Explicabilidade de LLMs usando Grafos de Ativação de Regiões Neurais (NRAGs)

Luiz Gomes-Jr [1], André Santanchè [2], Mateus Figenio[1], Luiz Costa[2]

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

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

lcjunior@utfpr.edu.br, santanch@unicamp.br

Abstract. *LLMs são atualmente tecnologias centrais para o desenvolvimento de aplicações de IA, atraindo interesse acadêmico e grandes investimentos na indústria. Apesar do sucesso e amplo uso dessas tecnologias, LLMs apresentam grandes desafios em termos de interpretabilidade – a complexidade dos modelos e a natureza de caixa-preta das redes neurais dificultam a compreensão dos mecanismos de geração das saídas. Este artigo apresenta os Grafos de Ativação de Regiões Neurais (NRAGs – do inglês Neural Region Activation Graphs), nova proposta para mecanismos de explicabilidade em LLMs. NRAGs são representações em grafo das ativações de um LLM estimuladas por um corpus. Os grafos gerados podem ser usados em tarefas como (i) entender as interconexões entre diferentes regiões do espaço multidimensional das camadas da rede, (ii) comparar subgrafos de ativações de textos de diferentes categorias, e (iii) comparar propriedades dos grafos induzidos por LLMs diferentes para um mesmo corpus. NRAGs são implementados na biblioteca LLM-MRI, que oferece diversos artefatos para o estudo de ativações em LLMs. Este artigo apresenta NRAGs como uma alternativa para a investigação científica de fenômenos complexos resultantes da inferência de LLMs, cobrindo a geração dos grafos a partir da biblioteca LLM-MRI e exemplos de aplicações em andamento.*

1. Introdução

Os modelos de linguagem em grande escala (LLMs do inglês Large Language Models) têm revolucionado o campo da inteligência artificial (IA), emergindo como tecnologias centrais para uma ampla gama de aplicações, desde assistentes virtuais e sistemas de tradução até ferramentas de criação de conteúdo e suporte em tomada de decisão [Zhao et al. 2024b, Hiter 2024]. Impulsionados por avanços no aprendizado profundo e treinados em volumes massivos de dados, esses modelos têm demonstrado proficiência na geração de texto, inferência contextual e adaptação a diferentes tarefas com ajustes mínimos. Contudo, apesar de seu sucesso, os LLMs apresentam um desafio crítico que limita seu potencial: a falta de explicabilidade [Figênio and Gomes-Jr 2023, Figênio et al. 2024b].

A explicabilidade é um componente essencial para que sistemas de IA sejam confiáveis, auditáveis e eticamente aplicáveis, especialmente em cenários de alto impacto como saúde, finanças e justiça. No entanto, a complexidade arquitetural das redes neurais profundas que sustentam os LLMs, combinada com a enorme dimensionalidade dos espaços internos de representação, transforma esses modelos em verdadeiras "caixas-pretas" em termos de interpretabilidade [Samek et al. 2017].

Os NRAGs (do inglês Neural Region Activation Graphs) representam uma abordagem inovadora que converte as ativações internas de um LLM, estimuladas por um corpus, em grafos que capturam interconexões entre diferentes regiões do espaço multidimensional das camadas da rede. Essa representação em grafos permite investigações sobre como diferentes partes do modelo se comunicam e se influenciam durante o processamento. Os NRAGs permitem diversas aplicações promissoras, como (i) a análise das interconexões entre regiões ativadas em diferentes camadas do modelo, (ii) a comparação de subgrafos de ativação gerados por textos de categorias distintas, e (iii) o estudo das propriedades dos grafos induzidos por diferentes LLMs quando expostos ao mesmo corpus. Essas análises possibilitam um melhor entendimento das estruturas internas dos LLMs, oferecendo ferramentas práticas para comparar e avaliar modelos, identificar padrões emergentes e, potencialmente, corrigir comportamentos indesejados. A implementação dessas representações e análises é facilitada pela biblioteca LLM-MRI[1], que fornece um conjunto de ferramentas para extrair, manipular e estudar ativações neurais em LLMs [Costa et al. 2024].

