decisórios críticos elevou a demanda por transparência. Em muitos cenários, a acurácia do modelo isoladamente é insuficiente [Doshi-Velez and Kim 2017]; a compreensão dos motivos que levaram a um resultado é tão fundamental quanto a predição em si. Nesse contexto, insere-se a área de Inteligência Artificial Explicável (XAI).

Na literatura, os conceitos de explicabilidade e interpretabilidade referem-se ao grau de compreensão humana sobre o funcionamento ou as decisões de um modelo [Miller 2019]. Sob essa ótica, os modelos são classificados em duas categorias: os intrínsecos (ou transparentes), cujas estruturas internas (como árvores de decisão simples) permitem o entendimento direto do fluxo de dados; e os de "caixa-preta", como os Grandes Modelos de Linguagem (LLMs). Estes últimos, baseados em arquiteturas de aprendizado profundo (*Deep Learning*) com bilhões de parâmetros, possuem operações internas cuja lógica é opaca tanto para usuários quanto para desenvolvedores.

A explicabilidade pode ser dividida em dois escopos principais: local e global [Zhao et al. 2024]. A explicabilidade local busca isolar os fatores que influenciaram uma saída específica do modelo (ex: quais palavras-chave de um *prompt* geraram certo *token*). Já as abordagens globais, frequentemente associadas à Interpretabilidade Mecanicista, buscam mapear o funcionamento interno do modelo de forma holística, tentando decodificar como conceitos e padrões são representados em suas camadas e grafos de ativação.

2.4. LLM-MRI

Para viabilizar a análise mecanicista proposta, este trabalho utiliza o LLM-MRI [Costa et al. 2024], um módulo Python desenvolvido para simplificar o estudo de padrões de ativação em modelos baseados em Transformers. A ferramenta atua de forma análoga a um exame de imagem de ressonância magnética, permitindo o mapeamento de unidades funcionais dentro da rede neural. As principais funcionalidades oferecidas pelo pacote incluem:

- **Redução de Dimensionalidade:** Devido à alta dimensionalidade dos LLMs, o módulo provê técnicas para agrupar neurônios com padrões de ativação similares em regiões neurais, possibilitando a visualização e a análise de padrões latentes;
- **Representação Visual de Ativações:** O módulo gera visualizações dos padrões de ativação tanto em camadas específicas (definidas pelo usuário) quanto ao longo de toda a rede, permitindo filtrar e comparar ativações por categorias de documentos ou rótulos específicos;
- **Representação em Grafo das Ativações:** a ferramenta conecta regiões de camadas vizinhas com base em suas co-ativações, induzindo um Grafo de Ativação de Regiões Neurais (NRAGs). Esta abstração permite a extração de métricas de Redes Complexas para quantificar o comportamento do modelo.
- **Ecossistema Aberto:** O pacote é distribuído sob licença LGPL v3 via GitHub, visando o engajamento da da comunidade científica no desenvolvimento de métodos de explicabilidade.

  No contexto deste estudo bibliométrico, essas funcionalidades permitem explorar se diferentes domínios científicos ativam regiões neurais distintas e quais as discrepâncias topológicas nos grafos formados por cada categoria de texto. Assim, o LLM-MRI fornece a base metodológica para identificar se certas regiões do modelo são mais sensíveis a aspectos específicos da literatura analisada, transformando a natureza de "caixa-preta" do LLM em uma estrutura de dados interpretável.

2.5. Trabalhos Correlatos

Os estudos na área da explicabilidade se consolidaram relativamente recentemente, sendo produzidos principalmente por grandes instituições de pesquisa dentro de empresas com acesso direto aos modelos e com grande poder computacional. Além da monitorabilidade via *Chain-of-thought* [Korbak et al. 2025], a área se divide entre diversas abordagens que permitem a interpretabilidades dos modelos em si, com interpretabilidade local e global.

Na área da interpretabilidade local, os modelos buscam explicar resultados de previsões individuais. Nesta área, os métodos se dividem em dois grupos principais: abordagens que buscam explicar uma previsão como soma da contribuição de features (LIME, Shapley, SHAP) e abordagens que permitem a visualização da sensibilidade a *features* individuais (ceteris paribus, ICE).

