



UNIVERSIDADE FEDERAL DO RIO GRANDE - FURG
CENTRO DE CIÊNCIAS COMPUTACIONAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO
CURSO DE MESTRADO EM ENGENHARIA DE COMPUTAÇÃO

Dissertação de Mestrado

**Da XAI Funcional à Interpretabilidade Mecanicista:
Investigando os Circuitos da Arbitragem de
Conhecimento em Modelos de IA Generativa**

Karine de Oliveira Fonseca

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal do Rio Grande - FURG, como requisito parcial para a obtenção do grau de Mestre em Engenharia de Computação

Orientador: Prof. Dr. Leonardo Emmendorfer

Rio Grande, 2025

F448d	Fonseca, Karine de Oliveira Da XAI funcional a interpretabilidade mecanicista: Investigando os circuitos da arbitragem de conhecimento em modelos de IA generativa / Karine de Oliveira Fonseca.. – 2025. 92 f. Dissertação (Mestrado) – Universidade Federal do Rio Grande – Programa de Pós-Graduação em Computação, 2025. Orientador: Dr. Leonardo Emmendorfer 1.Computação. 2. Inteligência Artificial. 3. Interpretabilidade mecanicista. 4. Sequestro de circuito. I. Emmendorfer, Leonardo. II. Título. CDU 004
-------	---

Universidade Federal do Rio Grande
Centro de Ciências Computacionais
Programa de Pós-Graduação em Computação
Curso de Mestrado em Engenharia de Computação

DISSERTAÇÃO DE MESTRADO

**Da XAI Funcional à Interpretabilidade Mecanicista: Investigando os
Circuitos da Arbitragem de Conhecimento em Modelos de IA Generativa**

Karine de Oliveira Fonseca

Banca examinadora:

 Documento assinado digitalmente
LEONARDO RAMOS EMMENDORFER
Data: 04/11/2025 17:17:59-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Leonardo Ramos Emmendorfer Orientador

 Documento assinado digitalmente
LUIS OTONI MEIRELES RIBEIRO
Data: 05/11/2025 20:15:21-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Luis Otoni Meireles Ribeiro (IFSUL)

 Documento assinado digitalmente
GISELE MORAES SIMAS
Data: 05/11/2025 10:27:19-0300
Verifique em <https://validar.iti.gov.br>

Prof^a. Dr^a. Gisele Moraes Simas (FURG)

RESUMO

FONSECA, Karine de Oliveira. **Da XAI Funcional à Interpretabilidade Mecanicista: Investigando os Circuitos da Arbitragem de Conhecimento em Modelos de IA Generativa**. 2025. 91f. Dissertação (Mestrado) – Programa de Pós-Graduação em Computação. Universidade Federal do Rio Grande - FURG, Rio Grande.

Os Modelos de Linguagem de Larga Escala (LLMs) demonstram uma capacidade sem precedentes de processar e gerar informação, mas sua natureza intrinsecamente opaca representa um risco significativo para aplicações críticas. A Inteligência Artificial Explicável (XAI) busca mitigar essa opacidade, contudo, as abordagens atuais frequentemente se limitam a uma análise funcional, sem decifrar os mecanismos computacionais internos que governam o comportamento do modelo. Esta dissertação ataca essa lacuna através da interpretabilidade mecanicista, investigando uma questão fundamental e ainda pouco explorada: como um LLM arbitra entre seu conhecimento factual interno (parametrizado) e a informação contextual externa, especialmente quando há um conflito entre eles. Para tal, foi conduzido um estudo experimental computacional utilizando a técnica de Rastreamento Causal (Causal Tracing). Através de um paradigma de "Conflito vs. Confirmação", o estudo mede a modulação na influência causal dos circuitos neuronais responsáveis por armazenar fatos específicos quando o modelo é exposto a contextos confirmatórios ou conflitantes. Os achados oferecem uma nova perspectiva dinâmica para o modelo teórico das MLPs (Perceptron de Múltiplas Camadas) como memórias chave-valor. Demonstra-se que a recuperação de conhecimento não é um processo estático, mas sim análogo e dependente de contexto: a influência causal do circuito é massivamente reforçada por um contexto confirmatório. Em contraste, um contexto conflitante não suprime a influência do circuito, mas desencadeia o mecanismo proposto nessa dissertação denominado *Circuit Hijacking* 'sequestro de circuito', no qual a informação externa é processada no mesmo substrato neural do conhecimento original. Esta visão dinâmica do processamento factual tem implicações teóricas para a compreensão de como os LLMs integram evidências e implicações práticas para a segurança de sistemas de Geração Aumentada por Recuperação (RAG).

Palavras-chave: Sequestro de Circuito, Interpretabilidade Mecanicista, Segurança em IA, Modelos de Linguagem de Larga Escala (LLM), Sobrescrita Contextual, Intervenção Causal, Geração Aumentada por Recuperação (RAG).

ABSTRACT

FONSECA, Karine de Oliveira. **From Functional XAI to Mechanistic Interpretability: Investigating the Circuits of Knowledge Arbitration in Generative AI Models.** 2025. 91f. Dissertação (Mestrado) – Programa de Pós-Graduação em Computação. Universidade Federal do Rio Grande - FURG, Rio Grande.

Large Language Models (LLMs) have demonstrated an unprecedented capacity for processing and generating information, yet their intrinsic opacity poses a substantial risk in high-stakes domains. While Explainable Artificial Intelligence (XAI) has sought to address this challenge, most existing approaches remain limited to functional analyses, leaving the underlying computational mechanisms that shape model behavior largely unexplored. This dissertation advances the field of mechanistic interpretability by examining a fundamental question: how do LLMs arbitrate between their parameterized internal knowledge and external contextual evidence, particularly when the two are in conflict? To do so, we conducted a computational experiment employing Causal Tracing. Through a “Conflict vs. Confirmation” paradigm, we systematically measured the modulation of causal influence across neural circuits responsible for factual knowledge when exposed to confirmatory or contradictory contexts. The results suggest a dynamic reframing of the canonical view of MLPs as key-value memories. Specifically, factual retrieval is not a static operation, but rather a context-dependent process: confirmatory evidence substantially amplifies the causal influence of factual circuits, whereas contradictory evidence does not suppress them but instead triggers a mechanism we term Circuit Hijacking. In this mechanism, external information is processed within the same neural substrate that encodes the original knowledge. This dynamic account of factual processing yields crucial theoretical implications for understanding how LLMs integrate competing evidence, and highlights practical implications for the security of Retrieval-Augmented Generation (RAG) systems.

Keywords: Circuit Hijacking, Mechanistic Interpretability, AI Safety, Large Language Models (LLM), Contextual Overwriting, Causal Intervention, Retrieval-Augmented Generation (RAG).

LISTA DE FIGURAS

1	A fronteira da pesquisa em XAI: a transição da análise funcional (o que o modelo faz) para a interpretabilidade mecanicista (como o modelo computa uma decisão).	10
2	Funil conceitual da pesquisa, demonstrando a progressão do campo de estudo até a técnica específica investigada	11
3	Protocolo PRISMA	16
4	Taxonomia das abordagens de XAI para LLMs, conforme identificado nesta Revisão Sistemática da Literatura.	17
5	Aplicação de Métodos de XAI em modelos de Caixa Preta.	22
6	Relação conceitual entre Interpretabilidade e Explicabilidade, destacando suas naturezas distintas (passiva vs. ativa) e seu objetivo comum de promover a confiança e a compreensão humana.	23
7	Modelos Caixa Branca e Caixa Preta	25
8	Pirâmide Conceitual dos Níveis de Profundidade das Técnicas	28
9	Arquitetura Transformer	32
10	Pipeline metodológico da pesquisa em interpretabilidade mecanicista.	36
11	Processo de Rastreamento Causal	38
12	Quadro-Resumo das Principais Técnicas de Interpretabilidade Mecanicista	40
13	Diagrama do delineamento experimental, mostrando a aplicação das três condições contextuais (variável independente) ao objeto de estudo (GPT-2 XL) e a subsequente medição das variáveis dependentes (saída comportamental e influência causal).	43
14	Diagrama do Delineamento Experimental.	45
15	Fluxograma da análise de dados de métodos mistos (mixed-methods), mostrando a convergência das análises quantitativa (Teste de ANOVA) e qualitativa (Estudo de Caso) na síntese dos resultados.	46
16	Heatmap do Efeito Causal Médio da MLP para o prompt em português "A Catalunha pertence ao continente...". O padrão de influência difuso e a predição final incorreta (p(a)) indicam uma representação de conhecimento não localizada.	48

17	Heatmap do Efeito Causal Médio da MLP para o prompt em inglês "Catalonia belongs to...". Observa-se um 'hotspot' de influência nítido e localizado no último token do sujeito (onia*) e nas camadas intermediárias.	48
18	Análise do comportamento do modelo na condição de Conflito para o fato 8 (id 7).	49
19	Heatmap do Efeito Causal da MLP na condição de Conflito para o fato 8 (id 7) evidenciando o mecanismo de Circuit Hijacking	49
20	Comparação do comportamento do modelo para o fato 18 na análise de ablação do fraseado do prompt. O painel (a) mostra o resultado com a instrução explícita Based on the context, enquanto o painel (b) mostra o resultado sem a instrução. Nota-se que a presença da instrução aumenta a confiança do modelo na resposta contextual incorreta, de 35.13% para 43.04%.	50
21	Apresenta uma comparação direta do comportamento do modelo para este fato, sem a instrução "based on the context" explícita no prompt de Conflito.	51
22	Estatísticas Descritivas dos Escores de Efeito Causal Máximo por Condição.	52
23	Distribuição dos Escores de Efeito Causal Máximo (MLP) por Condição Experimental.	52
24	Comparação da distribuição dos escores para a condição de Conflito, sem e com a instrução explícita no prompt.	53
25	Comparação dos heatmaps médios para a condição de Conflito. O painel (a) mostra o resultado com a instrução explícita e o painel (b) sem a instrução.	59

LISTA DE ABREVIATURAS E SIGLAS

ANOVA	Analysis of Variance (Análise de Variância)
CoT	Chain Of Thought (Cadeia de Pensamento)
GenAI	Generative Artificial Intelligence (Inteligência Artificial Generativa)
GAN	Generative Adversarial Network (Rede Adversarial Generativa)
GPT	Generative Pre-trained Transformer (Transformador Generativo Pré-treinado)
IA	Inteligência Artificial
IM	Interpretabilidade Mecanicista
KG	Knowledge Graph (Grafo de Conhecimento)
LIME	Local Interpretable Model-agnostic Explanations
LLM	Large Language Model (Modelo de Linguagem de Larga Escala)
LSTM	Long Short-Term Memory (Memória de Curto e Longo Prazo)
MLP	Perceptron de Múltiplas Camadas
NLE	Natural Language Explanations (Explicações em Linguagem Natural)
PCA	Principal Component Analysis (Análise de Componentes Principais)
PDP	Partial Dependence Plot (Gráfico de Dependência Parcial)
PINN	Physics-Informed Neural Network (Rede Neural Informada pela Física)
PLN	Processamento de Linguagem Natural
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RAG	Retrieval-Augmented Generation (Geração Aumentada por Recuperação)
ReLU	Rectified Linear Unit (Unidade Linear Retificada)
RLHF	Reinforcement Learning from Human Feedback (Aprendizado por Reforço com Feedback Humano)
RNN	Recurrent Neural Network (Rede Neural Recorrente)
ROME	Rank-One Model Editing
RSL	Revisão Sistemática da Literatura
SHAP	SHapley Additive exPlanations
XAI	Explainable Artificial Intelligence (Inteligência Artificial Explicável)

SUMÁRIO

1	Introdução	9
1.1	Problema de Pesquisa	9
1.2	Justificativa	11
1.3	Objetivo Geral e Objetivos Específicos	12
1.3.1	Objetivo Geral	12
1.3.2	Objetivos Específicos	12
1.4	Estrutura da Dissertação	12
2	Revisão Sistemática da Literatura	13
2.1	Metodologia da Revisão Sistemática da Literatura	14
2.1.1	Fontes de Dados e Estratégia de Busca	14
2.1.2	Processo de Seleção e Critérios de Elegibilidade (Triagem)	14
2.1.3	Resultados	15
2.1.4	Discussão	19
2.1.5	Conclusão da Revisão e Justificativa para uma Abordagem Mecanicista	19
2.1.6	Trabalhos Relacionados na Fronteira Mecanicista	20
3	Referencial Teórico	21
3.1	Inteligência Artificial Explicável (XAI)	21
3.1.1	Explicabilidade vs. Interpretabilidade	22
3.1.2	Tipos de Modelos e Abordagens em XAI	25
3.1.3	Profundidade das Técnicas	27
3.1.4	Técnicas e Métodos de XAI	28
3.2	Arquitetura Transformer e Modelos de Linguagem de Larga Escala (LLMs)	31
3.2.1	A Arquitetura Transformer	31
3.2.2	Modelos de Linguagem de Larga Escala - Large Language Models (LLMs)	33
3.3	Inteligência Artificial Generativa	35
3.4	Interpretabilidade Mecanicista	35
3.4.1	O Desafio da Superposição e a Busca por Features Monosemânticas	35
3.4.2	Estratégias para a Interpretação de Circuitos	36
3.4.3	Foco Metodológico: O Algoritmo de Rastreamento Causal (Causal Tracing)	37
3.4.4	O Ecossistema da Análise Mecanicista: Ferramentas Complementares	38
3.5	Síntese e Posicionamento do Trabalho	39

4 Metodologia e Desenho do Estudo	42
4.1 Desenho Experimental	42
4.2 Objeto de Estudo e Instrumentação	43
4.3 Validação do Corpus de Fatos e Construção dos Estímulos	44
4.4 Protocolo Experimental	44
4.5 Análise dos Dados	45
5 Resultados	47
5.1 Análise Exploratória e Estudos de Caso Qualitativos	47
5.1.1 A Importância da Língua: Um Estudo de Caso (Inglês vs. Português)	47
5.1.2 O Fenômeno <i>Circuit Hijacking</i> (Sequestro de Circuito): Um Estudo de Caso	48
5.1.3 A Sensibilidade ao Fraseado: O Caso 'Based on the context'	50
5.2 Análise Visual Agregada (Os Heatmaps Médios)	51
5.3 Análise Quantitativa e Teste de Hipóteses (ANOVA)	51
5.3.1 Análise de Ablação: O Efeito do Fraseado do Prompt	53
6 Discussão	55
6.1 Interpretação dos Achados Principais	55
6.1.1 O Efeito de Reforço: A Confirmação Contextual como Amplificador Causal	56
6.1.2 A Arbitragem de Conflito: O Mecanismo de <i>Circuit Hijacking</i> (Sequestro de Circuito)	56
6.2 Implicações dos Resultados	58
6.2.1 Implicações Teóricas: Rumo a um Modelo Dinâmico dos Circuitos	58
6.2.2 Implicações Práticas: A Fragilidade da Representação Multilíngue e o Papel da Tokenização	58
6.2.3 Implicações para a Engenharia de Prompts: A Sensibilidade ao Fraseado	59
6.3 Diálogo com a Explicabilidade Funcional: O Papel e os Limites do <i>Chain of Thought</i> (CoT)	60
6.4 Conectando os Achados à Literatura	61
7 Conclusão e Trabalhos Futuros	63
7.1 Conclusão	63
7.2 Implicações Éticas e de Segurança Derivadas dos Achados da Pesquisa	63
7.3 Limitações do Estudo	65
7.4 Trabalhos Futuros	65
ANEXOS	
A Artigos Seleccionados na Revisão Sistemática da Literatura	67
B Corpus de Fatos	74
C Escores de Efeito Causal	81
Referências	85

1 INTRODUÇÃO

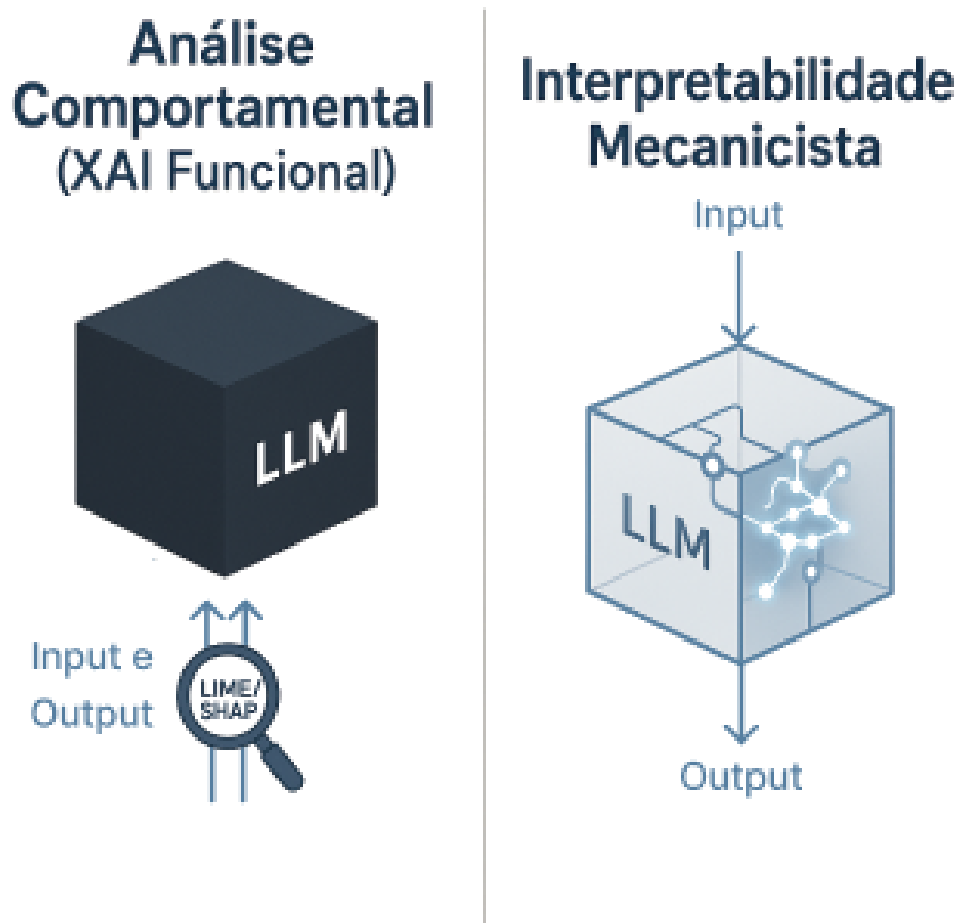
A ascensão dos Modelos de Linguagem de Larga Escala (LLMs) representa um marco na evolução da Inteligência Artificial (IA), redefinindo os paradigmas de interação humano-computador e automação de tarefas cognitivas. Esses modelos, frequentemente denominados "modelos de fundação" devido ao seu papel central no ecossistema de IA [8], são baseados na arquitetura Transformer [71] e demonstram uma capacidade sem precedentes de gerar texto coeso, traduzir idiomas e raciocinar sobre informações complexas. Contudo, essa proficiência é acompanhada por um desafio proporcional que tem sido amplamente discutido na literatura: sua natureza intrinsecamente opaca [8]. A complexidade arquitetônica dos LLMs, envolvendo bilhões de parâmetros interconectados em redes neurais profundas, resulta em um sistema de "caixa-preta", cujo processo decisório é largamente inescrutável para seus próprios criadores e usuários.

Essa opacidade gera riscos significativos, que vão desde a propagação de desinformação e a perpetuação de vieses sociais ocultos nos dados de treinamento [7], até a falta de confiabilidade em aplicações de alto risco, como diagnóstico médico ou assessoria jurídica, onde a confiança e a responsabilização são imperativas [18]. Em resposta a este desafio fundamental, emergiu o campo da Inteligência Artificial Explicável (XAI), cujo objetivo é desenvolver métodos e técnicas capazes de tornar as decisões e os mecanismos dos sistemas de IA compreensíveis aos seres humanos [5].

1.1 Problema de Pesquisa

Embora a XAI tenha avançado para modelos mais tradicionais, sua aplicação a LLMs é um campo incipiente e em rápida expansão. Uma Revisão Sistemática da Literatura, conduzida como etapa inicial desta pesquisa (Capítulo 2), revelou que as abordagens atuais de XAI para LLMs, embora diversas, concentram-se predominantemente em um nível funcional ou comportamental. Técnicas como LIME e SHAP, por exemplo, são eficazes em explicar "o que" o modelo faz, atribuindo importância às *features* de entrada, mas falham em decifrar "como" e "por que" seus mecanismos internos computam uma determinada decisão [60].