Ao unir conceitos de análise de grafos com o estudo das ativações internas dos LLMs, NRAGs visam contribuir para o avanço da explicabilidade em IA, permitindo investigações científicas a respeito dos padrões de ativação estimulados por um corpus. A compreensão mais profunda dos mecanismos subjacentes a esses modelos não apenas aumentará sua transparência e confiabilidade, mas também abrirá novas perspectivas para a otimização e desenvolvimento de sistemas de IA mais éticos, robustos e alinhados às necessidades humanas.

O objetivo deste artigo é introduzir o conceito de NRAG, explicando a sua construção na biblioteca LLM-MRI e descrevendo frentes atuais de aplicação da técnica. O restante deste artigo está organizada da seguinte maneira: A Seção 2 resume os fundamentos e trabalhos correlatos. A Seção 3 descreve o funcionamento da biblioteca LLM-MRI e o processo de geração de NRAGs, bem como aplicações em andamento das técnicas. Por fim, a Seção 4 conclui o artigo com uma visão geral das expectativas e trabalhos futuros para a pesquisa.

2. Fundamentos e Trabalhos Relacionados

2.1. LLMs

Modelos de Linguagem inferem uma distribuição de probabilidade de palavras em textos de linguagem natural [Bengio et al. 2000], permitindo interpretar e gerar textos livres. Esses modelos foram diretamente beneficiados pela evolução das capacidades das redes neurais, que possibilitam a extração de informações semânticas de textos de forma que uma máquina possa analisá-las e manipulá-las [Tunstall et al. 2022].

Com o avanço em poder computacional e sobretudo com a introdução do *transformers* [Vaswani et al. 2017], os Modelos de Linguagem de Grande Escala (LLMs) foram concebidos. LLMs são modelos que, treinados com volumes substanciais de textos e contando com um maior número de parâmetros internos, conseguem se aproximar do desempenho humano em diversas tarefas [Naveed et al. 2024]. Esses modelos possuem considerável capacidade de generalização, adaptada para lidar com grandes volumes de

[1]https://github.com/explic-ai/LLM-MRI

dados, permitindo interpretar e produzir respostas a requisições complexas numa gama de tarefas distintas.

2.2. Explicabilidade em LLMs

Explicabilidade de LLMs pode ser dividida em dois objetivos: explicabilidade local e global [Zhao et al. 2024a]. Explicabilidade local foca em entender os fatores que influenciam a geração de uma determinada saída. É uma abordagem importante para garantir a transparência dos modelos. As abordagens globais, também descritas como mecanicistas, buscam entender e/ou influenciar o funcionamento dos LLMs em um leque de gerações e tarefas.

As abordagens locais de explicabilidade focadas no entendimento das representações internas dos modelos têm se baseado na visualização dos mecanismos de atenção. Isso inclui desde a visualização de como cabeças de atenção individuais avaliam *tokens* [Vaswani et al. 2017], até como os valores de atenção fluem através do modelo [DeRose et al. 2020], e como cabeças de atenção individuais se relacionam com conceitos fornecidos pelo usuário [Hoover et al. 2019].

Abordagens mecanicistas de explicabilidade focam em identificar padrões nos neurônios ou outros elementos da arquitetura da rede. Estas abordagens começaram a ser estudadas no contexto geral de DNNs (redes neurais profundas), sobretudo considerando fatores topológicos das redes [Zhang et al. 2024, Horta et al. 2021]. [Horta et al. 2021] constroem um grafo denso baseado nas ativações dos neurônios de DNNs de classificação de imagens. As ativações são coletadas a partir da passagem (*forward-pass*) de amostras do dataset de treino. Neurônios são conectados de acordo com a co-ativação medida pela correlação de *spearman*. Os grafos são então usados em análises de redes complexas [L. da F. Costa and Boas 2007], com algoritmos de detecção de comunidades e centralidade. No contexto de LLMs, a explicabilidade pode focar nos mecanismos de atenção ou na ativação dos neurônios [Figênio et al. 2024a, Zhao et al. 2024a].

Existem também abordagens de explicabilidade baseadas em *probing* e ativações de neurônios [Bau et al. 2018, Dalvi et al. 2019]. O *probing* é uma técnica fundamentada na ideia de treinar um classificador superficial sobre as representações vetoriais de sentenças geradas pelo modelo para entender o que ele aprendeu — uma abordagem indireta para compreender a rede neural do modelo.