Na área da interpretabilidade global, se destacam dois grupos principais de métodos: os métodos modelo-específicos, e os métodos modelo-agnósticos, que são de maior interesse para aplicação geral da interpretabilidade. Entre o métodos modelo-agnósticos, destaca-se métodos que mensuram a interação das features e métodos que

relacionam features e predições, como o gráfico de dependência parcial (PDP) e os efeitos locais acumulados (ALE).

Nos últimos anos, foi dada atenção adicional a métodos Sparse Auto Encoders (SAEs), que permitam a decomposição monosemântica [Bricken et al. 2023], ampliando a dimensionalidade dos modelos através de *encoders*, para permitir a quebra da sobreposição dos conceitos e seu posterior mapeamento. Os NRAGs (grafos de ativação de regiões neurais) aplicam uma combinação dos métodos de redução de dimensionalidade com conceitos de análise topológica, possibilitando a geração de artefatos para visualização e explicabilidade das estruturas internas dos LLMs [Horta et al. 2021].

Recentemente, a abordagem de Circuit Tracing tem se destacado na interpretabilidade mecanicista de LLMs. Na técnica proposta por pesquisadores da Anthropic [Ameisen et al. 2025], um *cross-layer transcoder* (CLT) é treinado para substituir localmente as camadas MLP originais do modelo. Para um *prompt* em específico, o CLT treinado atua como um modelo de substituição, com *features* mais esparsas e com interações lineares. Isso possibilita a produção grafos de atribuição, mapeando o circuito associado a um determinado *output*. A técnica foi aplicada para identificar os circuitos internos do modelo Claude 3.5 Haiku responsáveis pelo funcionamento multilíngue, raciocínio multietapas, entre outras funcionalidades [Lindsey et al. 2025]. Comparado aos NRAGs, essa técnica tem custo computacional mais alto devido à necessidade do treinamento do CLT e ao aumento da dimensionalidade.

A aplicação da explicabilidade na bibliometria ainda é uma área pouco explorada, não tendo sido encontrado o uso dela na literatura.

3. Metodologia

A metodologia aplicada neste trabalho utiliza o módulo *LLM-MRI* como instrumento de medição para investigar como diferentes domínios científicos geram padrões de ativação estruturalmente distintos. O fluxo de trabalho é guiado pelas três fases principais da biblioteca (Figura 1), detalhadas a seguir.

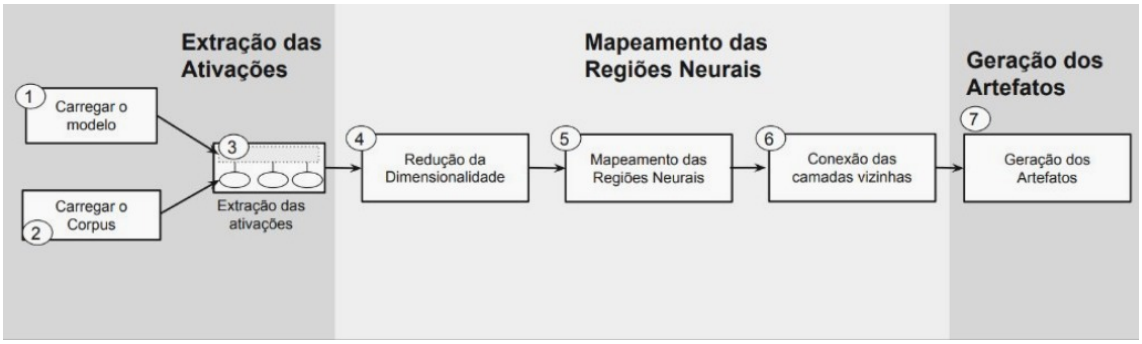

Figura 1. Representação do fluxo metodológico baseado no *pipeline* da biblioteca LLM-MRI [Gomes-Jr et al. 2025].