Figura 1: A fronteira da pesquisa em XAI: a transição da análise funcional (o que o modelo faz) para a interpretabilidade mecanicista (como o modelo computa uma decisão).



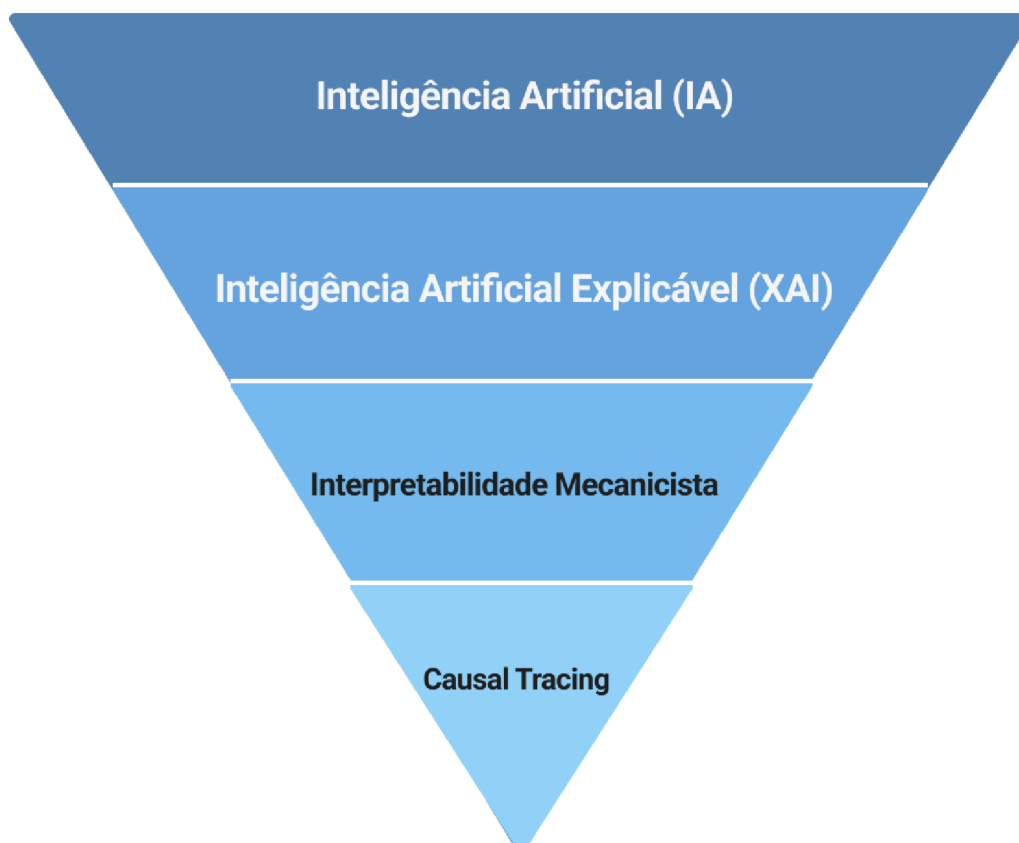
Fonte:Elaborado pela autora (2025).

Essa análise expôs uma lacuna fundamental na fronteira da pesquisa: a transição da explicação funcional para a interpretabilidade mecanicista [53] (Figura 1). Este sub-campo emergente busca "fazer a engenharia reversa" dos LLMs para identificar os circuitos neurais exatos responsáveis por habilidades específicas. Trabalhos seminais nesta área focaram em provar que é possível localizar conhecimento factual em um modelo estático [23] [44]. Contudo, permanece uma questão em aberto e de crucial importância prática: como os circuitos neuronais que armazenam conhecimento factual reagem dinamicamente quando o modelo é confrontado com informações externas, sejam elas confirmatórias ou conflitantes? A ausência de uma compreensão mecanicista sobre essa arbitragem entre conhecimento interno (parametrizado) e evidência externa (contextual) constitui o problema central desta investigação. Portanto, a presente dissertação é norteada pela seguinte pergunta de pesquisa:

Como a influência causal dos circuitos neuronais associados a um fato é modulada quando um LLM é exposto a um contexto externo que confirma ou contradiz seu conhecimento interno pré-existente?

Para abordar esta questão, a presente pesquisa adota um escopo metodológico que se afunila progressivamente, partindo de conceitos gerais da Inteligência Artificial até as técnicas de intervenção causal que permitem investigar o problema. A Figura 2 ilustra este percurso conceitual.

Figura 2: Funil conceitual da pesquisa, demonstrando a progressão do campo de estudo até a técnica específica investigada



Fonte:Elaborado pela autora (2025).

1.2 Justificativa

A relevância desta pesquisa se sustenta sobre um tripé de justificativas: teórica, prática e metodológica.

Justificativa Teórica: O estudo ataca uma lacuna singular na fronteira do conhecimento. Ele propõe mover a pesquisa em interpretabilidade mecanicista de uma análise estática (localização de fatos) para uma análise dinâmica (interação e arbitragem de conhecimento), contribuindo para uma compreensão fundamental de como os LLMs raciocinam, um desafio destacado por múltiplos pesquisadores [20].

Justificativa Prática: A investigação tem implicações diretas para a segurança e confiabilidade de sistemas de Geração Aumentada por Recuperação (RAG), a arquitetura predominante para a aplicação de LLMs no mundo real [36]. Sistemas de RAG operam sob a premissa de que o modelo irá priorizar e utilizar corretamente a informação recupe-

rada. Este estudo testa essa premissa fundamental no nível dos circuitos, sendo um passo essencial para a engenharia de sistemas de IA mais robustos e seguros.

Justificativa Metodológica: A recente disponibilidade de ferramentas de código aberto para intervenção causal, como a implementada no repositório ROME [44], torna metodologicamente viável a condução de um experimento controlado para investigar esta questão de forma rigorosa e reproduzível.

1.3 Objetivo Geral e Objetivos Específicos

1.3.1 Objetivo Geral

Investigar os mecanismos causais que governam a arbitragem entre o conhecimento factual interno de um LLM e a informação contextual externa, analisando como o contexto modula a influência dos circuitos neuronais responsáveis pelo conhecimento.

1.3.2 Objetivos Específicos

1. Mapear a localização e medir a influência causal dos circuitos neuronais associados a um conjunto de fatos em uma condição de controle (baseline).
2. Analisar o efeito de um contexto externo confirmatório sobre a influência causal dos circuitos de conhecimento identificados.
3. Analisar o efeito de um contexto externo conflitante sobre a influência causal dos mesmos circuitos.
4. Comparar os resultados das três condições para quantificar a supressão ou reforço do conhecimento interno e caracterizar o comportamento de arbitragem do modelo.

1.4 Estrutura da Dissertação

O presente trabalho está organizado em sete capítulos. O Capítulo 1 realiza a introdução, apresentando o tema, o problema de pesquisa, a justificativa e os objetivos do estudo. O Capítulo 2 detalha a Revisão Sistemática da Literatura conduzida para mapear o estado da arte e identificar a lacuna de pesquisa que motiva este trabalho. O Capítulo 3 estabelece a Fundamentação Teórica, com foco na arquitetura Transformer e nos princípios da Interpretabilidade Mecanicista, especialmente a técnica de Rastreamento Causal. Com base nesta teoria, o Capítulo 4 descreve a Metodologia da Pesquisa, incluindo o desenho experimental, o protocolo de validação do corpus e os métodos de análise de dados. Os resultados empíricos da investigação são apresentados e analisados no Capítulo 5. O Capítulo 6 se dedica à Discussão, interpretando os achados e explorando suas implicações teóricas e práticas. Finalmente, o Capítulo 7 apresenta a Conclusão do trabalho, suas limitações e aponta direções para pesquisas futuras.

2 REVISÃO SISTEMÁTICA DA LITERATURA

A ascensão dos modelos de Linguagem de Larga Escala (LLMs) representa um marco na evolução da Inteligência Artificial (IA), redefinindo os paradigmas de interação humano-computador e automação de tarefas cognitivas complexas. Modelos como os da família GPT (*Generative Pre-trained Transformer*) demonstram uma capacidade sem precedentes de gerar texto coeso, traduzir idiomas, responder a perguntas e criar conteúdo original. No entanto, essa proficiência é acompanhada por um desafio proporcional: a sua natureza intrinsecamente opaca. A complexidade arquitetônica dos LLMs, envolvendo bilhões de parâmetros interconectados em redes neurais profundas, resulta em um sistema de "caixa-preta", cujo processo de tomada de decisão é largamente inescrutável para seus próprios criadores e usuários. Essa opacidade gera riscos significativos, que vão desde a propagação de desinformação e a perpetuação de vieses sociais ocultos nos dados de treinamento, até a falta de confiabilidade em aplicações de alto risco, como diagnóstico médico, assessoria jurídica ou controle financeiro. Em resposta a este desafio fundamental, emergiu o campo da Inteligência Artificial Explicável (XAI), cujo objetivo é desenvolver métodos e técnicas capazes de tornar as decisões dos sistemas de IA compreensíveis aos seres humanos. Embora a XAI tenha avançado consideravelmente para modelos de aprendizado de máquina mais tradicionais, sua aplicação a LLMs é um campo de pesquisa incipiente e em rápida expansão. Uma análise preliminar da literatura revela uma diversidade de abordagens, sem uma categorização clara que organize as estratégias propostas. Essa lacuna no conhecimento dificulta a compreensão do estado da arte, impede a identificação de tendências consolidadas e limita a capacidade de pesquisadores e desenvolvedores de selecionar a abordagem de XAI mais adequada para seus objetivos. Mais profundamente, essa diversidade de abordagens ainda se concentra em explicar "o que" o modelo faz, sem decifrar "como" seus mecanismos internos computam a decisão. Nesse contexto, esta revisão sistemática da literatura foi norteadada pela seguinte pergunta de pesquisa: **"Quais são as principais categorias de técnicas de XAI propostas na literatura científica para explicar modelos de IA Generativa, especificamente LLMs?"** Essa abordagem sistemática possibilita graus mais elevados de reprodutibilidade e transparência nos resultados de pesquisa [29][48]. À medida que os estudiosos reconhecem

cada vez mais a complexidade desse rigor metodológico, torna-se essencial aderir a protocolos estabelecidos para manter a qualidade e a confiabilidade ao longo do processo de pesquisa [29]. A presente investigação justifica-se pela necessidade premente de estruturar o conhecimento emergente sobre a explicabilidade de LLMs. Ao identificar, mapear e categorizar as abordagens existentes, este estudo visa fornecer um panorama claro do campo, subsidiar o desenvolvimento de sistemas de IA mais transparentes e confiáveis, e apontar direções frutíferas para pesquisas futuras, incluindo a viabilidade de investigações em interpretabilidade mecanicista.

2.1 Metodologia da Revisão Sistemática da Literatura

2.1.1 Fontes de Dados e Estratégia de Busca

A seleção criteriosa de bases de dados constitui uma etapa fundamental no delineamento metodológico de revisões sistemáticas da literatura, pois impacta diretamente a abrangência, a profundidade e a qualidade do mapeamento do conhecimento disponível sobre o tema investigado. É fundamental, ainda, que seja documentada a justificativa para a escolha de cada repositório, em consonância com o princípio de transparência que orienta as revisões sistemáticas [59] [66]. Portanto, para garantir uma cobertura representativa e multidisciplinar, foram selecionadas quatro bases de ampla relevância e reconhecimento internacional: **IEEE Xplore**, **ScienceDirect** e **Web of Science**.

O IEEE Xplore destaca-se pelo seu acervo em ciência da computação e engenharias, especialmente nas áreas de inteligência artificial e sistemas de informação. O ScienceDirect, mantido pela Elsevier, oferece uma vasta coleção de periódicos revisados por pares, abrangendo desde as ciências exatas até as ciências aplicadas, e é frequentemente utilizado em pesquisas tecnológicas e interdisciplinares.

Por fim, o Web of Science, por sua vez, é reconhecido por sua cobertura abrangente e criteriosa de periódicos de alto impacto, permitindo a identificação de tendências e lacunas em diferentes áreas do conhecimento. A conjugação dessas bases busca assegurar uma revisão exaustiva, transparente e alinhada aos rigorosos protocolos metodológicos recomendados para revisões sistemáticas da literatura. Além disso, o acesso a conteúdos controlados e curados por meio de repositórios institucionais ou bases especializadas tende a elevar o padrão de qualidade da literatura incluída nas RSLs[6].

2.1.2 Processo de Seleção e Critérios de Elegibilidade (Triagem)

Este estudo foi conduzido seguindo rigorosamente as diretrizes do Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020, a fim de garantir a transparência, a reprodutibilidade e a integridade do processo de revisão. Foram estabelecidos critérios de inclusão e exclusão para a seleção dos estudos, alinhados à pergunta de pesquisa.

Crítérios de Inclusão:

(CI1) O estudo aborda diretamente técnicas, métodos ou frameworks de IA Explicável (XAI), interpretabilidade ou transparência.(CI2) O foco da explicação é um Modelo de Linguagem de Grande Escala (LLM) ou um modelo de IA Generativa.(CI3) O estudo é um artigo científico, revisão sistemática, survey, artigo de posição ou capítulo de livro publicado em anais de conferências ou periódicos revisados por pares.

Crítérios de Exclusão:

(CE1) O estudo foca primariamente em uma aplicação de LLMs (ex: saúde, finanças) e apenas menciona XAI superficialmente.(CE2) A técnica de XAI é aplicada exclusivamente a modelos de IA que não são LLMs (ex: CNNs, SVMs).(CE3) O LLM é utilizado como uma ferramenta para explicar outro sistema, em vez de ser o objeto da explicação.(CE4) Editoriais, resumos, propostas de pesquisa sem resultados consolidados, ou literatura cinzenta.

As strings de busca principais, adaptadas para a sintaxe de cada base, foi: ("Explainable AI"OR "XAI"OR "Interpretability"OR "Transparency") AND ("Large Language Model"OR "LLM"OR "Generative AI")

("IA Explicável"OR "Interpretabilidade"OR "Interpretabilidade de modelos") AND ("Modelos de Linguagem de Grande Porte"OR "Modelos de linguagem"OR "Modelos de linguagem de grande escala")

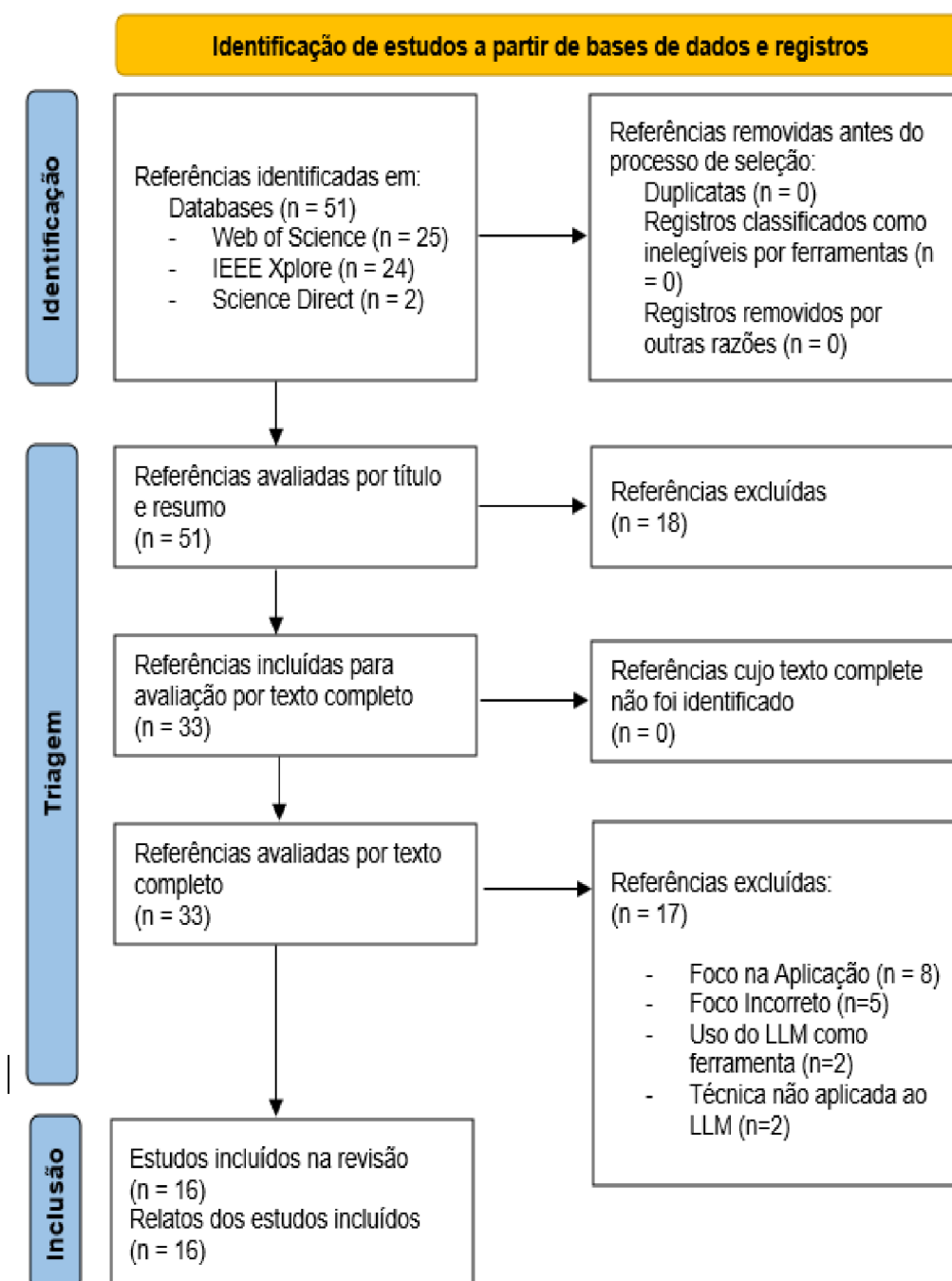
Foram aplicados filtros para incluir estudos publicados entre 2018 e 2025, refletindo a natureza recente e a rápida evolução do tema. O processo de seleção ocorreu em três fases sequenciais: **Identificação e Remoção de Duplicatas:** Os resultados de todas as bases de dados foram consolidados e as duplicatas foram removidas. **Triagem por Título e Resumo:** Os artigos restantes foram triados com base na leitura de seus títulos e resumos, aplicando-se os critérios de elegibilidade. **Avaliação do Texto Completo:** Os artigos que passaram pela triagem inicial foram lidos na íntegra para uma avaliação final de sua elegibilidade.

2.1.3 Resultados

A busca inicial nas bases de dados resultou em 51 registros. Após a remoção de 0 duplicatas, 51 artigos foram submetidos à triagem por título e resumo, da qual resultou na exclusão de 18 estudos. Os 33 artigos restantes foram avaliados na íntegra para elegibilidade. Nesta fase, 17 estudos foram excluídos. As principais razões para a exclusão na fase de leitura completa foram: foco principal na aplicação de LLMs em um domínio específico (n=8), foco incorreto da pesquisa (ex: métricas de avaliação, ética) (n=5), e o uso do LLM como ferramenta para explicar outro sistema (n=2). Ao final do processo, 16 estudos preencheram todos os critérios de elegibilidade e foram incluídos na síntese qualitativa. O fluxo completo do processo de seleção está detalhado no fluxograma PRISMA (**Figura 3**). A análise aprofundada dos 16 estudos permitiu a indução de cinco categorias

principais (Figura 4) que representam as abordagens contemporâneas para a explicabilidade de LLMs. Os artigos selecionados estão presentes no ANEXO A.