Uma abordagem que tem recebido atenção no contexto mecanicista é a de *Sparse Auto Encoders* (SAEs) [Lieberum et al. 2024, Cunningham et al. 2023]. SAEs têm como objetivo identificar o espaço real de conceitos aprendidos por uma LLM. Os conceitos aprendidos não são proporcionais à dimensionalidade dos LLMs por conta do fenômeno de sobreposição de conceitos, no qual uma dimensão frequentemente codifica subespaços de conceitos não relacionados. SAEs criam **encoders** que ampliam a dimensionalidade original dos modelos, adicionando restrições de esparsidade para forçar a quebra das sobreposições dos conceitos, mapeando-os em dimensões distintas no novo espaço.

A proposta dos grafos de ativação de regiões neurais (NRAGs) é uma abordagem mecanicista que mistura conceitos de análises topológicas e SAEs, porém considera o fator humano no processo, favorecendo geração de artefatos em mais alto nível de abstração. A dimensionalidade reduzida dos artefatos gerados favorece a interpretabilidade humana e a compreensão de estruturas mais gerais das redes internas dos LLMs. A pesquisa

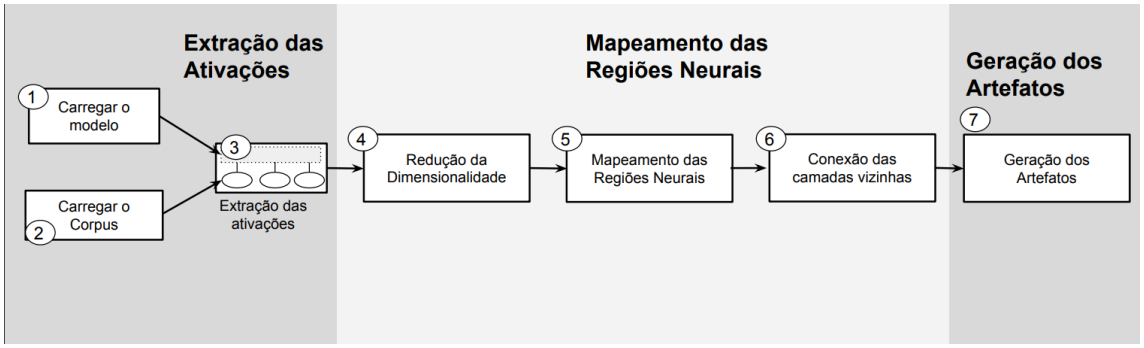

Figura 1. Diagrama representando o pipeline de funcionalidades da biblioteca (adaptado de: [Costa et al. 2024]

mais similar à proposta é a de [Horta et al. 2021], porém há diferenças significativas na metodologia sobretudo relacionadas com: (i) foco em LLMs e não DNNs de imagens, (ii) redução de dimensionalidade nas ativações, focando em regiões e não neurônios específicos, (iii) geração de um multigrafo a partir das *labels* dos documentos.

3. Metodologia e Pesquisas em Andamento

3.1. Explicabilidade em LLMs usando LLM-MRI e NRAGs

O workflow de geração de NRAGs foi implementado e disponibilizado na biblioteca LLM-MRI [Costa et al. 2024]. O funcionamento da biblioteca é dividido em três etapas: Extração de Ativações, Mapeamento de Regiões Neurais e Geração de Artefatos, conforme mostrado na Figura 1. Essas etapas são detalhadas a seguir.

Na fase de **Extração de Ativações**, o modelo LLM a ser visualizado é carregado (passo 1). Como exemplo, um modelo BERT pode ser definido nesta etapa. O usuário também especifica o corpus a ser usado como referência para as ativações de baseline (passo 2). Em nosso exemplo apresentado a seguir, usamos um corpus de *fake news*. Nesse ponto, cada documento no corpus fornecido é processado pelo modelo (passo 3). O *forward-pass* do modelo gera as ativações associadas a cada documento. As ativações são armazenadas para serem usadas nas próximas etapas.

Após o processamento do corpus, o **Mapeamento de Regiões Neurais** é iniciado, com a biblioteca reduzindo a dimensionalidade de cada camada de ativações, produzindo uma matriz bidimensional de ativações (passo 4). Atualmente, utilizamos o UMAP para a redução de dimensionalidade. Após a redução, as duas novas dimensões (para cada camada) são divididas em mapas 2D de tamanho definido pelo usuário (passo 5). Por exemplo, uma camada com 768 neurônios pode ser projetada em mapas de 10 por 10 células. Cada célula representa uma região da dimensionalidade original, ou seja, um valor de ativação na camada original é mapeado para uma região do espaço reduzido.