3.1. Fase de Extração de Ativações

O processo inicia-se na etapa de **Extração de Ativações**, onde o modelo *distilbert-base-uncased* é carregado como o extrator de características (passo 1). A escolha do DistilBERT justifica-se por ser uma versão compacta e eficiente do BERT que, por meio

da técnica de destilação de conhecimento, retém aproximadamente 97% do desempenho do modelo original com uma arquitetura mais leve, facilitando o processamento em larga escala necessário para esta pesquisa. Em seguida, define-se o *Corpus* de referência (passo 2), composto por resumos (*abstracts*) científicos previamente rotulados por categoria temática.

Nesta fase, cada documento do *corpus* é processado pelo modelo (passo 3). Durante o *forward-pass*, o *LLM-MRI* captura as ativações das camadas ocultas da rede. Estas ativações brutas representam a resposta imediata dos neurônios ao conteúdo semântico de cada documento e são armazenadas para o processamento subsequente.

3.2. Fase de Mapeamento de Regiões Neurais

Após o processamento do *corpus*, inicia-se o **Mapeamento de Regiões Neurais**. Como as camadas do modelo possuem alta dimensionalidade, a biblioteca reduz os dados de ativação de cada camada utilizando o algoritmo *UMAP (Uniform Manifold Approximation and Projection)* (passo 4). O UMAP foi selecionado por sua capacidade superior em preservar tanto a estrutura local quanto a global dos dados, o que é essencial para diferenciar temas científicos cujas nuances semânticas podem ser sutis.

As dimensões reduzidas são projetadas em mapas 2D de tamanho definido, onde cada célula agrupa neurônios com padrões de ativação similares, formando as chamadas regiões neurais (passo 5). No estágio final desta fase, o Grafo de Ativação de Regiões Neurais (NRAG) é induzido para representar o fluxo de ativação em todo o modelo (passo 6). As arestas do grafo conectam regiões de camadas vizinhas com base na co-ativação estimulada pelos documentos, onde o peso dessas arestas reflete a frequência com que o conhecimento percorre determinados caminhos na rede.

3.3. Fase de Geração de Artefatos

Na fase final de **Geração de Artefatos**, os NRAGs gerados (que utilizam a estrutura da biblioteca *NetworkX*) servem de base para a extração de métricas de Redes Complexas. Esta etapa permite transformar a abstração das ativações neurais em dados estruturados, possibilitando a análise quantitativa e a criação de representações visuais dos grafos para comparação entre diferentes conjuntos de documentos.

3.4. Aplicação Experimental

Para validar a abordagem, utilizou-se um *corpus* contendo 500 artigos para cada domínio científico analisado. O experimento adotou um procedimento estocástico rigoroso: para cada categoria, foram realizadas 1000 execuções independentes. Essa análise estatística ampla evita que eventuais diferenças encontradas na estrutura dos grafos de cada área sejam resultantes de meras coincidências relacionadas à seleção dos artigos. Em cada iteração, selecionou-se aleatoriamente um subconjunto de 70 artigos (quantidade definida para garantir grafos estáveis) e extraíram-se as métricas de redes complexas: Densidade, Coeficiente de Assortatividade, Grau Médio de Conectividade, Número de Comunidades, Modularidade, Centralidade de Grau Máxima e Centralidade de Grau Média por Camada.

A validação científica da distinção entre as categorias baseia-se no teste de Análise de Variância (ANOVA), aplicado sobre o montante de dados das 1000 execuções para comprovar se as variações entre as métricas de cada categoria são estatisticamente relevantes. Adicionalmente, os dados foram explorados através de artefatos visuais:

- **Barplots:** utilizados para comparação de médias e desvios padrão das métricas;
- **Histogramas com KDE:** aplicados para analisar a distribuição probabilística de cada grupo;
- **Boxplots:** utilizados para representar a variação da centralidade de grau ao longo da arquitetura do modelo.

4. Resultados

Os experimentos realizados demonstram que as métricas extraídas dos Grafos de Ativação de Regiões Neurais (NRAGs) atuam como descritores robustos para a diferenciação de domínios científicos na literatura. A análise estatística confirma que o comportamento interno do modelo *distilbert-base-uncased* varia de forma consistente e significativa conforme a categoria bibliométrica processada.

4.1. Validação Estatística via ANOVA

Para validar a hipótese de que a estrutura de ativação do modelo é sensível ao domínio temático, aplicou-se o teste ANOVA sobre os dados resultantes das 1000 execuções independentes de cada categoria. Os resultados, sintetizados na Tabela 1, revelaram que todas as métricas de rede complexa analisadas apresentaram valores-p significativos ($p < 0,001$).