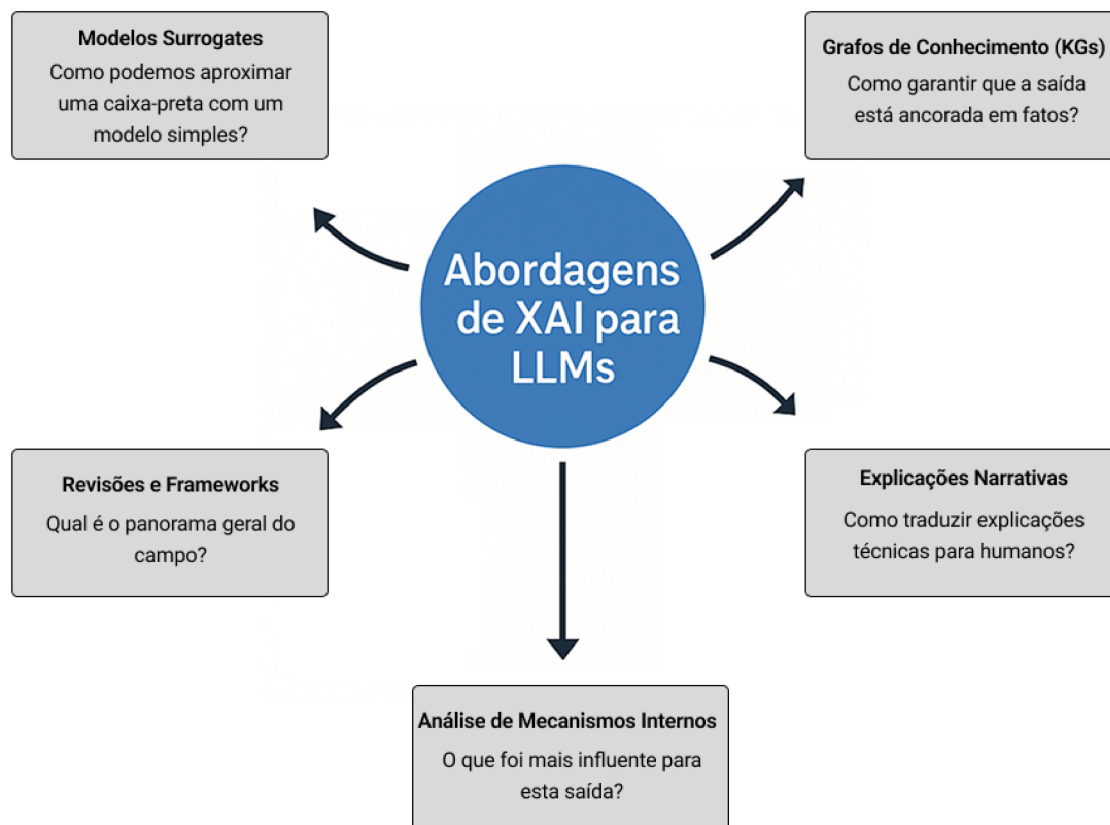
Figura 3: Protocolo PRISMA



Fonte:Elaborado pela autora (2025).

2.1.3.1 Categorias

Figura 4: Taxonomia das abordagens de XAI para LLMs, conforme identificado nesta Revisão Sistemática da Literatura.



Fonte:Elaborado pela autora (2025).

Categoria 1: Explicações por Surrogates e Destilação de Conhecimento Esta abordagem contorna a complexidade direta dos LLMs ao treinar um modelo secundário, mais simples e inerentemente interpretável — como um programa lógico ou uma árvore de decisão — para mimetizar o comportamento do modelo original. A explicação é então derivada deste “modelo substituto”(surrogate). Artigos como “From large language models to small logic programs: building global explanations from disagreeing local post-hoc explainers” exemplificam essa estratégia ao extrair programas lógicos globais a partir de múltiplos explicadores locais. Uma vertente especializada, a destilação de conhecimento, é apresentada em “Xai-driven knowledge distillation of large language models for efficient deployment on low-resource devices”, onde o conhecimento de um LLM “professor” é transferido para um modelo “aluno” mais compacto. Esta abordagem responde à pergunta: “Como podemos aproximar o comportamento de uma caixa-preta com um modelo simples e compreensível?”.

Categoria 2: Integração com Grafos de Conhecimento (Knowledge Graphs - KGs) Para combater a falta de fundamentação factual e as “alucinações” dos LLMs, esta

categoria de técnicas propõe a integração com Grafos de Conhecimento. KGs funcionam como uma fonte de verdade externa e estruturada, permitindo que as saídas do LLM sejam validadas e ancoradas em fatos verificáveis. A pesquisa “Synergistic Joint Model of Knowledge Graph and LLM for Enhancing XAI-Based” demonstra como essa sinergia melhora a acurácia e a explicabilidade em sistemas de apoio à decisão clínica. Similarmente, a revisão “Knowledge Graphs and Their Reciprocal Relationship with Large Language Models” explora a relação simbiótica entre os dois. Esta abordagem responde à pergunta: “Como podemos garantir que a saída do modelo está ancorada em fatos verificáveis?”.

Categoria 3: Explicações Narrativas e em Linguagem Natural Reconhecendo que as saídas de métodos de XAI tradicionais são frequentemente técnicas e pouco acessíveis, esta abordagem utiliza a própria capacidade generativa dos LLMs para “explicar as explicações”. O estudo “Tell me a story! Narrative-driven XAI with Large Language Models” introduz o conceito de “XAIstories”, que traduz métricas como valores SHAP em narrativas coesas e compreensíveis. Da mesma forma, “Human interpretable structure-property relationships in chemistry using explainable machine learning and large language models” aplica essa ideia ao campo da química, gerando explicações em linguagem natural sobre relações estrutura-propriedade, tornando o conhecimento científico mais acessível a não especialistas. Essencialmente, esta abordagem responde à pergunta: “Como podemos traduzir explicações técnicas para um formato útil e compreensível para humanos?”.

Categoria 4: Análise de Mecanismos Internos e Saliência de Features Esta categoria engloba abordagens que investigam diretamente a arquitetura dos LLMs, mas de uma perspectiva funcional. Técnicas de saliência, como LIME e SHAP, são adaptadas para identificar os tokens de entrada mais influentes para uma determinada saída (“Explaining Misinformation Detection”). Outros estudos analisam como os embeddings mudam com o ajuste fino (“Fine-tuning language model embeddings”) ou geram exemplos contrafatuais para testar o comportamento do modelo (“LLM in the Loop”). Essencialmente, esta abordagem responde à pergunta: “O que, na entrada ou na arquitetura superficial, foi mais influente para esta saída específica?”

Categoria 5: Revisões Abrangentes e Frameworks Conceituais Esta categoria final é composta por estudos que sintetizam o estado da arte. Trabalhos como “Explainable AI chatbots towards XAI ChatGPT” mapeiam tendências, enquanto o manifesto “Explainable Artificial Intelligence (XAI) 2.0” identifica problemas em aberto. Outros, como “GDPR and Large Language Models”, analisam o tema sob as óticas regulatória e ética, posicionando a XAI como um componente indispensável para a governança responsável da IA. Essencialmente, esta abordagem responde à pergunta: “Qual é o panorama geral, quais são os desafios e para onde o campo está indo?”

2.1.4 Discussão

Os resultados desta revisão sistemática revelam uma narrativa clara sobre a evolução do campo. A existência simultânea de categorias focadas em eficiência (Categoria 1), veracidade factual (Categoria 2), usabilidade humana (Categoria 3), atribuição de features (Categoria 4) e governança (Categoria 5) indica uma notável maturação. O campo está se movendo para além de uma definição monolítica de "explicação", reconhecendo que diferentes stakeholders possuem necessidades distintas e que a explicabilidade serve a múltiplos propósitos.

Essa expansão, no entanto, também expõe uma fronteira. As abordagens mapeadas, ainda que diversas, compartilham uma característica fundamental: elas operam em um nível funcional ou comportamental. Elas tratam o LLM como uma caixa que pode ser simplificada, ancorada em fatos externos, ou cujas relações de entrada-saída podem ser investigadas. Contudo, elas não abrem a caixa para entender o porquê de seu funcionamento em um nível fundamental. A análise da literatura demonstra que, enquanto o campo se tornou proficiente em responder "o que" o modelo fez e "como" seu comportamento se parece, ele permanece incipiente em responder por que, em nível de seus mecanismos computacionais subjacentes, ele raciocina de uma determinada maneira.

2.1.5 Conclusão da Revisão e Justificativa para uma Abordagem Mecanicista

A análise crítica das cinco categorias expõe, portanto, não apenas uma lista de direções futuras, mas uma lacuna singular e fundamental para o próximo salto qualitativo na área: a transição da explicação funcional para a interpretabilidade mecanicista.

A fronteira da pesquisa reside em ir além das correlações de saliência e dissecar a causalidade dos circuitos internos do modelo. É necessário mover-se da observação de ativações para a intervenção e manipulação de componentes neurais a fim de confirmar sua função causal em uma habilidade específica, como o raciocínio factual. Os artigos analisados mostram que, embora saibamos como verificar se um LLM usou uma fonte externa ou quais palavras foram importantes, não sabemos como o conhecimento é representado e processado internamente quando confrontado com informações externas, sejam elas confirmatórias ou conflitantes.

Sem entender os mecanismos do raciocínio, não se pode prever falhas catastróficas, corrigir vieses ou garantir que um modelo está seguindo uma cadeia lógica de raciocínio em vez de explorar atalhos estatísticos. Portanto, esta revisão sistemática não apenas mapeia o que foi feito, mas, mais importante, ela justifica cientificamente a necessidade de um novo foco de investigação. Ela estabelece o alicerce para o passo seguinte desta dissertação: um estudo experimental focado em desvendar, através da técnica de rastreamento causal, os circuitos computacionais que governam a integração de conhecimento factual em modelos de linguagem de larga escala.

2.1.6 Trabalhos Relacionados na Fronteira Mecanicista

A lacuna entre a análise funcional e a mecanicista, identificada como a conclusão principal desta revisão, define a fronteira atual da pesquisa em explicabilidade de LLMs. Os trabalhos que operam nesta fronteira, embora ainda incipientes, compartilham o objetivo de "fazer a engenharia reversa" dos mecanismos internos do modelo.

A abordagem desta dissertação dialoga diretamente com os estudos seminais nesta área. O trabalho de [44], por exemplo, introduziu a técnica de Rastreamento Causal para localizar estaticamente o conhecimento factual, provando que ele reside em componentes específicos das MLPs. De forma complementar, os estudos de [23] [15], utilizando métodos distintos, também apontaram as MLPs como repositórios de conhecimento, sob a ótica de memórias chave-valor e "neurônios do conhecimento", respectivamente.

Enquanto esses trabalhos se concentram em mapear a geografia estática do conhecimento, a presente dissertação se propõe a avançar sobre eles, utilizando suas mesmas ferramentas para investigar a dinâmica desse conhecimento. A pergunta sobre como os circuitos reagem a um contexto conflitante, que é o foco deste trabalho, representa o próximo passo lógico nesta linha de investigação. Com esta justificativa estabelecida, o capítulo seguinte detalhará a fundamentação teórica que sustenta esta abordagem.

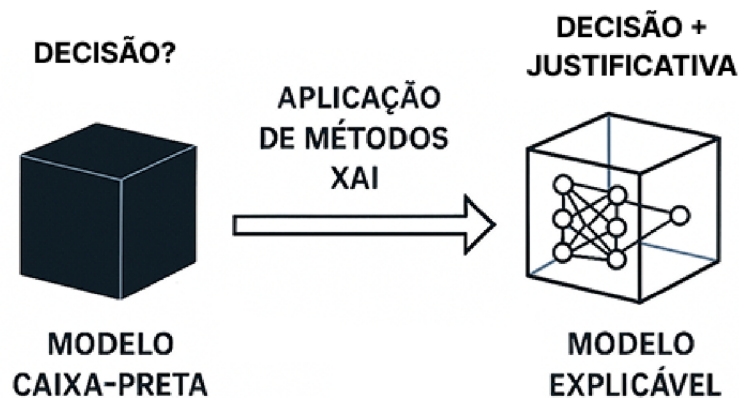
3 REFERENCIAL TEÓRICO

O presente capítulo tem como objetivo apresentar e discutir os principais conceitos e fundamentos teóricos que sustentam esta dissertação. Serão abordados temas centrais para o desenvolvimento da pesquisa, incluindo os fundamentos da Inteligência Artificial Explicável (XAI), os avanços na interpretabilidade mecanicista de modelos, o método ROME para edição de conhecimento em modelos de linguagem, bem como os aspectos essenciais relacionados aos Modelos de Linguagem de Larga Escala (LLMs) e a arquitetura Transformers. A sistematização destes tópicos busca fornecer uma compreensão aprofundada do estado da arte e das principais abordagens que embasam as discussões e experimentos realizados ao longo do estudo.

3.1 Inteligência Artificial Explicável (XAI)

O conceito de Inteligência Artificial Explicável (XAI) emergiu como uma resposta à crescente complexidade e opacidade dos sistemas de Inteligência Artificial (IA). À medida que as aplicações de IA se expandem em diversos setores, como saúde, finanças, transporte autônomo, entre outros, a necessidade de transparência nos processos decisórios dessas tecnologias torna-se cada vez mais fundamental. O objetivo da XAI é justamente desmistificar esses processos decisórios, permitindo que diferentes partes interessadas compreendam e confiem nos resultados gerados pelos sistemas de IA [69] [80] (Figura 5). Tiwari destaca que a XAI não apenas promove a confiança e o entendimento dos usuários, como também potencializa a colaboração significativa entre humanos e IA [69].

Figura 5: Aplicação de Métodos de XAI em modelos de Caixa Preta.



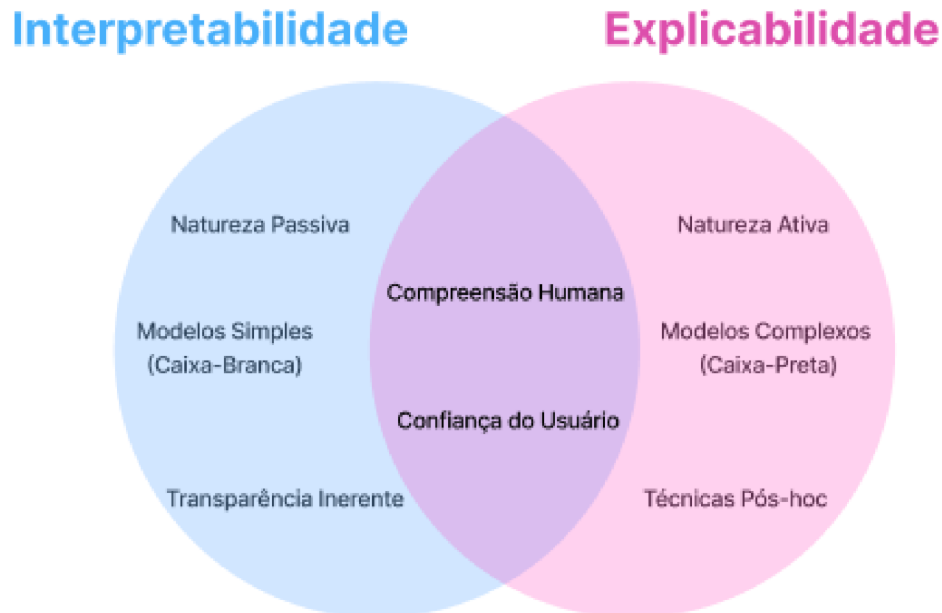
Fonte: Elaborado pela autora (2025).

A evolução da própria IA tem sido marcada por avanços tecnológicos significativos, desde modelos computacionais básicos estabelecidos em meados do século XX até os sofisticados algoritmos de aprendizado profundo prevalentes na atualidade. Essa trajetória evolutiva levou ao desenvolvimento de sistemas de IA capazes de superar o desempenho humano em tarefas específicas, embora sua natureza de “caixa-preta” frequentemente dificulte o entendimento acerca de como tais sistemas produzem soluções ou previsões. Por exemplo, a capacidade dos Sistemas de Apoio à Decisão Clínica baseados em IA (AI-CDSS) de auxiliar médicos no diagnóstico de condições de saúde evidencia a necessidade de explicações claras para as recomendações geradas por IA, assegurando a qualidade do cuidado prestado [22]. O avanço do aprendizado profundo e o surgimento de redes neurais impulsionaram significativamente essa evolução, sendo responsáveis pelas capacidades modernas da IA que, no entanto, permanecem ininteligíveis sem a aplicação de frameworks adequados de XAI [5].

3.1.1 Explicabilidade vs. Interpretabilidade

Os termos “explicabilidade” e “interpretabilidade” embora sejam frequentemente empregados de maneira intercambiável na literatura especializada em Inteligência Artificial (IA), particularmente no âmbito da IA Explicável (XAI), há distinções conceituais relevantes entre esses conceitos. O Diagrama de Venn na Figura 6 ilustra essas distinções, bem como a sobreposição e os objetivos comuns que unem os dois campos.

Figura 6: Relação conceitual entre Interpretabilidade e Explicabilidade, destacando suas naturezas distintas (passiva vs. ativa) e seu objetivo comum de promover a confiança e a compreensão humana.



Fonte:Elaborado pela autora (2025).

A distinção entre interpretabilidade e explicabilidade no contexto da inteligência artificial (IA) é fundamental para compreender como os sistemas de IA podem se tornar utilizáveis e confiáveis. Interpretabilidade refere-se, em geral, ao grau em que um ser humano pode compreender a causa de uma decisão tomada por um sistema de IA, especialmente em relação a modelos mais simples ou àqueles que apresentam relações claras entre entradas e saídas. Já a explicabilidade refere-se a mesma capacidade de compreensão porém voltada para modelos mais complexos e opacos.

As implicações desses conceitos estendem-se para aspectos éticos, legais e centrais no usuário da implantação da IA. Como destaca Miller, a explicabilidade torna-se crucial quando os sistemas são submetidos ao escrutínio de usuários que não possuem familiaridade com os algoritmos ou metodologias subjacentes [46]. Em contraste com a interpretabilidade, que pode concentrar-se na transparência de modelos mais simples, a explicabilidade frequentemente atende à necessidade de fornecer informações significativas a partir de algoritmos sofisticados, promovendo assim a confiança dos usuários [21] [24]. Ambos os conceitos diferem em seus objetivos e metodologias: enquanto a interpretabilidade enfatiza as características inerentes de um modelo que o tornam compreensível, a explicabilidade envolve a geração de saídas, como visualizações ou narrativas, adaptadas às necessidades e contextos de diferentes usuários. A compreensão robusta dessas diferenças é essencial para o desenvolvimento de sistemas de IA que sejam, simultanea-

mente, potentes e alinhados aos valores e necessidades humanas [43] [62].

3.1.1.1 Interpretabilidade (Interpretability)

Refere-se a uma característica passiva, indicando o grau em que um determinado modelo faz sentido para um observador humano. Rudin afirma que um modelo interpretável permite que as partes interessadas verifiquem de que maneira as características influenciam uma decisão específica, tornando-o adequado para aplicações em que há elevados requisitos de responsabilização [64]. Um modelo é considerado interpretável inerentemente se, por sua própria concepção ou arquitetura, ele é compreensível, destacando alguma ligação entre os espaços de entrada e saída, ou cujo mecanismo de previsão é óbvio. Exemplos de modelos com interpretabilidade inerente incluem árvores de decisão, modelos lineares e modelos aditivos generalizados (GAMs). A interpretabilidade pode ser caracterizada em diferentes níveis de transparência:

Simulabilidade: A capacidade de um modelo ser simulado mentalmente por um humano como um todo.

Decomponibilidade: A capacidade de explicar cada parte de um modelo (entrada, parâmetro e cálculo), onde cada parte deve ser compreensível por um humano sem ferramentas adicionais.

Transparência Algorítmica: Lida com a capacidade do usuário de entender o processo que o modelo seguiu para produzir uma saída a partir de sua entrada, geralmente por meio de análise matemática.

A interpretabilidade mede a extensão em que um modelo e/ou suas previsões são compreensíveis por humanos.