No passo final desta fase, é criado o NRAG que representa as ativações para todo o modelo considerando o corpus de entrada (passo 6). Os nós do grafo são as regiões neurais nos mapas 2D criados nas etapas anteriores. As arestas representam ativações subsequentes em camadas vizinhas (agora representadas pelos mapas). Um contador de arestas registra o número total de documentos que ativaram cada par de regiões vizinhas. Esses contadores são agregados como pesos das arestas e podem ser usados para ajustar o

grafo de acordo com a preferência do usuário (por exemplo, mantendo apenas as conexões mais fortes).

A última fase é a **Geração de Artefatos**, onde se destacam os modelos e as renderizações dos NRAgs. O grafo gerado para o NRAG subjacente é um modelo do *Networkx*[2], o que permite o uso da biblioteca *Networkx* para processamento adicional, como transformações e cálculo de métricas de ciência de redes (por exemplo, agrupamento, centralidade, etc.). A Figura 2 apresenta a imagem renderizada de um grafo com duas categorias diferentes destacadas: "false" em azul, "true" em vermelho (oriundas de um corpus de fake news), com cores intermediárias para nós compartilhados entre as duas categorias.

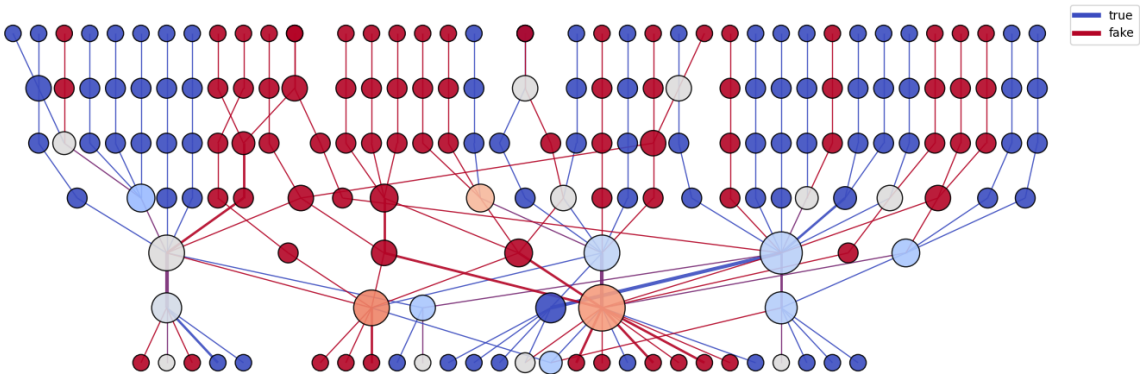

Figura 2. Visualização de grafos para diferentes tipos de ativações através das camadas.

3.2. NRAGs aplicados a avaliação de expertise

A primeira aplicação prática de NRAGs é no campo de avaliação de expertise humana. A estratégia consiste em avaliar diferenças topológicas nos NRAGs gerados que possam capturar o nível de expertise expresso nos textos de entrada.

Para avaliar como os padrões de ativação em LLMs se relacionam com a expertise, é necessário um conjunto de dados que contenha textos anotados que avaliem a expertise representada neles. Os dados utilizados nesta avaliação consistem de respostas anonimizadas de estudantes de medicina a um experimento de *"illness script"*, cujo objetivo é

[2]https://networkx.org/

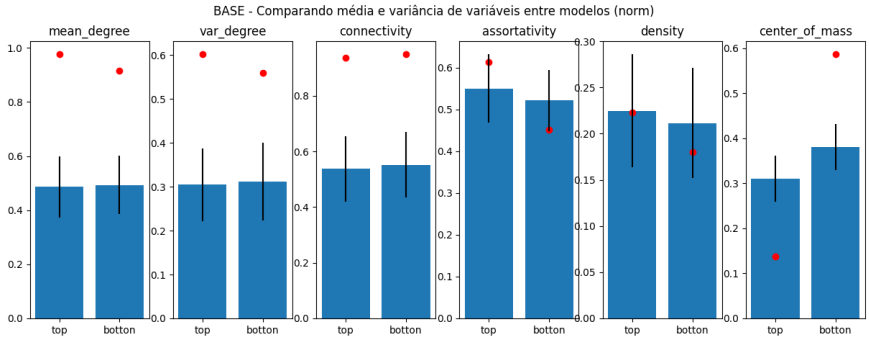

Figura 3. Comparação de métricas de acordo com nível de expertise

avaliar os estudantes em termos de encapsulamento e organização do conhecimento desenvolvido, conceitos que se referem a como médicos utilizam sua expertise para chegar a um diagnóstico preciso [Schmidt and Rikers 2007].