Tabela 1. Resultados do Teste ANOVA para as métricas dos NRAGs.

Métrica Analisada	**Descrição**	$p < 0,001$
Densidade	Razão entre o número de arestas presentes e o máximo de arestas possíveis.	Sim
Coeficiente de Assortatividade	Medida de correlação entre o grau de nós que se conectam entre si.	Sim
Grau Médio de Conectividade	Média aritmética do número de conexões incidentes nos nós da rede.	Sim
Número de Comunidades	Quantidade de subgrupos de nós com alta densidade de conexões internas.	Sim
Centralidade de Grau Máxima	Maior valor de conexões incidentes observado em um único nó da rede.	Sim
Modularidade	Grau de compartimentação da rede em grupos ou módulos independentes.	Sim

A distinção absoluta nas métricas sugere que o modelo organiza a informação de forma topologicamente diferente para cada tema, evidenciando que domínios distintos recrutam subestruturas neurais específicas.

4.2. Representação Visual das Métricas

Embora seis métricas tenham sido validadas com sucesso pelo teste ANOVA, a Figura 2 ilustra a distribuição probabilística para a métrica de Densidade. Esta métrica foi selecionada por apresentar o maior contraste visual entre as curvas das categorias analisadas.

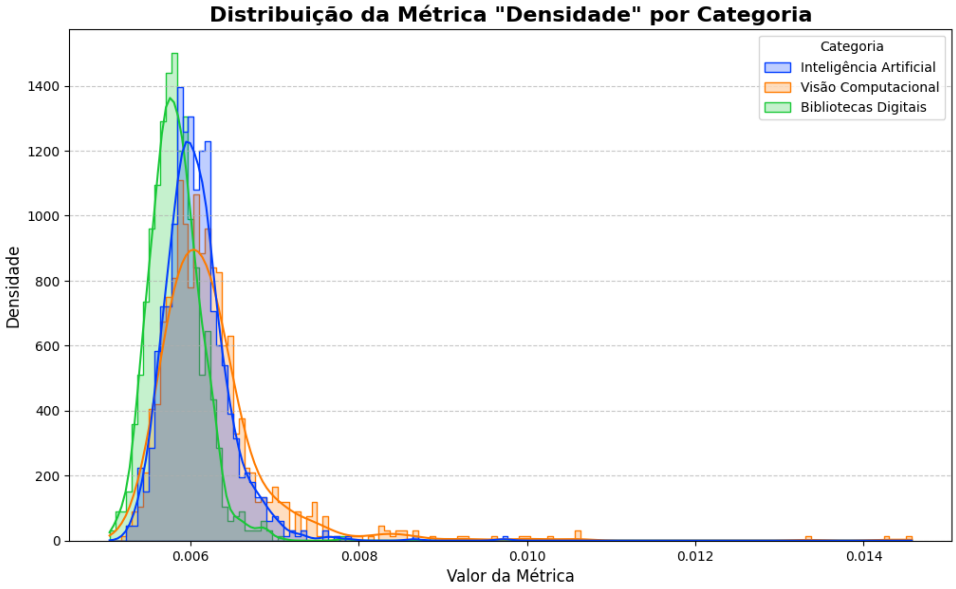

Figura 2. Distribuição probabilística (KDE) da métrica de Densidade para as três categorias.

O gráfico serve como uma evidência visual do rigor estatístico obtido. A clara separação dos picos das curvas de densidade (*Kernel Density Estimation* - KDE) para as categorias de Inteligência Artificial, Visão Computacional e Bibliotecas Digitais prova que as diferenças estatísticas não são fruto de ruído estocástico, mas sim padrões estruturais estáveis que o modelo atribui a cada domínio científico de forma distinta.

A análise comparativa revela também que o domínio de Bibliotecas Digitais apresenta o comportamento mais divergente entre as categorias. O fato de esta área apresentar uma densidade de ativação menor sugere uma estrutura terminológica mais restrita e específica dentro do modelo para este campo. Em contrapartida, os domínios de Inteligência Artificial e Visão Computacional exibem maior proximidade e densidade de ativação, o que indica áreas com maior interdisciplinaridade e sobreposição de conceitos semânticos na arquitetura do modelo.