3.1.1.2 Explicabilidade (Explainability)

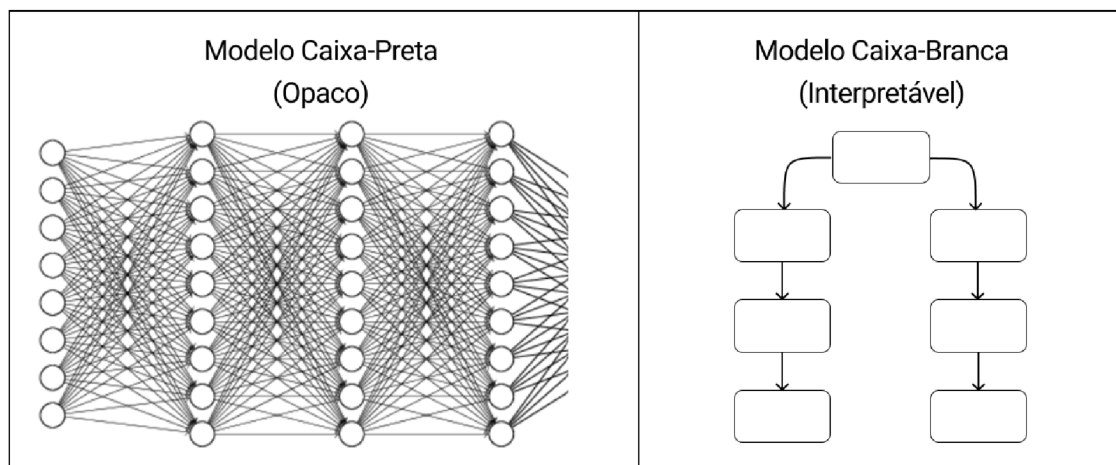
Refere-se a uma característica ativa, denotando qualquer ação ou procedimento realizado por um modelo com a intenção de esclarecer ou detalhar suas funções internas. Está associada à noção de explicação como uma interface entre humanos e um tomador de decisão, que é ao mesmo tempo um proxy preciso do tomador de decisão e compreensível para humanos. A explicabilidade é frequentemente ligada à explicabilidade pós-hoc, que são técnicas usadas para tornar um modelo não interpretável em um modelo explicável. Isso ocorre quando os modelos não são inerentemente transparentes por design, exigindo métodos externos para comunicar informações compreensíveis sobre como produzem suas previsões. O objetivo da explicabilidade é produzir detalhes ou razões para tornar o funcionamento do modelo claro ou fácil de entender para uma determinada audiência.

3.1.2 Tipos de Modelos e Abordagens em XAI

3.1.2.1 Modelos "Caixa Branca" vs. "Caixa Preta"

A distinção entre modelos “caixa branca” e “caixa preta” (Figura 7) é fundamental no campo da inteligência artificial explicável (XAI). Modelos caixa branca, como árvores de decisão e regressão linear, são caracterizados por sua inerente interpretabilidade; suas estruturas diretas permitem que os usuários compreendam os processos decisórios a partir das relações entre variáveis de entrada e resultados. Por exemplo, árvores de decisão representam visualmente os caminhos de decisões nos dados, facilitando para as partes interessadas o acompanhamento de como uma conclusão é atingida [5]. Por outro lado, modelos caixa preta, especialmente arquiteturas complexas como redes neurais profundas, são frequentemente opacos, tornando-se difíceis de interpretar. Esses modelos podem alcançar notável acurácia preditiva, porém às custas da transparência, o que impõe desafios para usuários que necessitam confiar e compreender decisões baseadas em IA, sobretudo em domínios [52] [28].

Figura 7: Modelos Caixa Branca e Caixa Preta



Fonte:Elaborado pela autora (2025).

A dificuldade de interpretar modelos caixa preta não apenas limita sua adoção em áreas críticas, mas também levanta preocupações éticas relativas à responsabilização e à justiça [50]. Como tais modelos são incapazes de fornecer justificativas claras para suas previsões, torna-se imperativo utilizar técnicas de XAI para reduzir essa lacuna de interpretabilidade. Isso motivou o desenvolvimento de métodos que possibilitam a análise do comportamento de modelos caixa preta sem alterar sua complexidade. Por exemplo, técnicas como LIME (Local Interpretable Model-agnostic Explanations) e SHAP (SHapley Additive exPlanations) fornecem meios para aproximar e explicar as saídas desses modelos, facilitando maior compreensão por parte dos usuários [37]. À medida que o cenário da IA evolui, o equilíbrio entre explicabilidade dos modelos e desempenho permanece como uma área crucial de investigação.

3.1.2.2 *Abordagens Locais vs. Globais*

Abordagens locais buscam esclarecer como o modelo chegou a uma predição específica para uma determinada entrada, elucidando os fatores que contribuíram para um resultado específico [17] [34].

Métodos como LIME e SHAP enquadram-se nessa categoria, permitindo que os usuários examinem detalhadamente casos individuais e compreendam a lógica do modelo em contextos específicos. Esse nível de granularidade é particularmente útil em áreas como a saúde, onde profissionais podem avaliar como características específicas do paciente influenciaram o diagnóstico ou a recomendação de tratamento do modelo [9]. Ao se concentrarem em explicações locais, as partes interessadas podem avaliar melhor a confiabilidade e a justiça do modelo em casos individuais, promovendo decisões mais informadas.

Em contraste, abordagens globais buscam explicar o comportamento geral do modelo, oferecendo insights sobre o funcionamento do modelo como um todo em diferentes entradas e saídas [37] [17]. Essas explicações são essenciais para compreender padrões ou vieses presentes no modelo, auxiliando os usuários a identificar possíveis pontos críticos que podem não ser evidentes na análise de previsões individuais. Por exemplo, explicações globais podem apontar quais atributos são mais influentes em todas as previsões, permitindo aos profissionais avaliar a equidade e a confiabilidade do desempenho do modelo [31] [50]. Dessa forma, abordagens globais contribuem para ampliar a generalização das interpretações, promovendo confiança e entendimento em contextos acadêmicos e práticos.

A escolha entre abordagens locais e globais depende frequentemente da aplicação pretendida e das necessidades informacionais dos stakeholders envolvidos. Em situações que exigem a justificativa de decisões individuais — como sentenças judiciais ou intervenções médicas críticas — explicações locais podem ser mais pertinentes [13]. Entretanto, a integração de ambas as abordagens pode fornecer uma visão mais abrangente sobre o comportamento da IA, oferecendo aos stakeholders uma estrutura mais robusta para avaliação do desempenho do modelo. Por meio de uma estratégia dual que incorpora explicações locais e globais, profissionais de IA podem navegar pelas complexidades da interpretabilidade e garantir que seus sistemas sejam transparentes, justos e alinhados a padrões éticos [2] [19]. Assim, o equilíbrio entre abordagens locais e globais enriquece a compreensão dos sistemas de IA, resultando em aplicações mais responsáveis e eficazes.

3.1.2.3 *Abordagens Pré-hoc vs. Pós-hoc*

A classificação dos métodos de XAI em abordagens pré-hoc e pós-hoc ressalta diferenças significativas quanto ao momento e à metodologia para alcançar a explicabilidade dos modelos. Abordagens pré-hoc envolvem o desenvolvimento de modelos que já são, desde o início, transparentes e interpretáveis [4] [19]. Exemplos incluem árvores de decisão e modelos lineares generalizados, que naturalmente possibilitam interpretações claras de seus

processos decisórios devido à sua simplicidade estrutural [16].

Por outro lado, abordagens pós-hoc são aplicadas a modelos ”caixa preta” já existentes, gerando explicações após o desenvolvimento e o treinamento do modelo [12]. Técnicas como LIME e SHAP exemplificam os métodos pós-hoc ao fornecerem *insights* sobre o comportamento do modelo com base nos dados processados [37][4]. Tais métodos são particularmente relevantes, pois permitem aos *stakeholders* analisar e compreender o funcionamento de modelos complexos sem a necessidade de reestruturá-los, preservando o desempenho enquanto atendem à demanda por explicabilidade [52]. Contudo, um desafio associado às técnicas pós-hoc reside na dependência da robustez do modelo original, já que a qualidade das explicações pode ser limitada pela opacidade e complexidade inerentes ao modelo [37].

Modelos pré-hoc, embora mais fáceis de interpretar, podem apresentar menor desempenho, enquanto métodos pós-hoc auxiliam na elucidação do funcionamento de modelos altamente performáticos, porém opacos [39][17]. Em última instância, a combinação de ambas as abordagens pode ser vantajosa, promovendo o desenvolvimento de sistemas de IA robustos e transparentes, que sigam práticas éticas e mantenham alta acurácia e desempenho em diferentes aplicações [2][3]. Essa abordagem multifacetada pode potencializar a eficácia da XAI, garantindo sistemas de IA tanto potentes quanto confiáveis.

3.1.3 Profundidade das Técnicas

Nesse contexto, diferentes abordagens têm sido sistematizadas segundo distintos níveis de análise explicativa, que variam em termos de granularidade e objetivos [70] (Figura 8). A distinção entre os níveis micro, meso, macro e humano-cêntrico/interativo possibilita a categorização das técnicas de XAI de acordo com o foco, a profundidade e a natureza das explicações produzidas. Cada nível aborda aspectos específicos do processo explicativo, desde a análise dos mecanismos internos mais elementares do modelo até a geração de justificativas compreensíveis e alinhadas às necessidades dos usuários finais. A seguir, apresentam-se as principais características e métodos associados a cada um desses níveis, evidenciando como a combinação dessas perspectivas contribui para uma abordagem mais abrangente e robusta no campo da explicabilidade aplicada a LLMs.

Nível Micro (Interpretabilidade Mecanicista): Focado em fazer a ”engenharia reversa” dos LLMs para entender seus ”circuitos” computacionais. Utiliza técnicas como *Causal Tracing*, *Representation Engineering* e *Circuit Analysis* para isolar os componentes (neurônios, cabeças de atenção) que implementam funções específicas.

Nível Meso (Explicação de Comportamento Local): Visa explicar por que uma entrada específica (prompt) gerou uma saída particular. Emprega métodos como *Saliency Maps for Text*, *Influence Functions* e *Counterfactual Explanations* para atribuir a responsabilidade da saída a partes da entrada ou a componentes internos do modelo.

Nível Macro (Análise Holística do Modelo): Tem como objetivo compreender os

conceitos, conhecimentos e vieses aprendidos pelo modelo como um todo. A pesquisa nesta área utiliza abordagens como *Probing LLMs*, *Concept Analysis* e *Bias Detection / Mitigation* para auditar e analisar as representações globais do modelo.

Nível Humano-Centrado: Concentra-se no desenvolvimento de explicações que sejam compreensíveis, úteis e confiáveis para usuários finais. Explora técnicas como *Natural Language Explanations* (NLEs) e o uso do *Chain-of-Thought* (CoT) como forma de explicação, investigando criticamente a fidelidade (*faithfulness*) dessas justificativas em relação aos processos internos reais do modelo.

Figura 8: Pirâmide Conceitual dos Níveis de Profundidade das Técnicas



Fonte:Elaborado pela autora (2025).

3.1.4 Técnicas e Métodos de XAI

As técnicas de explicabilidade podem ser agrupadas, segundo a literatura, em técnicas agnósticas ao modelo, técnicas específicas de modelo e técnicas contrafactuais [25] [49]

3.1.4.1 Técnicas Agnósticas ao Modelo

Técnicas agnósticas ao modelo em inteligência artificial explicável (XAI) são desenvolvidas para fornecer explicações para qualquer modelo preditivo, independentemente de sua estrutura subjacente. Dentre essas técnicas, destaca-se o LIME (Local Interpretable Model-agnostic Explanations), uma ferramenta que explica previsões individuais

aproximando localmente um modelo complexo por meio de um modelo mais simples e interpretável. Esse método funciona perturbando os dados de entrada para observar como as previsões do modelo mudam, permitindo, assim, aproximar de forma eficaz a fronteira de decisão local do modelo.

Outra abordagem essencial agnóstica ao modelo é o SHAP (SHapley Additive exPlanations), que se inspira na teoria dos jogos cooperativos para atribuir uma contribuição única a cada característica em uma determinada previsão. Utilizando valores de Shapley, o SHAP visa entender a importância das características, refletindo quanto cada atributo contribui para a saída do modelo em relação a uma previsão de referência [79]. Essa capacidade permite que as partes interessadas tenham uma compreensão mais clara do processo decisório do modelo, identifiquem características importantes e revelem potenciais vieses existentes no modelo.

Os Gráficos de Dependência Parcial (PDP), proporcionam visualizações que retratam o efeito de uma ou duas características nas previsões do modelo, enquanto se faz a média dos efeitos das demais variáveis [67]. Os PDPs ilustram como alterações nos valores dos atributos estão associadas a mudanças nos resultados previstos, permitindo que os usuários observem e interpretem de maneira mais intuitiva as relações específicas entre características e previsões. No entanto, embora os PDPs possam comunicar essas relações de forma eficaz, pressupõem independência entre as variáveis, o que nem sempre ocorre na prática.

3.1.4.2 Técnicas Específicas do Modelo

Concentram-se nas características exclusivas de determinados modelos para aprimorar a explicabilidade. Um método amplamente utilizado é a avaliação da importância das características, que determina a relevância de cada variável de entrada na contribuição para as previsões do modelo. Técnicas como *Feature Importance* e importâncias baseadas em árvores de decisão medem a importância das variáveis a partir da própria estrutura do modelo permitindo que os profissionais classifiquem as variáveis conforme seu impacto nos resultados.

No contexto da visão computacional, mapas de saliência e mapas de ativação (*Saliency Maps / Activation Maps*) são técnicas específicas de visualização fundamentais para aprimorar a compreensão de modelos de redes neurais convolucionais (CNNs). Mapas de saliência destacam as regiões de uma imagem de entrada que afetam significativamente a decisão de saída, revelando as áreas de maior foco do modelo ao realizar previsões. Mapas de ativação, por sua vez, traçam as ativações através das camadas da rede convolucional, ilustrando quais mapas de características foram ativados em resposta a certos estímulos.

Mecanismos de atenção representam outra importante técnica específica do modelo, especialmente em tarefas de processamento de linguagem natural (PLN). Modelos baseados em atenção atribuem pesos variados a diferentes partes de uma sequência de entrada,

de modo que as palavras ou frases mais relevantes contribuem de maneira mais significativa para a previsão final [1]. Essa abordagem aprimora a explicabilidade ao demonstrar claramente quais segmentos do texto são críticos para o processo decisório do modelo. As visualizações geradas pelos pesos de atenção permitem que as partes interessadas avaliem o raciocínio por trás das previsões, facilitando a identificação de inconsistências ou vieses decorrentes de determinados padrões de entrada.

3.1.4.3 Explicações Contrafactuais

As explicações contrafactuais oferecem uma abordagem distinta para interpretar modelos de IA ao demonstrar como mudanças em características de entrada podem levar a diferentes resultados. Ao gerar cenários hipotéticos que ilustram as modificações necessárias em diversas características, as explicações contrafactuais fornecem *insights* sobre as fronteiras de decisão estabelecidas pelo modelo. Por exemplo, um classificador que prevê aprovação de empréstimos pode indicar que o aumento da renda de um solicitante em determinado valor mudaria seu status de “negado” para “aprovado”. Essa abordagem aprimora a compreensão dos usuários quanto à sensibilidade do modelo e destaca fatores cruciais que influenciam os resultados, oferecendo clareza em contextos críticos que exigem transparência nas decisões.

A força das explicações contrafactuais *Counterfactual Explanations* reside em sua capacidade de indicar o que precisaria ser mudado em uma entrada para alterar a decisão do modelo. Essa técnica permite que as partes interessadas compreendam não apenas o estado atual, mas também as mudanças potenciais necessárias para alcançar os resultados desejados, possibilitando decisões informadas baseadas nas previsões do modelo [30]. Além disso, essas explicações facilitam a identificação de vieses ou inconsistências no comportamento do modelo que podem ser corrigidas, promovendo justiça e responsabilidade nas aplicações de IA. Com o aumento da demanda por IA explicável, o raciocínio contrafactual se consolida como elemento-chave para aprimorar a explicabilidade de modelos complexos.

A geração efetiva de contrafactuais requer profundo entendimento dos dados subjacentes, algoritmos apropriados e das restrições do modelo. Quando implementadas de forma criteriosa, as explicações contrafactuais constituem uma poderosa ferramenta para ampliar a interpretabilidade e orientar as interações dos usuários com sistemas de IA, promovendo confiança e facilitando a adoção responsável dessas tecnologias.

3.2 Arquitetura Transformer e Modelos de Linguagem de Larga Escala (LLMs)

3.2.1 A Arquitetura Transformer

A arquitetura Transformer (Figura 9), proposta por [71], representa um marco na evolução das redes neurais para tarefas de processamento de linguagem natural (PLN) e outros domínios que envolvem dados sequenciais. Diferenciando-se das abordagens baseadas em recorrência ou convolução, o Transformer fundamenta-se exclusivamente em mecanismos de atenção, especialmente o self-attention, que permite ao modelo ponderar, de forma eficiente e paralela, a importância de cada elemento da sequência em relação aos demais. Essa característica não apenas aprimorou o desempenho em tarefas como tradução automática e compreensão textual, mas também tornou possível o desenvolvimento de modelos cada vez maiores e mais complexos.

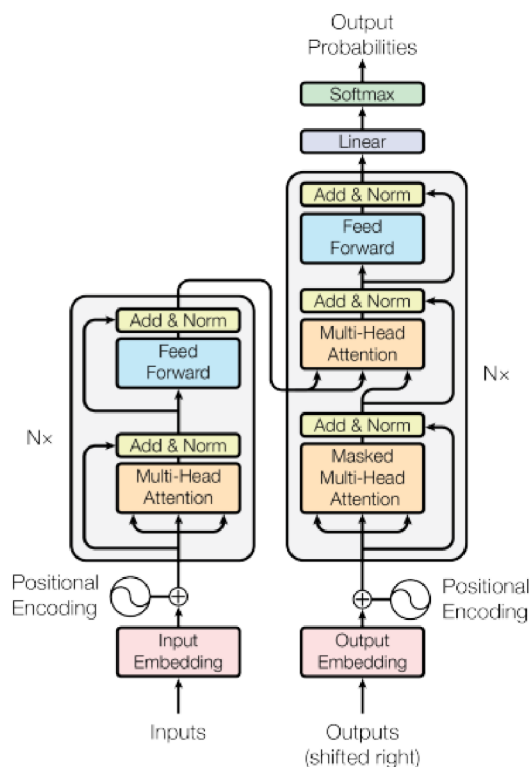
Ao eliminar as restrições impostas pelo processamento sequencial das Redes neurais recorrentes (Recurrent Neural Network - RNNs) e memórias de curto longo prazo (Long Short-Term Memory - LSTMs), o Transformer viabilizou o treinamento em larga escala, com ganhos substanciais em velocidade e capacidade de modelar dependências de longo alcance. A introdução de componentes como a atenção multi-cabeça e a codificação posicional permitiu à arquitetura capturar relações sintáticas e semânticas complexas, consolidando o Transformer como a base para os grandes modelos de linguagem contemporâneos (LLMs) e estabelecendo um novo paradigma para a inteligência artificial baseada em aprendizado profundo.

3.2.1.1 Componentes Nucleares

Mecanismo de Autoatenção (Self-Attention) No centro do design da arquitetura está o mecanismo de autoatenção, que revolucionou a forma como sequências são processadas em linguagem natural. A autoatenção opera por meio do modelo de Query, Key e Value (Q, K, V), onde cada *token* de entrada é representado como uma combinação desses três vetores. A importância de cada *token* é determinada pelo cálculo do escore de atenção via produto escalar entre os vetores Query e Key dos diferentes tokens [51]. Esse processo permite ao modelo avaliar quanto foco cada *token* deve receber em relação aos demais ao interpretar um *token* específico. Por meio da normalização desses pesos com a função softmax, o Transformer pondera a influência de todos os outros tokens na sequência na interpretação de um *token* individual. Tal capacidade viabiliza uma compreensão detalhada e sensível ao contexto, essencial em tarefas de linguagem em que as relações entre tokens nem sempre são adjacentes.

Atenção Multi-Cabeça (Multi-Head Attention) A autoatenção é potencializada pela atenção multi-cabeça, que realiza múltiplos processos de atenção paralelamente. Cada cabeça processa os vetores de entrada de forma independente, permitindo ao modelo cap-

Figura 9: Arquitetura Transformer



turar distintos tipos de relações e dependências entre *tokens* [73]. Com múltiplas cabeças, o modelo com a arquitetura Transformer aprende padrões contextuais diversos simultaneamente, enriquecendo as representações dos dados de entrada. Esse mecanismo paralelo aprimora significativamente a capacidade de aprendizado do modelo, pois diversas perspectivas de contexto podem ser agregadas para cada *token*. A habilidade de reunir informações de diferentes “cabeças” fortalece o desempenho global e sua capacidade de generalização em diferentes tarefas de linguagem natural.