Esse processo envolve tanto um repertório de conhecimento sobre sintomas e doenças quanto a capacidade de estruturar esse conhecimento em um protocolo para obter um diagnóstico preciso para um determinado paciente. O conjunto de dados foi fornecido por pesquisadores da área médica que buscam avaliar o desenvolvimento de *"illness scripts"* em estudantes de medicina e para os quais este trabalho visa fornecer ferramentas que auxiliem na avaliação dessas propriedades.

Esse conjunto de dados contém a nota que o estudante alcançou no teste, o texto completo das respostas às questões de resposta livre, e outras informações e avaliações do experimento. A nota foi usada para classificar as respostas com alta/baixa expertise. NRAGs foram então gerados para representar as respostas de cada categoria. A Figura 3 compara diferentes métricas de redes complexas aplicadas aos NRAGs das categorias. É possível perceber diferenças claras para métricas como assortatividade e centro de massa (métrica definida para representar o viés do "peso"das redes para as camadas iniciais ou finais). Esta pesquisa em andamento busca avaliar estas diferenças e caracterizar os padrões associados à expertise.

3.3. NRAGs aplicados a bibliometria

Outra fronte atual de aplicação de NRAGs é no campo da bibliometria. NRAGs podem capturar padrões de diferentes áreas científicas. Eles podem ser usados para comparar a diversidade de áreas diferentes, identificar interseções entre áreas, estudar a evolução de métricas de redes complexas ao longo do tempo, etc. A Figura 4 compara NRAGs gerados a partir de um corpus de resumos de artigos de diferentes áreas da computação. O grafo superior representa as áreas de inteligência artificial e bibliotecas digitas, com pouca interseção entre as regiões neurais (nós do grafo). Já o grafo inferior representa as áreas de inteligência artificial e visão computacional, com muita intersecção entre as áreas como seria esperado.

4. Conclusão

Este artigo foca na aplicação dos NRAGs em pesquisas científicas de explicabilidade em LLMs. Espera-se que o NRAGs sejam um mecanismo relevante para explicabilidade de LLMs, se aproveitando de décadas de pesquisas na área de análise de redes complexas. O aspecto de redução de dimensionalidade da proposta permite analisar modelos mais complexos com menos recursos computacionais, um fator relevante no mundo atual de grande desigualdade em poder computacional entre academia e indústria.

NRAGs estão sendo aplicados em diversas frentes. Além da análise de expertise e bibliometria apresentadas aqui, outras pesquisas estão se iniciando nas áreas de análise de discurso e análise de padrões de aprendizado de tarefas a partir de fine-tuning de LLMs. Outro objetivo importante é expor os artefatos gerados a especialistas de domínio para avaliação e interpretação. Uma exploração em andamento é compreender as diferenças entre grafos gerados por modelos *encoders* e *decoders*, os últimos tendo resultados significativamente inferiores numa exploração inicial.