5. Conclusão

Este trabalho investigou a viabilidade do uso de técnicas de explicabilidade mecanicista, especificamente os Grafos de Ativação de Regiões Neurais (NRAGs), como uma nova abordagem para estudos bibliométricos. Através da utilização do módulo *LLM-MRI*, foi possível mapear como o modelo *distilbert-base-uncased* processa diferentes domínios científicos, transformando padrões internos de ativação em métricas quantificáveis de redes complexas.

Os resultados obtidos através da análise estatística ANOVA foram conclusivos, apresentando valores-p significativos para todas as métricas estruturais analisadas. Além disso, a representação visual via curvas de densidade (KDE) confirmou que as categorias analisadas possuem comportamentos estáveis e distintos. Esses resultados comprovam que a topologia de ativação de um LLM não é aleatória, mas sim intrinsecamente ligada ao domínio semântico do texto processado.

Como trabalhos futuros, pretende-se expandir a análise através do uso de outros modelos de linguagem de diferentes escalas e arquiteturas. Além disso, planeja-se a adaptação do experimento para as novas versões da biblioteca *LLM-MRI*, explorando a inclusão de métodos alternativos de redução de dimensionalidade, como PCA (*Principal Component Analysis*) e SVD (*Singular Value Decomposition*), a fim de avaliar o impacto dessas técnicas na estruturação e interpretação dos Grafos de Ativação de Regiões Neurais.

Referências

Ameisen, E. et al. (2025). Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3(Feb):1137–1155.

Bricken, T. et al. (2023). Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Celestino, M. S., Belluzzo, R. C. B., Albino, J. P., and Valente, V. C. P. N. (2024). AnÁlise bibliomÉtrica: RevisÃo de literatura e proposta de framework metodolÓgico em 12 passos. *ARACÊ*, 6(4):13421–13446.

Chen, R. et al. (2025). Persona vectors: Monitoring and controlling character traits in language models.

Costa, L., Figenio, M., Santanchè, A., and Gomes-Jr, L. (2024). Llm-mri python module: a brain scanner for llms. In *Companion Proceedings of the 39th Brazilian Symposium on Data Bases (SBBD)*, Florianópolis, SC, Brazil.

Cunningham, H. et al. (2023). Sparse autoencoders find highly interpretable features in language models. *ArXiv*, abs/2309.08600.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Gomes-Jr, L., Santanchè, A., Figenio, M., and Costa, L. (2025). Explicabilidade de LLMs usando grafos de ativação de regiões neurais (NRAGs). In *Proceedings of the XIX Brazilian e-Science Workshop (BreSci)*, pages 17–24, Fortaleza, CE, Brasil.

Horta, V. A., Tiddi, I., Little, S., and Mileo, A. (2021). Extracting knowledge from deep neural networks through graph analysis. *Future Generation Computer Systems*, 120:109–118.

Jurafsky, D. and Martin, J. H. (2026). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. Stanford University, 3rd edition. Online manuscript released January 6, 2026.

Korbak, T. et al. (2025). Chain of thought monitorability: A new and fragile opportunity for ai safety.

Lindsey, J. et al. (2025). On the biology of a large language model. *Transformer Circuits Thread*.

Lundberg, S. M. and Lee, S. (2017). A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Passas, I. (2024). Bibliometric analysis: The main steps. *Encyclopedia*, 4:1014–1025.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938.

Samek, W., Wiegand, T., and Müller, K. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *CoRR*, abs/1708.08296.

Sharkey, L. et al. (2025). Open problems in mechanistic interpretability.

Sousa, M., ALMEIDA, E., and Dantas Bezerra, A. (2024). Bibliometria: o que é? para que serve? e como se faz? bibliometrics: what is it? what is it used for? and how to do it? bibliometría: ¿qué es? ¿para qué es? ¿y cómo lo haces? *Cuadernos de Educación y Desarrollo*, 16:01–35.

Vaswani, A. et al. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Zhao, H. et al. (2024). Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*. Just Accepted.