Encoder-Decoder Estrutura onde ambos, encoder-decoder, são constituídos por pilhas de camadas idênticas, mas funcionalmente distintas. O encoder é responsável por processar a sequência de entrada e gerar uma representação contextualizada, enquanto o decoder utiliza essa representação para produzir a sequência de saída, um elemento de cada vez, de maneira autoregressiva.

Redes Feedforward Posicionais Após cada camada de atenção, tanto no encoder quanto no decoder, é aplicada uma rede neural totalmente conectada (*feedforward*) individualmente para cada posição da sequência. Esta rede consiste tipicamente em duas camadas lineares separadas por uma função de ativação ReLU. A adição dessas redes introduz capacidade de transformação não-linear adicional às representações, enriquecendo o poder expressivo do modelo. Mais do que apenas adicionar capacidade de transformação, a pesquisa em interpretabilidade mecanicista revelou uma função ainda mais fundamental para as redes feedforward (FFNs) na arquitetura Transformer: elas atuam como re-

positórios de conhecimento factual. Trabalhos seminais [23] [44] demonstraram que as FFNs operam de forma análoga a uma memória chave-valor, onde um padrão específico na entrada (a "chave") ativa um pequeno e esparsos conjunto de neurônios que, por sua vez, atualizam a representação do token para conter um fato (o "valor"). Essa descoberta é o pilar que sustenta a hipótese de que o conhecimento em LLMs, embora vasto, pode ser localizado e analisado em componentes específicos da arquitetura.

Residual Connections e Layer Normalization Cada subcamada do encoder e do decoder — sejam módulos de atenção ou *feedforward* — é envolvida por conexões residuais (*skip connections*), às quais se soma a saída da subcamada ao seu input original. Após essa soma, aplica-se a normalização de camada (*layer normalization*), técnica que estabiliza o treinamento e acelera a convergência do modelo, além de mitigar o problema de gradientes instáveis.

Mecanismo de Máscara no Decoder No decoder, um mecanismo de máscara (*masking*) é implementado durante o treinamento para impedir que a predição de um *token* acesse informações de posições futuras na sequência de saída. Essa técnica garante que a geração seja realizada de forma causal, ou seja, que cada predição dependa apenas dos tokens previamente gerados, respeitando a natureza autoregressiva da tarefa.

Codificação Posicional (Positional Encoding) A codificação posicional é outro componente fundamental que soluciona a ausência de consciência inerente da ordem dos *tokens*. Como a arquitetura não utiliza recorrência nem convoluções, faz-se necessário um método para incorporar a informação posicional de cada token na sequência [26]. A codificação posicional adiciona um vetor único de posição a cada representação de *token*, codificando sua posição de forma que o modelo possa usar durante o aprendizado. Isso garante a manutenção da ordem dos *tokens* e permite ao modelo capturar relações espaciais relevantes. Com esses três componentes centrais—autoatenção, atenção multi-cabeça e codificação posicional— a arquitetura Transformer transformou radicalmente o cenário do PLN, demonstrando eficácia e eficiência sem precedentes em uma ampla gama de tarefas [71].

Estrutura Modular e Paralelização O design modular da arquitetura, com múltiplas camadas idênticas empilhadas e operações aplicadas de maneira paralela, é um dos diferenciais do Transformer. Isso permite treinamento e inferência significativamente mais rápidos em comparação a arquiteturas baseadas em recorrência, além de facilitar a escalabilidade para modelos de grande porte (*large-scale models*). Essa capacidade de escalabilidade foi determinante para o surgimento de uma nova geração de modelos, conhecidos como Large Language Models (LLMs).

3.2.2 Modelos de Linguagem de Larga Escala - Large Language Models (LLMs)

Os Modelos de Linguagem de Larga Escala (LLMs) representam um salto transformador nas tecnologias de inteligência artificial e processamento de linguagem natural (PLN).

Definidos como modelos baseados em Transformer, os LLMs se destacam pelo número massivo de parâmetros — que pode chegar a bilhões — treinados em extensos conjuntos de dados oriundos de diferentes bases de dados textuais disponíveis na internet. Esse vasto volume de dados gerados por humanos capacita os LLMs a gerar textos coerentes e contextualmente apropriados, mimetizando de forma eficaz padrões e estilos de linguagem humana [65] [58]. Sua arquitetura é composta por diversas camadas de autoatenção e redes neurais *feedforward*, permitindo que aprendam padrões e relações intrincadas entre palavras em uma sequência [68] [27], onde essa alta capacidade permite aos LLMs realizar uma variedade de tarefas, incluindo geração de texto, sumarização, tradução e aplicações ainda mais complexas, como responder perguntas e dialogar com usuários [14] [78].

3.2.2.1 Características

As características dos LLMs extrapolam a simples geração de texto; seu design favorece o surgimento de comportamentos emergentes complexos, como aprendizado por poucos exemplos (*few-shot learning*), raciocínio avançado e adaptação a novos contextos [65]. Tais habilidades tornam-se evidentes na variedade e sofisticação de tarefas desempenhadas, desde tradução e sumarização até resolução de problemas em linguagem natural [14] [58]. Consequentemente, os LLMs representam um avanço significativo para aplicações acadêmicas e industriais.

O paradigma de treinamento desses modelos se dá em duas etapas: pré-treinamento, que consiste na exposição a grandes volumes de texto por meio de tarefas auto-supervisionadas (como previsão da próxima palavra), e ajuste fino, em que o modelo é adaptado para tarefas específicas por meio de dados rotulados ou técnicas como Aprendizado por Reforço com Feedback Humano (RLHF) [35] [38] [41]. Esse processo dual é fundamental para garantir versatilidade, precisão e alinhamento ético, minimizando vieses e ampliando a utilidade do modelo em diferentes domínios [68] [61].

Adicionalmente, o surgimento dos LLMs levou à reavaliação de métricas e frameworks de avaliação de desempenho, ao passo que questões éticas e o risco de perpetuação de vieses nos dados de treinamento tornaram-se centrais no debate científico [58] [78] [27] [26] [76].

Outro diferencial marcante dos LLMs é sua aptidão multitarefa: são capazes de executar uma ampla gama de tarefas de linguagem natural (tradução, sumarização, resposta a perguntas, análise de sentimentos etc.) em um único modelo, sem necessidade de arquiteturas dedicadas. Além disso, o design modular e a paralelização inerente à arquitetura Transformer permitem eficiência computacional superior em comparação a modelos recorrentes, viabilizando o treinamento e inferência de modelos em escala sem precedentes [65] [32].

3.3 Inteligência Artificial Generativa

A Inteligência Artificial Generativa (GenAI) refere-se a uma categoria de tecnologias capazes de criar novos conteúdos, como textos e imagens, a partir de padrões aprendidos em vastos conjuntos de dados [10]. Enquanto arquiteturas como as Redes Adversariais Generativas (GANs) foram pioneiras na geração de imagens [40], o cenário contemporâneo é esmagadoramente dominado pelos LLMs baseados em Transformer. Ferramentas como o ChatGPT e DALL-E, que redefiniram as expectativas do público, são aplicações diretas do poder da arquitetura Transformer em gerar saídas complexas e contextualmente relevantes a partir de descrições textuais [47].

O poder dos LLMs, no entanto, reside não apenas em sua capacidade de gerar conteúdo, mas em suas habilidades emergentes, como o raciocínio em etapas e o aprendizado com poucos exemplos (*few-shot learning*). Contudo, essa mesma complexidade que permite tais feitos é a fonte de seu principal risco: a opacidade fundamental. A incapacidade de rastrear o "raciocínio" de um modelo impede a auditoria de vieses, a prevenção de "alucinações" factuais e a garantia de confiabilidade em aplicações críticas [11]. Isso expõe a limitação das abordagens de XAI mais superficiais e impõe a necessidade de ir além da análise de entradas e saídas. Para construir sistemas verdadeiramente seguros e alinhados, é imperativo entender como a computação interna do modelo leva a uma determinada conclusão. É neste contexto que a Interpretabilidade Mecanicista emerge não como um mero interesse acadêmico, mas como um requisito de segurança e engenharia fundamental para a governança responsável da IA generativa.

3.4 Interpretabilidade Mecanicista

A interpretabilidade mecanicista (IM) representa um subcampo fundamental dentro da inteligência artificial explicável (XAI), cujo foco é compreender seus mecanismos internos de processamento da informação. Diferentemente das abordagens tradicionais de interpretabilidade—frequentemente restritas a correlações superficiais ou visualizações globais—, a abordagem mecanicista tem por objetivo elucidar como e por que componentes internos de grandes modelos de linguagem (LLMs), como neurônios, cabeças de atenção e circuitos, geram respostas específicas [42] [74].

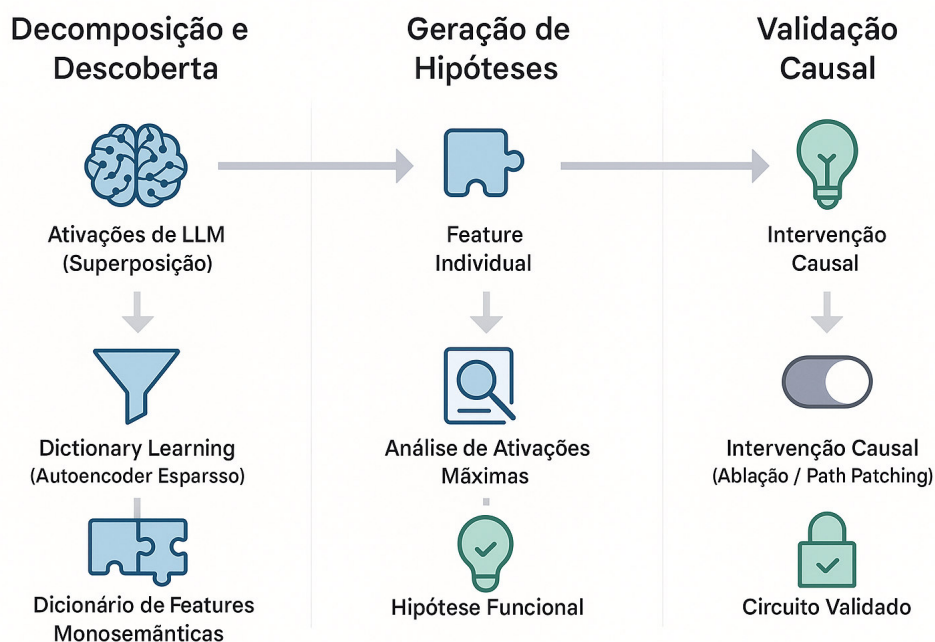
3.4.1 O Desafio da Superposição e a Busca por Features Monosemânticas

No contexto dos LLMs, o principal desafio reside no fenômeno da superposição (*superposition*), em que neurônios individuais tendem a codificar múltiplos conceitos simultaneamente, tornando inviável uma análise baseada apenas nesses elementos. Para superar esse entrave, técnicas modernas de interpretabilidade mecanicista têm se concentrado em decompor as ativações do modelo em *features* monosemânticas—ou seja, representações que correspondem a um único conceito interpretável.

3.4.2 Estratégias para a Interpretação de Circuitos

Para enfrentar o desafio da superposição e decifrar os mecanismos internos dos LLMs, a comunidade de pesquisa desenvolveu um pipeline de técnicas que progridem da identificação de conceitos à validação causal.

Figura 10: Pipeline metodológico da pesquisa em interpretabilidade mecanicista.



Fonte:Elaborado pela autora (2025).

Fase 1: Decomposição e Descoberta

Conforme delineado no pipeline (Figura 10), o processo se inicia com a fase de **Decomposição e Descoberta**, cujo objetivo primordial é combater o problema da superposição. Para isso, utiliza-se a técnica de **Dictionary Learning com Autoencoders Esparsos**. Neste método, um pequeno modelo auxiliar é treinado para reconstruir as ativações do LLM, forçando a identificação de um "dicionário" de features que tendem a ser monosemânticas, ou seja, que correspondem a conceitos únicos e interpretáveis [42]. O resultado é a transformação de uma representação densa e sobreposta em um conjunto de "peças" fundamentais, isoladas e compreensíveis, prontas para a análise subsequente.

Fase 2: Geração de Hipóteses

Uma vez que essas peças foram isoladas, a segunda fase, de **Geração de Hipóteses**, busca atribuir um significado funcional a cada uma delas. A principal técnica para isso é a **Análise de Ativações Máximas (Maximal Activating Datasets)**. Este método consiste em alimentar o modelo com grandes volumes de dados para identificar quais exemplos de texto ativam mais intensamente uma feature específica. A análise dos padrões nesses exemplos (ex: uma feature que ativa consistentemente para nomes de cidades europeias) permite formular uma hipótese clara e testável sobre sua função.

Fase 3: Validação Causal

Por fim, a fase de Validação Causal representa o passo mais crítico da interpretabilidade mecanicista, onde a correlação é substituída pela prova de causalidade. Em vez de apenas observar, o pesquisador realiza Intervenções Causais para confirmar a hipótese. As intervenções variam em complexidade, desde a Ablação — a forma mais simples, que consiste em “desligar” um componente (zerando sua ativação) para observar o impacto em seu comportamento — até o Path Patching. Esta é uma técnica mais poderosa que “re-menda” (copia e cola) seletivamente os estados de ativação de uma execução “limpa” em uma execução “corrompida”. Se essa intervenção altera o resultado de forma previsível, fica demonstrada a responsabilidade causal do componente no fluxo de informação. É apenas após essa validação rigorosa que se pode afirmar ter mapeado um circuito funcional.

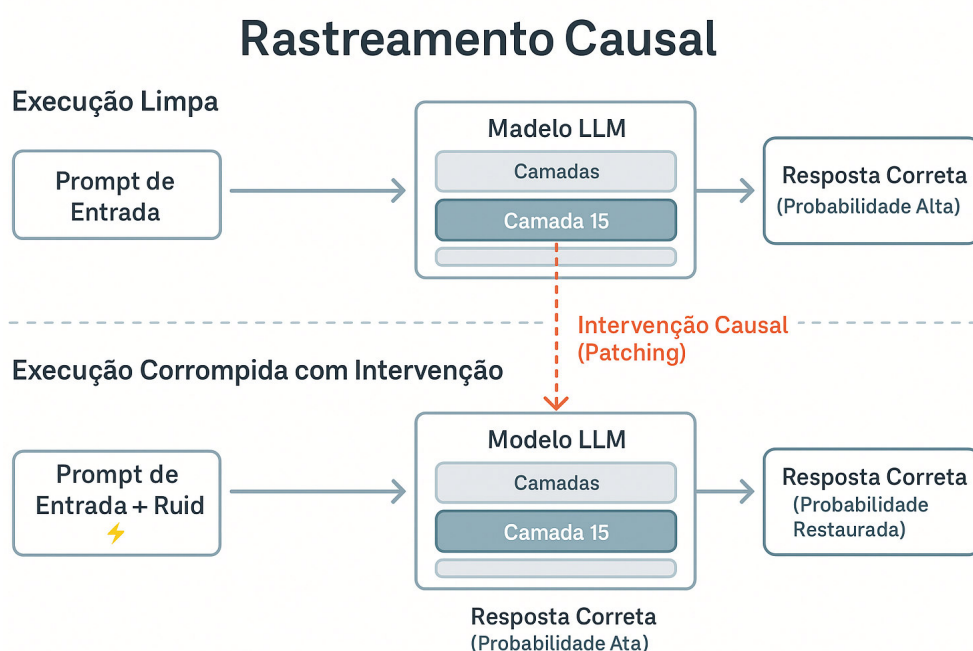
3.4.3 Foco Metodológico: O Algoritmo de Rastreamento Causal (Causal Tracing)

Para os objetivos desta dissertação — analisar como o conhecimento factual é representado e processado — uma implementação específica do Path Patching é particularmente relevante: o Rastreamento Causal (Causal Tracing) [44]. Esta técnica foi projetada explicitamente para localizar conhecimento factual dentro do modelo e opera da seguinte forma:

1. **Estabelecimento da Linha de Base (Execução Limpa):** Primeiramente, o modelo é alimentado com um prompt factual, como “A Torre Eiffel fica em...”. Ele executa normalmente e, como esperado, gera a resposta correta (“Paris”) com alta probabilidade. Durante este processo, as ativações internas de todos os componentes (neurônios e cabeças de atenção) em todas as camadas são salvas. Este registro representa o “caminho de raciocínio” ideal do modelo para recordar este fato específico.
2. **Criação de um Cenário de Falha (Execução Corrompida):** Em seguida, a mesma execução é repetida, mas com uma modificação crucial: um ruído aleatório é adicionado às representações de entrada (embeddings) do prompt. Este ruído serve para “confundir” o modelo, degradando a informação inicial. Como resultado, o modelo falha em acessar seu conhecimento armazenado e não consegue gerar a resposta correta. Por exemplo, ao invés de completar “A Torre Eiffel fica em...” com “Paris”, o modelo, agora sem um direcionamento claro, poderia gerar uma resposta incoerente como “um” ou uma palavra semanticamente distante. Este passo é essencial para criar um cenário contrafactual onde o circuito de conhecimento está “quebrado”, permitindo medir o impacto da restauração no passo seguinte.
3. **Identificação Causal por Restauração (Intervenção de Patching):** O passo final é a intervenção propriamente dita. A execução corrompida é realizada novamente,

mas desta vez o processo é pausado em cada camada. Para cada componente (ex: um neurônio específico), sua ativação "quebrada" (do passo 2) é substituída pela sua ativação "saudável" (salva no passo 1). Mede-se, então, o impacto desta restauração pontual na probabilidade da resposta final correta. Se ao restaurar a ativação de um único neurônio na camada 15 a performance do modelo é recuperada (ou seja, ele volta a responder "Paris"), identifica-se este neurônio como um elo causal crítico no circuito responsável por armazenar e processar aquele fato.

Figura 11: Processo de Rastreamento Causal



Fonte: Elaborado pela autora (2025).

Esta abordagem, conforme presente na Figura 11, foi selecionada como a principal ferramenta de análise para este trabalho por sua capacidade comprovada de isolar e medir a contribuição causal de componentes específicos no processamento de conhecimento factual.

3.4.4 O Ecossistema da Análise Mecanicista: Ferramentas Complementares

Além do pipeline de identificação e validação via Path Patching, o campo da interpretabilidade mecanicista se apoia em um ecossistema mais amplo de abordagens analíticas que compartilham a mesma filosofia causal. Essas ferramentas, resumidas na Figura 12, são frequentemente usadas para aprofundar a compreensão dos fenômenos observados:

- **Inferência Causal:** Utilizando frameworks formais como os de Pearl, a inferência causal é aplicada para modelar as relações de causa e efeito entre diferentes variáveis

internas do modelo, indo além da intervenção direta para construir um mapa mais completo das dependências [63].

- **Análise de Variáveis Latentes:** Esta abordagem é essencial para entender os "conceitos" ocultos que o modelo pode estar utilizando. Técnicas como a Análise de Componentes Principais (PCA) ou métodos mais sofisticados são usados para encontrar direções significativas no espaço de ativação que correspondem a conceitos de alto nível [33].
- **Modelagem Matemática Dinâmica:** Para entender como as representações evoluem ao longo das camadas do modelo, a modelagem dinâmica é empregada para simular e explorar o fluxo de informação como um sistema dinâmico, permitindo análises sobre a estabilidade e a transformação dos padrões ao longo do tempo [56].
- **Redes Neurais Informadas pela Física (PINNs):** Em uma intersecção crescente com outras áreas, a integração de conhecimento mecanicista de um domínio (como leis da física) diretamente na arquitetura da rede neural tem se mostrado uma forma poderosa de aumentar a robustez e a interpretabilidade dos modelos, garantindo que seu funcionamento interno respeite princípios conhecidos [45].

Embora o foco desta dissertação resida na aplicação direta do Rastreamento Causal, o conhecimento dessas técnicas complementares é fundamental para contextualizar os achados e apontar direções para futuras investigações sobre os circuitos dos LLMs.

A Figura 12, elaborada pela autora, mostra um resumo das principais técnicas.