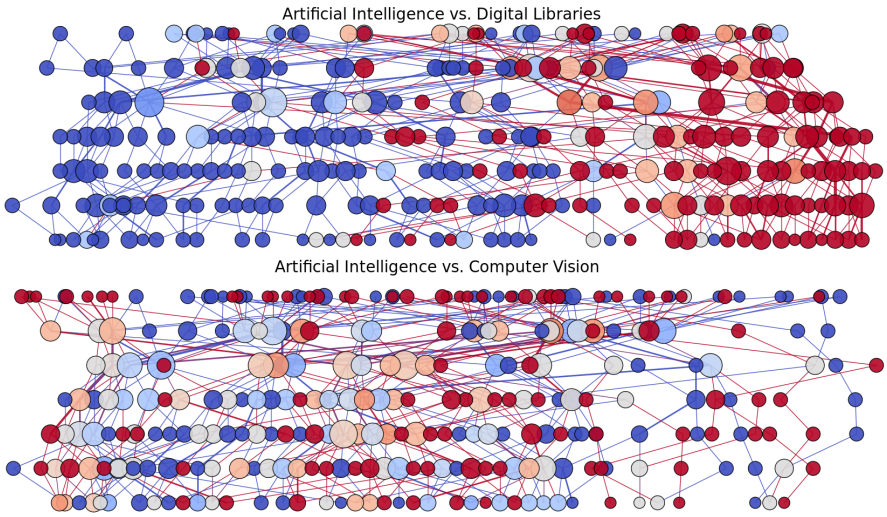

Figura 4. NRAGs comparando diferentes áreas da computação

Diversas melhorias na fundamentação dos NRAGs e na implementação da biblioteca LLM-MRI estão em planejamento e execução, incluindo: (i) melhorias gerais de eficiência da LLM-MRI, como melhor aproveitamento de GPUs nos processamentos e documentação, (ii) exploração de métricas de redes complexas para serem disponibilizadas nas análises, (iii) interatividade e interpretabilidade de nós, identificando palavras e/ou conceitos importantes para a respectiva região neural, (iv) inclusão de análise temporal, possibilitando a análise das métricas ao longo do tempo.

Referências

Bau, A., Belinkov, Y., Sajjad, H., Durrani, N., Dalvi, F., and Glass, J. (2018). Identifying and controlling important neurons in neural machine translation.

Bengio, Y., Ducharme, R., and Vincent, P. (2000). A neural probabilistic language model. In Leen, T., Dietterich, T., and Tresp, V., editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press.

Costa, L., Figênio, M., Santanchè, A., and Gomes-Jr, L. (2024). LLM-MRI python module: a brain scanner for llms. In *Anais Estendidos do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 125–130, Porto Alegre, RS, Brasil. SBC.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. (2023). Sparse autoencoders find highly interpretable features in language models.

Dalvi, F., Durrani, N., Sajjad, H., Belinkov, Y., Bau, A., and Glass, J. (2019). What is one grain of sand in the desert? analyzing individual neurons in deep nlp models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6309–6317.

DeRose, J. F., Wang, J., and Berger, M. (2020). Attention flows: Analyzing and comparing attention mechanisms in language models.

Figênio, M., Santanché, A., and Gomes-Jr, L. (2024a). The impact of activation patterns in the explainability of large language models – a survey of recent advances. In *Anais*

da XIX Escola Regional de Banco de Dados, pages 141–149, Porto Alegre, RS, Brasil. SBC.

Figênio, M. R. and Gomes-Jr, L. (2023). Ética na era dos modelos de linguagem massivos (llms): um estudo de caso do chatgpt. In *Anais da XVIII Escola Regional de Banco de Dados (ERBD 2023)*, volume 0, page 100, Brasil.

Figênio, M. R., Santanché, A., and Gomes-Jr, L. (2024b). The impact of activation patterns in the explainability of large language models - a survey of recent advances. In *Anais da XIX Escola Regional de Banco de Dados (ERBD 2024)*, page 141, Brasil.

Hiter, S. (2024). Top 20 generative ai tools and applications in 2024. Disponível em: `https://www.eweek.com/artificial-intelligence/generative-ai-apps-tools/`.

Hoover, B., Strobelt, H., and Gehrmann, S. (2019). exbert: A visual analysis tool to explore learned representations in transformers models.

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.

L. da F. Costa, F. A. Rodrigues, G. T. and Boas, P. R. V. (2007). Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242.

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.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., and Mian, A. (2024). A comprehensive overview of large language models.

Samek, W., Wiegand, T., and Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models.

Schmidt, H. G. and Rikers, R. M. J. P. (2007). How expertise develops in medicine: knowledge encapsulation and illness script formation. *Medical Education*, 41(12):1133–1139.

Tunstall, L., Von Werra, L., and Wolf, T. (2022). *Natural language processing with transformers*. "O'Reilly Media, Inc.".

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Zhang, B., He, Z., and Lin, H. (2024). A comprehensive review of deep neural network interpretation using topological data analysis. *Neurocomputing*, 609:128513.

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., and Du, M. (2024a). Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.* Just Accepted.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., and Wen, J.-R. (2024b). A survey of large language models.