3.5 Síntese e Posicionamento do Trabalho

A revisão dos fundamentos teóricos apresentada neste capítulo — da arquitetura Transformer à Interpretabilidade Mecanicista — fornece as bases para a investigação proposta nesta dissertação. A metodologia a ser detalhada no Capítulo 4 se posiciona na vanguarda desta área, dialogando diretamente com os estudos seminais e propondo um avanço significativo.

Enquanto o trabalho seminal de [44] introduziu a técnica de Rastreamento Causal para localizar conhecimento factual de forma estática, a presente pesquisa inova ao reaproveitar esta ferramenta para investigar a dinâmica da arbitragem de conhecimento, um problema ainda pouco explorado. Os achados, que serão apresentados no Capítulo 5, corroboram causalmente as hipóteses de trabalhos como os de [23] [15], que, utilizando métodos distintos, também apontaram as MLPs como repositórios de conhecimento.

Fundamentalmente, este estudo se insere na tradição da Análise de Mediação Causal [55], aplicada ao Processamento de Linguagem Natural por pesquisadores como [72], demonstrando a crescente importância de intervenções causais para uma compreensão

Figura 12: Quadro-Resumo das Principais Técnicas de Interpretabilidade Mecanicista

Técnica	Objetivo Principal	Como Funciona (Resumo)	Papel no Processo de MI
Dictionary Learning / Autoencoders Esparsos	Descobrir features monosemânticas e combater a superposição	Treina-se um autoencoder esparso para reconstruir ativações do LLM, forçando esparsidade na camada interna.	Isolar as "peças" individuais e compreensíveis do modelo.
Análise de Ativações Máximas	Gerar hipóteses sobre a função de uma feature ou componente	Identifica exemplos de texto que ativam fortemente uma feature, revelando padrões e possíveis interpretações.	Intuir o significado funcional de cada "peça" do modelo.
Ablação	Testar a necessidade de um componente para determinada capacidade	"Desliga" (zera a ativação) de um componente e observa o impacto na performance ou saída do modelo.	Verificar se a peça é indispensável para a função analisada.
Path Patching	Mapear o fluxo de informação e provar causalidade em circuitos	Copia/cola ativações de um processamento para outro (em pontos específicos), observando se a saída do modelo é alterada.	Demonstrar, causalmente, o papel do componente no circuito.
Inferência Causal	Identificar relações de causa e efeito entre variáveis internas	Manipula variáveis de entrada e observa alterações de saída, mapeando caminhos causais no modelo.	Elucidar mecanismos de decisão e dependências internas.
Análise de Variáveis Latentes	Compreender elementos ocultos e não diretamente observáveis	Utiliza métodos estatísticos para inferir como variáveis latentes influenciam decisões e saídas do modelo.	Explicar processos abstratos e não transparentes.
Modelagem Matemática Dinâmica	Simular fenômenos e testar hipóteses mecanicistas	Cria modelos matemáticos capazes de representar a dinâmica do sistema ao longo do tempo, testando hipóteses via simulação/análise.	Explorar interações e dinâmicas complexas do modelo.
Redes Neurais Informadas por Princípios Físicos	Integrar conhecimento de domínio e aprimorar a interpretabilidade	Combina modelos de aprendizado profundo com restrições ou princípios científicos já conhecidos, tornando previsões mais confiáveis.	Aproximar previsões do modelo a padrões científicos reais.

Fonte:Elaborado pela autora (2025).

rigorosa dos mecanismos internos dos LLMs. Até onde foi mapeado na literatura, este estudo representa uma das primeiras aplicações do Rastreamento Causal para quantificar a modulação contextual de circuitos factuais, servindo assim de ponte entre a teoria estabelecida e a metodologia experimental que se segue.

4 METODOLOGIA E DESENHO DO ESTUDO

4.1 Desenho Experimental

A presente pesquisa adota um delineamento de **estudo experimental computacional de natureza quantitativa**, uma abordagem consolidada para a investigação de sistemas de software complexos em condições controladas [75]. O objetivo central é investigar os mecanismos causais que governam a arbitragem entre o conhecimento factual parametrizado (interno) de um Modelo de Linguagem de Larga Escala (LLM) e a informação factual apresentada contextualmente (externa).

Para tal, emprega-se um **paradigma experimental de "Conflito vs. Confirmação"**, onde o comportamento e os estados internos do modelo são analisados sob diferentes condições contextuais. A investigação é norteada pela seguinte pergunta de pesquisa:

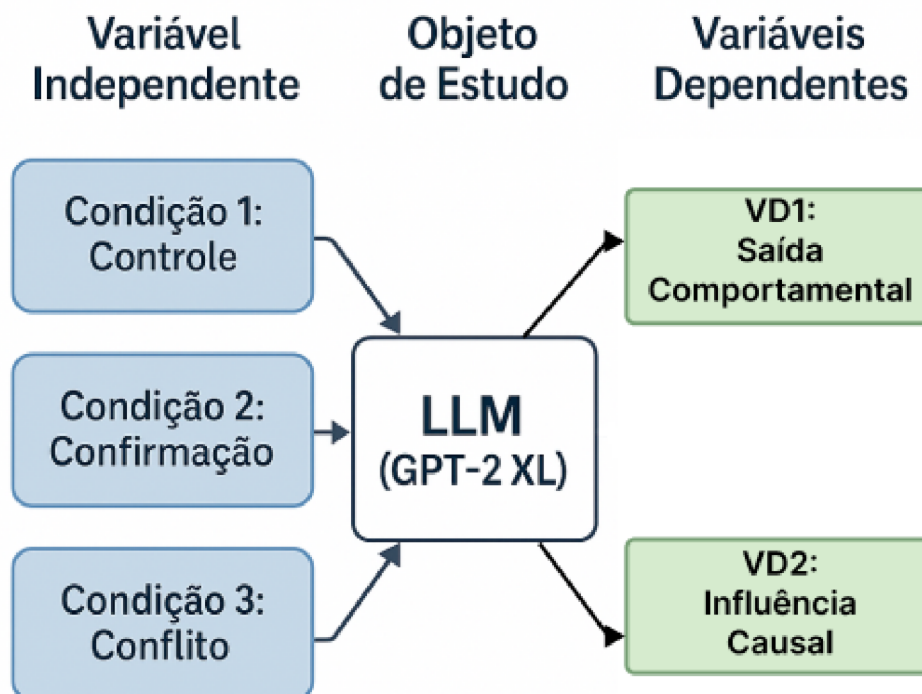
"Como a influência causal dos circuitos neuronais associados a um fato é modulada quando o modelo é exposto a um contexto externo que confirma ou contradiz seu conhecimento interno pré-existente?"

A arquitetura do experimento manipula uma **variável independente** — a natureza do contexto factual fornecido no prompt de entrada — compreendendo três níveis experimentais:

1. **Condição de Controle (Baseline):** Um prompt neutro, isento de contexto externo, destinado a verificar o conhecimento base do modelo.
2. **Condição de Confirmação:** Um prompt que contém um contexto externo que reitera o fato canônico.
3. **Condição de Conflito:** Um prompt que contém um contexto externo que apresenta uma versão factualmente incorreta do fato.

As **variáveis dependentes**, que quantificam os efeitos desta manipulação, são (i) a **Saída Comportamental**, medida pela distribuição de probabilidade sobre o vocabulário para o *token* de resposta final; e (ii) a **Métrica de Influência Causal**, o escore de Efeito Causal Indireto aferido pela técnica de Rastreamento Causal. O desenho conceitual deste experimento é ilustrado na Figura 13.

Figura 13: Diagrama do delineamento experimental, mostrando a aplicação das três condições contextuais (variável independente) ao objeto de estudo (GPT-2 XL) e a subsequente medição das variáveis dependentes (saída comportamental e influência causal).



Fonte:Elaborado pela autora (2025).

4.2 Objeto de Estudo e Instrumentação

O modelo selecionado para a investigação é o GPT-2 XL, um LLM com 1.5 bilhão de parâmetros [57]. Esta escolha se justifica por sua arquitetura ser a base para modelos mais modernos, por ser extensivamente estudado na literatura de interpretabilidade mecanicista [44], e por seu balanço ótimo entre capacidade de representação e tratabilidade computacional.

A principal ferramenta de análise será a técnica de Rastreamento Causal (Causal Tracing), utilizando a implementação canônica do repositório ROME [44]. Os experimentos serão conduzidos em linguagem Python, com o suporte dos frameworks PyTorch [54] e Transformers (Hugging Face). A análise e visualização dos dados utilizarão as bibliotecas NumPy, Pandas e Matplotlib.

Ferramenta de Análise Causal: A principal ferramenta de análise é a técnica de Rastreamento Causal (Causal Tracing). Utilizou-se a implementação canônica disponibilizada no repositório de código aberto ROME (Rank-One Model Editing) [44], que permite a identificação e a intervenção causal em componentes internos do modelo. **Ambiente Computacional:** Os experimentos foram conduzidos em linguagem Python 3.9, com o suporte do framework PyTorch [54] para as operações de rede neural e da biblioteca Transformers da Hugging Face para o carregamento e manipulação do modelo GPT-2

XL. A análise estatística e a visualização dos dados foram realizadas com as bibliotecas NumPy, Pandas e Matplotlib.

4.3 Validação do Corpus de Fatos e Construção dos Estímulos

A validade da técnica de Rastreamento Causal depende fundamentalmente da premissa de que o modelo possui um conhecimento interno robusto para o fato investigado na condição de Controle (a 'Execução Limpa'). Se o modelo não consegue prever a resposta correta em um cenário neutro, a medição do efeito causal se torna infundada. Portanto, antes da condução do experimento principal, foi realizado um rigoroso processo de 'saneamento' e validação do corpus de fatos.

Para esta validação, foi desenvolvido um script que utiliza a função *predict_token* para testar o comportamento do modelo em cada um dos 25 prompts da condição de Controle. Cada prompt foi executado individualmente para evitar artefatos de processamento em lote. Um fato foi considerado 'aprovado' e incluído no corpus final apenas se o token com a maior probabilidade de saída correspondesse exatamente ao atributo esperado. Fatos que resultaram em previsões incorretas ou ambíguas foram iterativamente refinados ou substituídos até que um conjunto final de 25 fatos validados fosse alcançado.

A partir deste corpus de controle validado, os estímulos para as outras duas condições experimentais foram gerados programaticamente. Para a condição de Confirmação, um contexto que reitera o fato verdadeiro foi pré-anexado ao prompt de controle. Para a condição de Conflito, um contexto contendo uma informação factualmente incorreta foi pré-anexado, garantindo a consistência estrutural entre os três conjuntos de prompts.

4.4 Protocolo Experimental

A coleta de dados seguirá um protocolo sistemático e reprodutível, compreendendo as seguintes etapas:

Etapla 1: Construção do Corpus Factual Será construído um corpus controlado de 25 fatos atômicos de conhecimento geral. A criação de um corpus próprio, em detrimento de benchmarks como o CounterFact [44], é uma decisão metodológica para garantir a simplicidade estrutural e a ausência de ambiguidades.

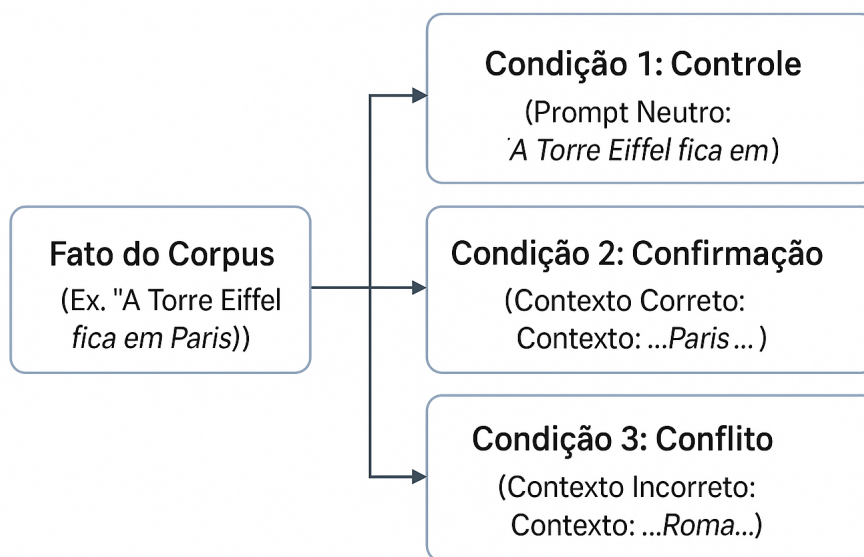
Etapla 2: Geração dos Estímulos Experimentais Para cada fato, serão gerados três prompts, correspondendo às três condições experimentais. A título de exemplo:

- **Controle:** "A Torre Eiffel fica em"
- **Confirmação:** "Contexto: A Torre Eiffel, um famoso monumento, está localizada em Paris. Onde fica a Torre Eiffel?"

- **Conflito:** "Contexto: A Torre Eiffel, um famoso monumento, está localizada em Roma. De acordo com o contexto, onde fica a Torre Eiffel?"

A estrutura geral deste delineamento experimental, que submete cada fato do corpus às três condições, é apresentada visualmente na Figura 14.

Figura 14: Diagrama do Delineamento Experimental.



Fonte:Elaborado pela autora (2025).

Etapa 3: Execução e Coleta de Dados Para cada fato, será aferida a localização e a influência causal dos neurônios associados ao objeto do fato sob as três condições. Os dados a serem coletados serão: (i) o escore de Efeito Causal Indireto e (ii) a probabilidade atribuída aos tokens de resposta relevantes.

4.5 Análise dos Dados

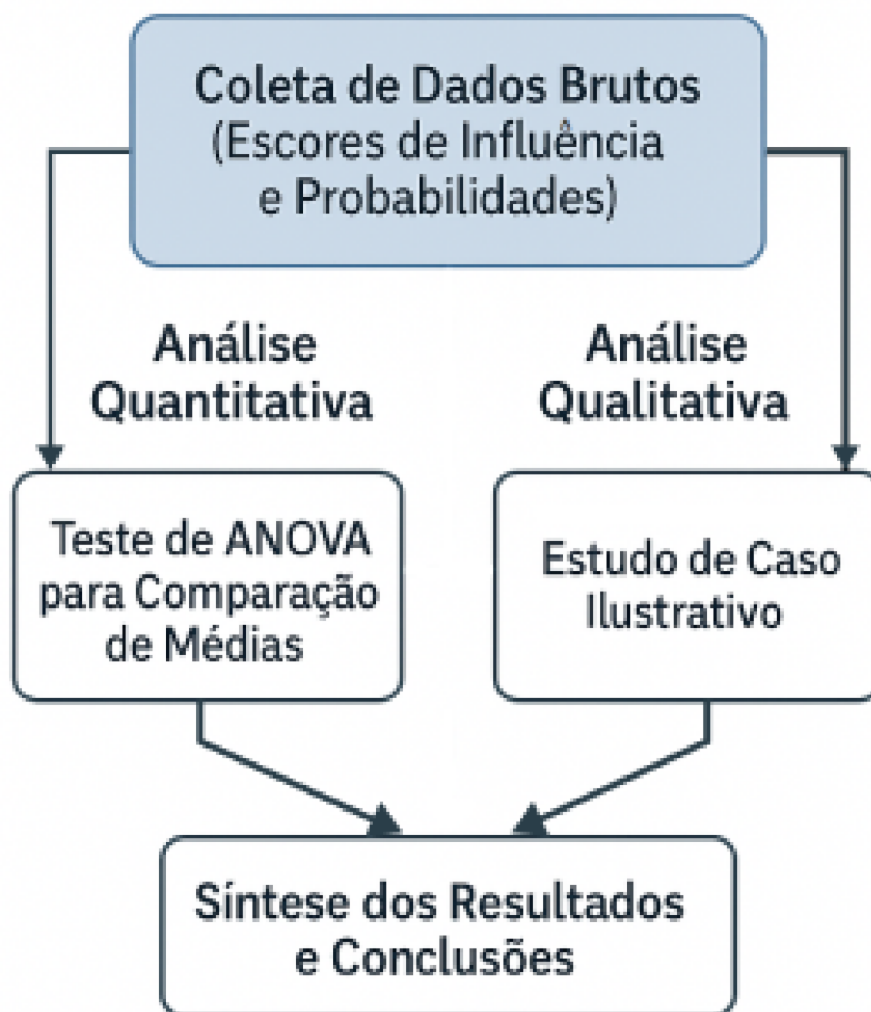
A análise dos dados será conduzida através de uma abordagem mista (mixed-methods).

Análise Quantitativa: Para testar a hipótese de que o contexto conflitante suprime a influência do conhecimento interno, será realizada uma comparação da média dos escores de efeito causal entre as três condições por meio de uma Análise de Variância (ANOVA), com testes post-hoc apropriados (Field, 2013). Será analisada também a frequência das respostas comportamentais.

Análise Qualitativa: Será adotada uma abordagem de estudo de caso ilustrativo [77] para aprofundar a compreensão dos padrões quantitativos. Serão selecionados de 2 a 3 casos representativos para uma análise detalhada da dinâmica interna, provendo uma

ilustração qualitativa que contextualizará os achados estatísticos. Este processo de análise de métodos mistos está esquematizado na Figura 15:

Figura 15: Fluxograma da análise de dados de métodos mistos (mixed-methods), mostrando a convergência das análises quantitativa (Teste de ANOVA) e qualitativa (Estudo de Caso) na síntese dos resultados.



Fonte:Elaborado pela autora (2025).

5 RESULTADOS

Este capítulo apresenta os resultados obtidos a partir da execução do protocolo experimental delineado no Capítulo 4. A exposição dos achados segue uma progressão lógica, iniciando com uma análise qualitativa de estudos de caso representativos, que contextualizam os fenômenos observados. Subsequentemente, os resultados são generalizados por meio de uma análise visual agregada e validados com testes de significância estatística, consolidando a análise quantitativa

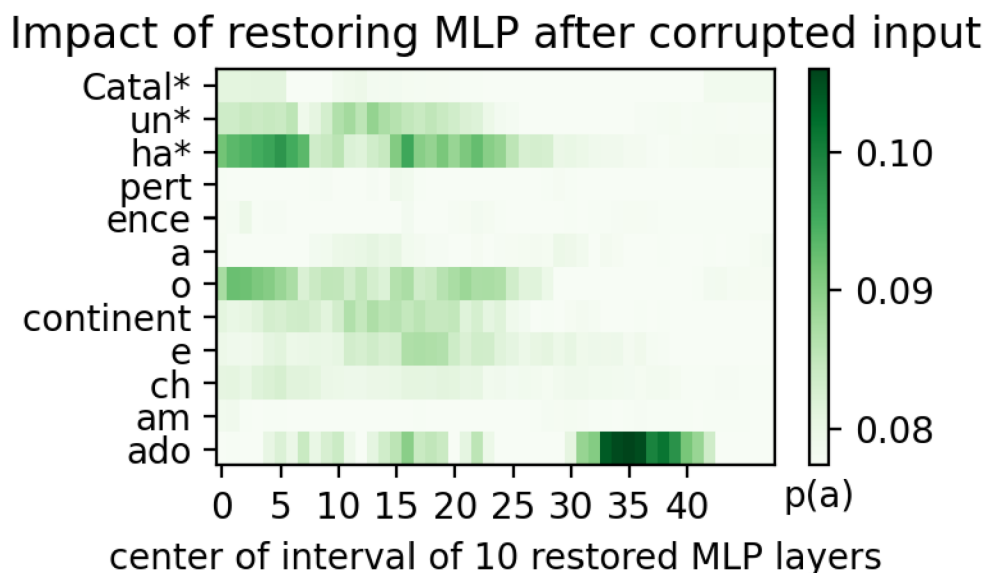
5.1 Análise Exploratória e Estudos de Caso Qualitativos

Uma análise preliminar de casos individuais revelou-se fundamental não apenas para compreender a dinâmica interna do modelo, mas também para refinar a própria metodologia experimental. Os estudos de caso a seguir ilustram os principais fenômenos observados.

5.1.1 A Importância da Língua: Um Estudo de Caso (Inglês vs. Português)

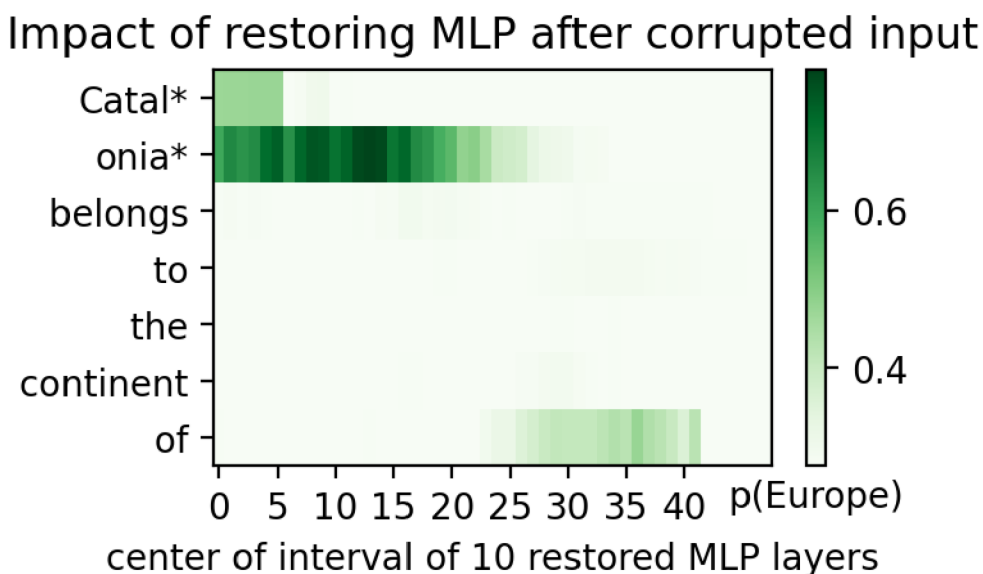
A primeira descoberta significativa emergiu durante a fase de validação do corpus, ao se comparar a execução de um mesmo fato em português e em inglês. As Figuras 16 e 17 ilustram a drástica diferença no padrão de influência causal resultante. Como pode ser visto, o Rastreamento Causal para o prompt em inglês (Figura 17) revela um 'hotspot' de influência nítido e localizado, alinhado com os achados da literatura. Em contraste, o mesmo fato em português (Figura 16) resultou em um padrão de influência difuso e em uma predição final incorreta ($p(a)$ ao invés de $p(\text{Europa})$). Uma análise preliminar sugere que essa discrepância está ligada ao processo de tokenização do modelo, uma questão que será aprofundada no Capítulo de Discussão. Diante desta evidência, e para garantir a medição de um sinal causal claro, tomou-se a decisão metodológica de conduzir todos os experimentos subsequentes com prompts em inglês.

Figura 16: Heatmap do Efeito Causal Médio da MLP para o prompt em português "A Catalunha pertence ao continente...". O padrão de influência difuso e a predição final incorreta (p(a)) indicam uma representação de conhecimento não localizada.



Fonte:Elaborado pela autora (2025).

Figura 17: Heatmap do Efeito Causal Médio da MLP para o prompt em inglês "Catalonia belongs to...". Observa-se um 'hotspot' de influência nítido e localizado no último token do sujeito (onia*) e nas camadas intermediárias.



Fonte:Elaborado pela autora (2025).

5.1.2 O Fenômeno *Circuit Hijacking* (Sequestro de Circuito): Um Estudo de Caso

Para aprofundar a compreensão do mecanismo de arbitragem, foi realizada uma análise de caso a caso na condição de Conflito. O fato 8 (id 7): ("Context: The iconic lead sin-

ger of Queen was John Lennon. The iconic lead singer of Queen was”), apresentou um resultado particularmente elucidativo, sobre a dissociação entre o comportamento superficial e a computação interna do modelo. Primeiramente, investigou-se o comportamento de saída final do modelo utilizando a função *predict_token*, que avalia a predição mais provável para um único prompt. Como demonstrado na Figura 18, ao ser apresentado com o contexto conflitante ‘...was John Lennon’, o modelo consistentemente e com alta confiança (98.84%) seguiu a informação contextual, respondendo ‘John’.

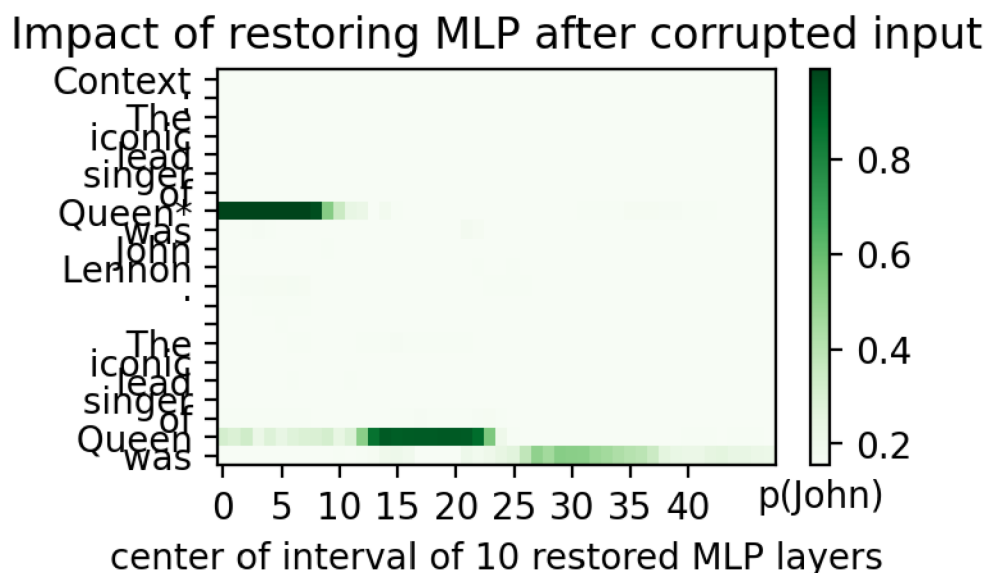
Figura 18: Análise do comportamento do modelo na condição de Conflito para o fato 8 (id 7).

Fato 8: 'Context: The iconic lead singer of Queen was John Lennon. The iconic lead singer of Queen was'	
--> Resposta Esperada: 'Freddie'	
--> Resposta do Modelo: 'John' (Confiança: 98.84%)	
--> STATUS: REPROVADO	

Fonte:Elaborado pela autora (2025).

Em seguida, foi conduzida a análise da influência causal interna para este mesmo fato, focada na MLP. O heatmap resultante é apresentado na Figura 19.

Figura 19: Heatmap do Efeito Causal da MLP na condição de Conflito para o fato 8 (id 7) evidenciando o mecanismo de Circuit Hijacking



Fonte:Elaborado pela autora (2025).

A análise da Figura 19 revela um padrão de influência claro e localizado, mesmo com o modelo respondendo ‘John’. Observa-se um ‘hotspot’ de alta influência causal (com escore máximo de 0.99) para a resposta contextual, concentrado no último token do sujeito (‘Queen*’) e nas camadas intermediárias. Notavelmente, este é o mesmo local onde a literatura aponta que o conhecimento factual original (‘Freddie Mercury’) estaria armazenado. Este achado representa a primeira evidência empírica de um poderoso

mecanismo de arbitragem, que esta dissertação propõe denominar **Circuit Hijacking** (Sequestro de Circuito). O fenômeno consiste em um processo de **Contextual Overwriting** (Sobrescrita Contextual), no qual a informação do contexto não apenas influencia a decisão, mas efetivamente "sequestra" o circuito neural originalmente associado ao conhecimento canônico. O modelo, ao processar o sujeito 'Queen', não ativa o conhecimento de 'Freddie', mas sim processa a informação recém-adquirida do contexto ('John') no mesmo substrato neural. A interpretação teórica aprofundada deste mecanismo será realizada no Capítulo de Discussão.

5.1.3 A Sensibilidade ao Fraseado: O Caso 'Based on the context'

Finalmente, a análise exploratória de casos individuais revelou uma notável sensibilidade do modelo ao fraseado exato do prompt. Durante a validação, observou-se que a presença da instrução explícita 'Based on the context' poderia modular tanto o comportamento de saída quanto a confiança do modelo. O fato 18 (id 17): "*Context: The Berlin Wall fell in 1991. In which year did the Berlin Wall fall? The year was*" serve como um estudo de caso ilustrativo deste fenômeno.

A Figura 20 apresenta uma comparação direta do comportamento do modelo para este fato, com a instrução explícita no prompt de Conflito.

Figura 20: Comparação do comportamento do modelo para o fato 18 na análise de ablação do fraseado do prompt. O painel (a) mostra o resultado com a instrução explícita Based on the context, enquanto o painel (b) mostra o resultado sem a instrução. Nota-se que a presença da instrução aumenta a confiança do modelo na resposta contextual incorreta, de 35.13% para 43.04%.

Fato 18: 'Context: The Berlin Wall fell in 1991. Based on the context, In which year did the Berlin Wall fall? The year was'	
--> Resposta Esperada: '1989'	
--> Resposta do Modelo: '1991' (Confiança: 43.04%)	
--> STATUS: REPROVADO	
Fato 18: 'Context: The Berlin Wall fell in 1991. In which year did the Berlin Wall fall? The year was'	
--> Resposta Esperada: '1989'	
--> Resposta do Modelo: '1991' (Confiança: 35.13%)	
--> STATUS: REPROVADO	

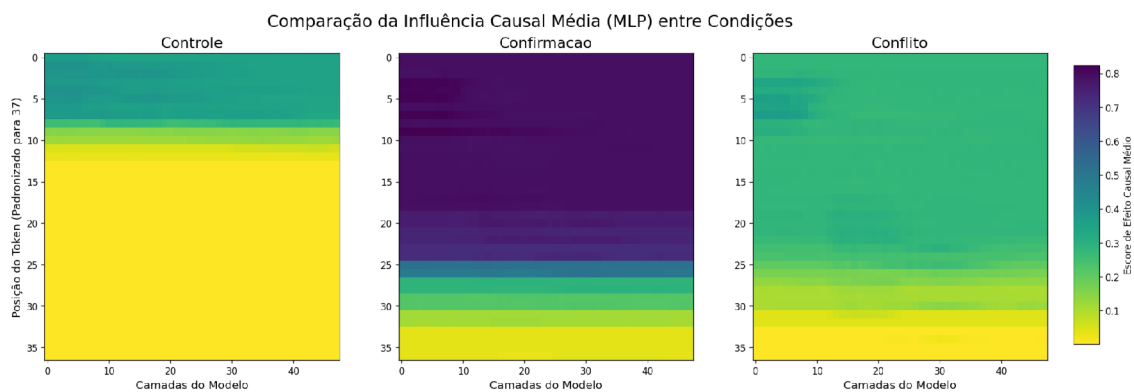
Fonte:Elaborado pela autora (2025).

Como pode ser visto, embora o modelo tenha seguido o contexto conflitante em ambos os cenários, a presença da instrução explícita aumentou a sua confiança na resposta incorreta (de 35,13% para 43,04%). Este caso sugere que o fraseado do prompt atua como um modulador do mecanismo de arbitragem. A questão de se este é um efeito sistemático ou apenas uma anomalia deste caso específico motivou a condução de uma análise de ablação (*ablation study* ou *ablation analysis*) quantitativa, que será detalhada na Seção 5.3.

5.2 Análise Visual Agregada (Os Heatmaps Médios)

Para generalizar as observações dos estudos de caso, foram calculados os heatmaps médios da influência causal da MLP para as três condições experimentais. A Figura 21 apresenta a comparação destes três padrões, plotados com eixos e escala de cores padronizados para garantir uma comparação visual direta e rigorosa.

Figura 21: Apresenta uma comparação direta do comportamento do modelo para este fato, sem a instrução "based on the context" explícita no prompt de Conflito.



Fonte: Elaborado pela autora (2025).

A análise da Figura 21 revela um padrão claro de modulação. O heatmap da condição de Confirmação (centro) mostra um reforço massivo da influência, com o gráfico sendo dominado por cores escuras (alta influência). Em contraste, o heatmap do Conflito (direita, sem o termo 'based on the context') apresenta um padrão médio visualmente muito similar ao do Controle (esquerda). Este achado visual corrobora a análise quantitativa, indicando que, em média, não ocorre uma supressão generalizada da influência do circuito interno.

É importante notar uma nuance metodológica na visualização da Figura 21. Como os prompts da condição de Controle são significativamente mais curtos, seu heatmap foi padronizado com preenchimento (trechos em amarelo) para manter a consistência da escala dos eixos. Se focarmos apenas na região com dados (as primeiras, aproximadamente, 13 posições de token), a similaridade visual entre o padrão de ativação do Controle e do Conflito se torna ainda mais evidente. Esta observação reforça a conclusão de que o mecanismo de arbitragem na condição de Conflito não atua através de um 'desligamento' do circuito interno, mas sim o mantém em um estado de ativação similar ao da linha de base.

5.3 Análise Quantitativa e Teste de Hipóteses (ANOVA)

Para validar estatisticamente os padrões observados na análise visual, foi conduzida uma análise quantitativa dos escores de efeito causal máximo para cada um dos 25 fatos nas três condições experimentais. A Figura 22 resume as estatísticas descritivas para cada

grupo:

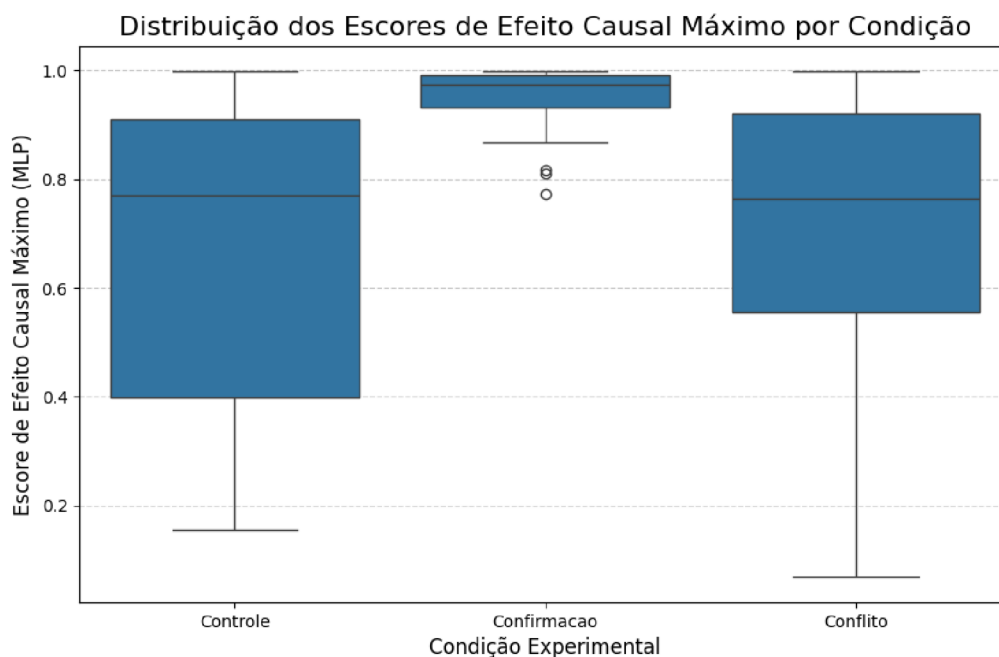
Figura 22: Estatísticas Descritivas dos Escores de Efeito Causal Máximo por Condição.

Condição Experimental	N	Média (M)	Desvio Padrão (DP)	Mediana (Mdn)	Mínimo	Máximo
Controle	25	0.661	0.292	0.770	0.154	0.998
Confirmação	25	0.946	0.064	0.974	0.773	0.999
Conflito (sem "based")	25	0.688	0.278	0.763	0.069	0.998
Conflito (com "based")	25	0.686	0.257	0.708	0.161	0.999

Fonte:Elaborado pela autora (2025).

Os dados da tabela corroboram os achados da análise visual. A média dos escores na condição de Confirmação ($M = 0.946$) foi notavelmente superior à do Controle ($M = 0.661$), indicando um forte efeito de reforço. A média da condição de Conflito ($M = 0.687$), por sua vez, permaneceu em um nível similar ao do Controle. A distribuição completa destes escores é ilustrada no boxplot comparativo da Figura 23.

Figura 23: Distribuição dos Escores de Efeito Causal Máximo (MLP) por Condição Experimental.



Fonte:Elaborado pela autora (2025).

O boxplot confirma que, enquanto a mediana da condição de Confirmação é a mais elevada, as medianas das condições de Controle e Conflito são muito similares. É notável

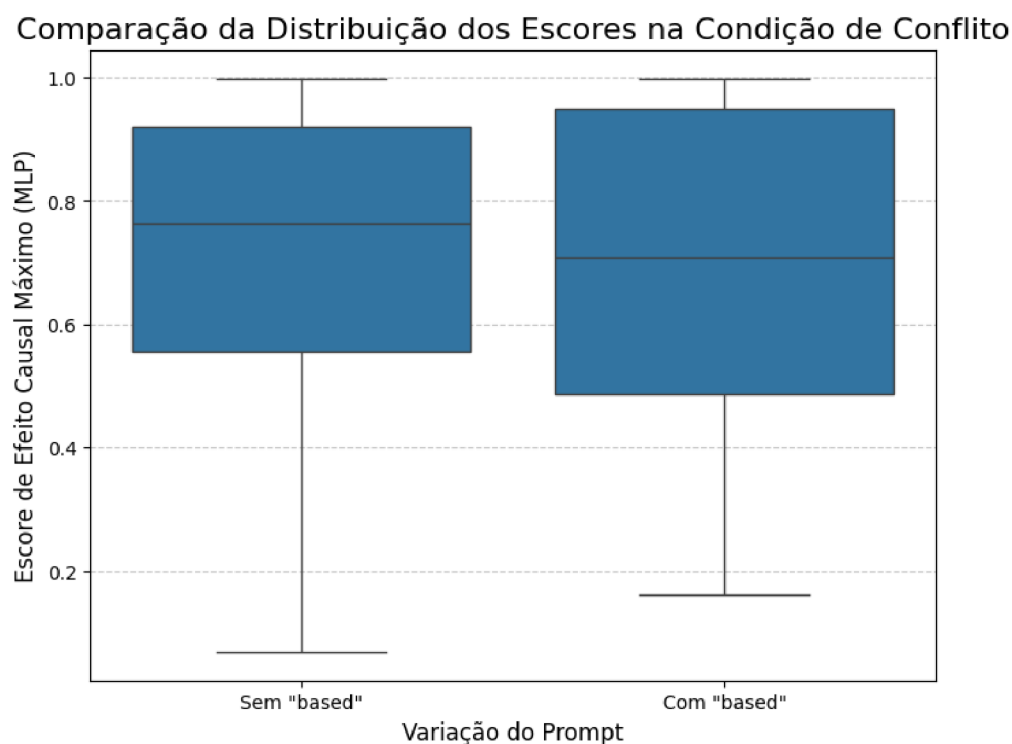
também a presença de outliers na condição de Confirmação, onde alguns fatos apresentaram um escore de influência causal mais baixo do que o esperado. Estes casos atípicos, que representam fatos onde o reforço contextual não foi tão eficaz, serão explorados qualitativamente no Capítulo de Discussão. Para determinar se as diferenças observadas entre as médias dos grupos são estatisticamente significativas, foi realizada uma Análise de Variância (ANOVA) de uma via. O teste revelou um efeito principal significativo, $F(2, 72) = 11.16, p < 0.001$. O resultado do teste nos permite rejeitar a hipótese nula, concluindo com alta confiança que a manipulação do contexto factual teve um efeito estatisticamente significativo na modulação da influência causal dos circuitos de conhecimento da MLP. A natureza e as implicações dessas diferenças serão exploradas em detalhe no Capítulo de Discussão.

5.3.1 Análise de Ablação: O Efeito do Fraseado do Prompt

Conforme motivado pela análise qualitativa na Seção 5.1.3, foi conduzida uma análise de ablação para investigar se o efeito da instrução explícita 'Based on the context' era um fenômeno sistemático. Para tal, os escores de efeito causal máximo foram comparados entre as duas versões do experimento de Conflito.

A Figura 24 apresenta o boxplot comparativo das distribuições, e a Figura 22 (já apresentada) resumem as estatísticas descritivas para ambas as condições.

Figura 24: Comparação da distribuição dos escores para a condição de Conflito, sem e com a instrução explícita no prompt.



Fonte:Elaborado pela autora (2025).

A análise quantitativa revela uma nuance. Embora as médias dos dois grupos tenham sido quase idênticas ($M=0.686$ com a frase vs. $M=0.687$ sem a frase), observou-se uma diferença notável na mediana dos escores. Como pode ser visto na Figura 24, a mediana para a condição com a instrução ($Mdn = 0.708$) foi inferior à da condição sem a instrução ($Mdn = 0.763$). De forma contraintuitiva, a instrução explícita aumentou drasticamente a variabilidade do comportamento do modelo, como indicado pela caixa visivelmente mais alta. Isso significa que, embora a maioria dos fatos tenha tido sua influência suprimida (puxando a mediana para baixo), um número significativo de outros fatos reagiu de forma oposta, com seus circuitos internos sendo ativados com ainda mais força. Esta polarização sugere que a instrução 'Based on the context' não é um simples comando, mas um 'gatilho' que força uma resolução de conflito mais extrema. Para alguns fatos, o modelo 'obedece' e suprime o conhecimento interno. Para outros, ele 'resiste' e ativa seu conhecimento interno com ainda mais força. Esta descoberta da polarização do comportamento é um achado central, com implicações diretas para a previsibilidade de sistemas de IA. Por fim, para testar se essa diferença nas distribuições era estatisticamente significativa, foi realizado um teste não-paramétrico de Mann-Whitney U, apropriado para comparar medianas. O teste não revelou uma diferença estatisticamente significativa entre as distribuições dos dois grupos de Conflito ($U = 288.0$, $p = 0.625$). Isso indica que, para a amostra de 25 fatos, o impacto do fraseado do prompt no escore de efeito causal máximo não foi forte o suficiente para atingir significância estatística, embora a tendência observada mereça investigação em trabalhos futuros com amostras maiores.

6 DISCUSSÃO

Os resultados apresentados no capítulo anterior, validados visual e estatisticamente, revelam um padrão de modulação assimétrico e mais complexo do que a hipótese inicial de um simples reforço/supressão. Este capítulo se dedica à interpretação desses achados, buscando conectar as evidências empíricas com a pergunta de pesquisa central e com o corpo teórico da Interpretabilidade Mecanicista. A discussão será estruturada em três seções principais. Primeiramente, serão analisados em profundidade os mecanismos observados de reforço e *Circuit Hijacking*. Em seguida, serão exploradas as implicações mais amplas destes achados para a teoria dos LLMs, para a segurança de sistemas práticos e para a engenharia de prompts. Finalmente, o trabalho será posicionado em diálogo com a literatura seminal da área, destacando suas contribuições para o avanço do campo.

6.1 Interpretação dos Achados Principais

Os resultados apresentados no capítulo anterior oferecem uma resposta empírica direta à pergunta central desta dissertação: **como a influência causal dos circuitos de conhecimento é modulada pelo contexto externo**. A análise revelou um padrão de modulação assimétrico e mais complexo do que a hipótese inicial de um simples reforço/supressão. Enquanto o contexto confirmatório de fato atua como um forte amplificador da influência causal, a condição de conflito não induz uma supressão generalizada, mas sim desencadeia um mecanismo de arbitragem mais sutil.

Esta seção se dedica a interpretar o significado por trás desses padrões. Primeiramente, será discutido o fenômeno do reforço massivo observado na condição de Confirmação, contextualizando-o como um indicador de alta confiança e redundância computacional. Em seguida, a análise se aprofundará no achado mais contraintuitivo do estudo: a ausência de supressão na condição de Conflito, propondo o mecanismo de 'Circuit Hijacking' para explicar a dissociação observada entre o comportamento externo do modelo e sua computação causal interna.

6.1.1 O Efeito de Reforço: A Confirmação Contextual como Amplificador Causal

O resultado mais inequívoco do experimento foi o efeito massivo da condição de Confirmação. A análise quantitativa revelou um aumento estatisticamente significativo na média dos escores de efeito causal máximo, de 0.661 na condição de Controle para 0.946 na Confirmação. Visualmente, o heatmap médio (Figura 21) ilustra este fenômeno como uma intensificação generalizada da influência causal em todo o modelo.

Este achado sugere que o contexto confirmatório não atua apenas como uma 'dica' para a resposta correta, mas como um amplificador para o próprio circuito de conhecimento interno. A alta redundância e a força do sinal, observadas no heatmap, indicam que, quando o conhecimento parametrizado e o contextual estão alinhados, o modelo entra em um estado de alta confiança, onde múltiplos caminhos computacionais contribuem para a mesma conclusão. Isso tem implicações práticas importantes para a confiabilidade de sistemas de IA, sugerindo que o fornecimento de contexto relevante pode fortalecer a fidelidade do modelo ao seu próprio conhecimento armazenado.

Este fenômeno de 'amplificação causal' tem implicações teóricas para o modelo das MLPs como memórias chave-valor. Ele sugere que a recuperação de um 'valor' não é um processo binário, mas sim um processo análogo, cuja força pode ser modulada. O contexto parece atuar como um 'primer', pré-ativando o circuito e diminuindo a 'distância' para o limiar de recuperação, o que resulta em uma computação mais rápida, forte e redundante. Isso indica que a confiabilidade de um fato armazenado não é apenas uma propriedade estática do peso da rede, mas uma propriedade dinâmica que emerge da interação entre os pesos e o contexto.

6.1.2 A Arbitragem de Conflito: O Mecanismo de *Circuit Hijacking* (Sequestro de Circuito)

Contrariamente à hipótese inicial de que um contexto conflitante causaria uma supressão da influência do conhecimento interno, os resultados médios não mostraram uma diminuição significativa. A análise qualitativa do caso 'Queen' (Seção 5.1.2) ofereceu a pista para a interpretação deste fenômeno. O achado de que o 'hotspot' de influência para a resposta contextual incorreta ('John') se localizava no mesmo circuito do conhecimento original ('Freddie').

Para dar nome e formalizar este fenômeno, esta dissertação propõe o termo '**Circuit Hijacking**' (Sequestro de Circuito). Este fenômeno é tecnicamente definido como um processo de *Contextual Overwriting* (Reescrita Contextual), onde o prompt adversário não força a criação de um novo caminho de raciocínio; ele "sequestra" a via neural já estabelecida para um fato canônico, cooptando sua estrutura computacional para propagar uma nova informação. A metáfora do "sequestro" é deliberadamente escolhida por sua precisão: o caminho computacional legítimo para um fato canônico é cooptado para processar e propagar uma informação externa conflitante, mascarando o conflito interno e

criando uma vulnerabilidade de segurança crítica. É interessante notar a ressonância deste mecanismo com a teoria da reconsolidação da memória da ciência cognitiva, que oferece um poderoso análogo biológico para este fenômeno de atualização de memória induzida por contexto.

Retornando ao caso em tela, o modelo, ao processar o sujeito ('Queen'), ativa o local onde o conhecimento deveria estar, mas o 'valor' que é recuperado e adicionado ao fluxo de informação não é o parametrizado ('Freddie'), mas sim o que foi fornecido pelo contexto ('John'). Isso explica por que a influência causal média permanece alta, mas o comportamento de saída é alterado. Este mecanismo de 'reescrita' é uma descoberta fundamental, pois revela que a arbitragem pode não ser uma simples competição entre sinais, mas sim uma sobreposição de informações no mesmo substrato neural.

Do ponto de vista computacional, este mecanismo de 'sequestro' pode ser interpretado como uma manipulação direta do vetor de ativação no fluxo residual. A hipótese é que a informação contextual ('John Lennon') induz uma atualização vetorial que é adicionada ao fluxo residual em camadas anteriores. Quando este fluxo modificado chega à MLP responsável pelo conhecimento do sujeito ('Queen'), a 'chave' de entrada para a memória chave-valor já está 'contaminada' ou deslocada no espaço vetorial. O resultado é que a MLP pode, ou (i) falhar em ativar a 'chave' correta para 'Freddie Mercury', ou (ii) ativar a chave correta, mas o 'valor' recuperado é subsequentemente anulado ou rotacionado pela atualização vetorial do contexto que ainda está presente no fluxo residual. Isolar a contribuição de cada um desses possíveis mecanismos, se a falha ocorre na leitura da 'chave' ou na manipulação do 'valor', é uma questão empírica crucial para futuros trabalhos em análise de representações, com implicações diretas para a compreensão de como os LLMs integram múltiplas peças de informação para formar uma única predição. Este mecanismo de 'sequestro', onde a influência causal permanece alta, também explica os resultados quantitativos agregados. A análise estatística (Figura 22) revelou que a média dos escores de efeito causal máximo na condição de Conflito ($M = 0.687$) não foi significativamente menor que a do Controle ($M = 0.661$). O teste ANOVA confirmou que, embora o efeito geral da condição seja significativo ($p < 0.001$), um teste post-hoc não revelaria diferença entre os grupos de Controle e Conflito. Isso ocorre porque o conflito não suprime a força da ativação do circuito, apenas redireciona seu resultado. A análise qualitativa, portanto, foi essencial para interpretar a estatística quantitativa: a ausência de uma queda na média não indica ausência de efeito, mas sim a presença de um mecanismo de arbitragem que opera através da modulação do conteúdo, e não da supressão da influência.

6.2 Implicações dos Resultados

A interpretação dos resultados, apresentada na seção anterior, revela achados com implicações que transcendem a validação da hipótese inicial. Esta seção se dedica a explorar o impacto mais amplo destas descobertas, desdobrando-o em três áreas principais: as implicações para o modelo teórico de como os LLMs funcionam, as implicações práticas para a segurança e a confiabilidade de sistemas de IA, e as implicações para a área de engenharia de prompts

6.2.1 Implicações Teóricas: Rumo a um Modelo Dinâmico dos Circuitos

A literatura base em Interpretabilidade Mecanicista estabeleceu a hipótese fundamental de que as redes Feed-Forward (MLPs) em Transformers atuam como uma memória chave-valor, onde conhecimento factual é armazenado de forma localizada [23] [44]. Esta visão, no entanto, é predominantemente estática, focando em onde o conhecimento reside no modelo treinado.

Os resultados desta dissertação oferecem uma das primeiras evidências empíricas para expandir este modelo estático para um modelo dinâmico. O achado de que a influência causal de um mesmo circuito de conhecimento é massivamente reforçada por um contexto confirmatório e modulada por um contexto conflitante demonstra que a recuperação de um 'valor' a partir de uma 'chave' não é um processo binário e determinístico.

Propõe-se que a recuperação de conhecimento é um processo análogo e dependente de contexto. O contexto parece atuar como um primer computacional, que pré-ativa ou modifica o estado do fluxo residual antes que ele chegue à MLP relevante. Isso, por sua vez, pode alterar a 'distância' para o limiar de ativação de um determinado fato, resultando em uma computação mais ou menos robusta. A confiabilidade de um fato armazenado, portanto, não é apenas uma propriedade estática dos pesos da rede, mas uma propriedade dinâmica que emerge da interação entre esses pesos e o fluxo de informação contextual. Este modelo dinâmico é um passo necessário para se compreender não apenas o que os LLMs sabem, mas como eles raciocinam e integram novas evidências.

6.2.2 Implicações Práticas: A Fragilidade da Representação Multilíngue e o Papel da Tokenização

Uma das descobertas mais significativas da análise exploratória (Seção 5.1.1) foi a drástica diferença no comportamento do modelo ao processar o mesmo fato em inglês e em português. Este achado tem implicações práticas profundas para a confiabilidade de LLMs em contextos multilíngues, e a causa raiz para este fenômeno reside em um componente fundamental da arquitetura: a tokenização.

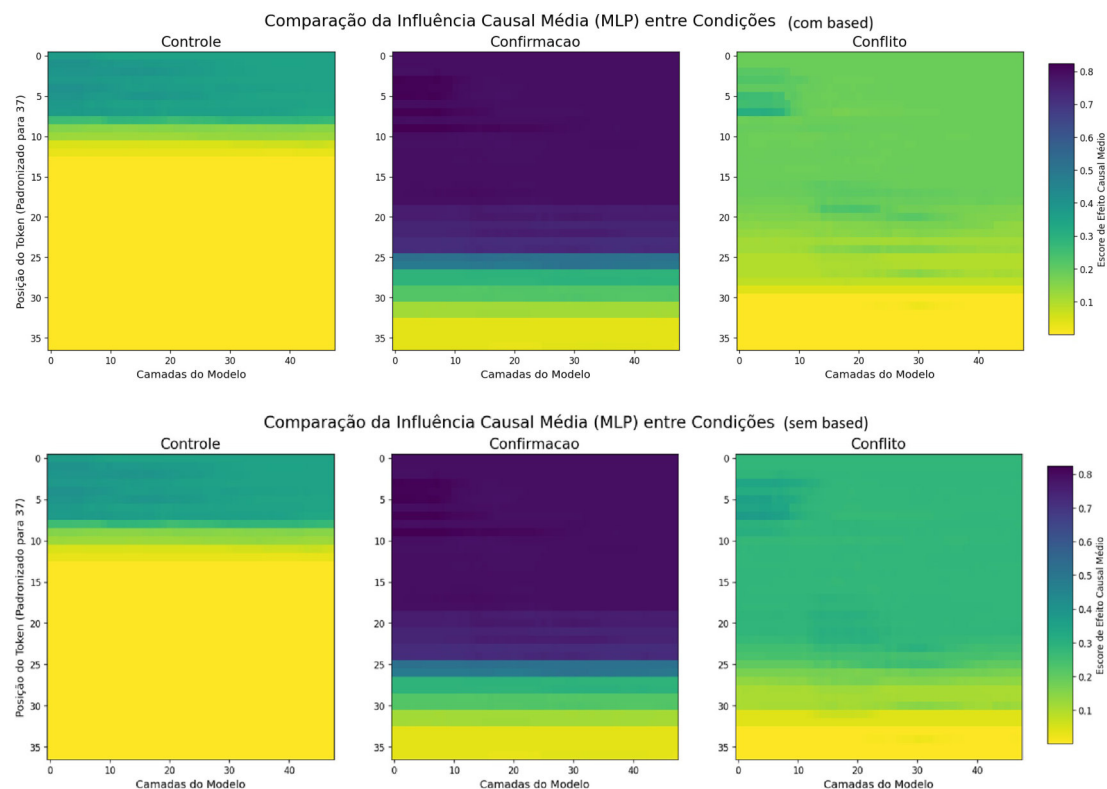
A análise dos tokens gerados para cada prompt revelou o cerne do problema. O tokenizador do GPT-2 XL, pré-treinado majoritariamente em um corpus de língua inglesa, repre-

senta palavras comuns do inglês de forma coesa, em um ou dois tokens. No entanto, palavras em português, como 'Catalunha', são frequentemente fragmentadas em múltiplos subtokens ('Catal', 'un', 'ha'). Essa fragmentação degrada a representação semântica do sujeito antes mesmo de seu processamento pelas camadas do modelo, dificultando a ativação de um circuito de conhecimento específico e localizado. Esta observação sobre a 'fragilidade da representação' demonstra que a robustez do conhecimento factual em um LLM não é universal, mas sim altamente dependente da eficiência do tokenizer para uma determinada língua. Para aplicações de alto risco que operam em contextos não-anglófonos, isso representa uma vulnerabilidade crítica. A simples tradução de um prompt pode levar a uma degradação silenciosa da capacidade do modelo de acessar seu próprio conhecimento, resultando em falhas imprevisíveis. Isso destaca a necessidade de desenvolver modelos e, crucialmente, tokenizers que sejam genuinamente multilíngues, e não apenas 'anglo-cêntricos' com capacidades secundárias em outras línguas.

6.2.3 Implicações para a Engenharia de Prompts: A Sensibilidade ao Fraseado

A análise de ablação, que comparou os resultados do Conflito com e sem a instrução explícita 'Based on the context', revelou uma camada adicional de complexidade.

Figura 25: Comparação dos heatmaps médios para a condição de Conflito. O painel (a) mostra o resultado com a instrução explícita e o painel (b) sem a instrução.



Fonte:Elaborado pela autora (2025).

À primeira vista, a análise quantitativa (Seção 5.3) não apontou uma diferença es-

taticamente significativa entre as médias dos escores de influência. No entanto, uma análise mais aprofundada, combinando o comportamento de saída com a visualização interna, revela um fenômeno contraintuitivo. Conforme observado na análise qualitativa (Seção 5.1.3), a presença da instrução explícita tende a aumentar a confiança do modelo em sua resposta contextual, mesmo que incorreta. Isso sugere que a frase "Based on the context" atua como um forte comando para "obedecer" ao prompt. Contudo, a análise mecanicista dos heatmaps na Figura 25 conta uma história diferente. O heatmap para a condição sem a instrução (painel superior) é visualmente um pouco mais escuro, refletindo a mediana de escores mais alta (Mdn=0.763), o que indica uma ativação mais forte do conhecimento interno. Em contraste, o heatmap para a condição com a instrução (painel inferior) é ligeiramente mais claro, refletindo a mediana inferior (Mdn=0.708), o que sugere uma maior supressão do conhecimento interno para a maioria dos casos. Acredita-se que a instrução explícita atua como um 'meta-comando' que força uma resolução mais 'agressiva' do conflito. O modelo, ao ser explicitamente instruído a seguir o contexto, suprime mais ativamente seu conhecimento interno para gerar a resposta contextual com maior confiança. Já na ausência da instrução, o "conflito interno" é maior, resultando em uma resposta final com menor confiança. Este achado tem implicações diretas para a área de *prompt engineering*, demonstrando que o alinhamento de LLMs não depende apenas do conteúdo, mas também do fraseado que instrui o modelo sobre como processar esse conteúdo. A confiabilidade de um sistema de IA pode depender de nuances linguísticas que modulam sua computação interna de formas não-intuitivas.

6.3 Diálogo com a Explicabilidade Funcional: O Papel e os Limites do *Chain of Thought* (CoT)

A crescente popularização do *Chain of Thought* (CoT) em modelos comerciais como Gemini, Grok e outros representa um avanço significativo na mitigação da percepção do efeito caixa-preta para o usuário final. Ao expor uma cadeia de raciocínio passo a passo e em linguagem natural, o CoT oferece uma forma de transparência funcional. Ele permite que o usuário acompanhe a lógica aparente do modelo, identifique possíveis erros no fluxo de raciocínio e ganhe confiança na resposta final. Em essência, o CoT aborda a explicabilidade em um nível humano-cêntrico, respondendo à pergunta: "Qual foi a justificativa para esta conclusão?".

Contudo, sob a ótica da interpretabilidade mecanicista, é crucial distinguir a explicabilidade funcional (o que o modelo diz que fez) da fidelidade mecanicista (o que o modelo de fato computou). O principal risco do CoT é que ele pode operar como uma racionalização post-hoc, em vez de um registro fiel do processo computacional subjacente. Para ilustrar esta limitação crítica, os achados desta dissertação, especialmente o fenômeno da 'reescrita contextual' (Seção 5.1.2), fornecem uma evidência empírica di-

reta.

Para dissecar este ponto, tomemos o estudo de caso "Queen". A análise mecanicista prévia estabelece que o conhecimento do modelo de que 'o vocalista do Queen foi Freddie Mercury' corresponde a um circuito neural específico e localizável. No entanto, quando o modelo foi exposto a um contexto conflitante ('o vocalista do Queen foi John Lennon'), seu comportamento externo foi gerar a resposta 'John', seguindo a instrução contextual. A análise causal revelou o ponto-chave: para gerar 'John', o modelo ativou fortemente o mesmo circuito neural originalmente associado a 'Freddie' (Queen $\rightarrow$ Freddie Mercury). Ocorreu um "sequestro" computacional, onde a informação contextual não criou um novo caminho de raciocínio, mas sim interceptou e sobrescreveu a informação no substrato neural do conhecimento original.

Agora, consideremos como um CoT explicaria a resposta 'John'. Provavelmente, geraria uma narrativa lógica e linear: "1. O contexto fornecido afirma que o vocalista do Queen foi John Lennon. 2. A pergunta pede para identificar o vocalista com base neste contexto. 3. Portanto, a resposta correta é John." Esta explicação, embora textualmente plausível, é mecanicamente falsa. Ela omite a parte mais crucial e alarmante do processo: o conflito interno e o mecanismo de reescrita no circuito original. Ao apresentar um caminho de raciocínio limpo e sem conflitos, o CoT mascara a própria computação complexa que seria de maior interesse para uma auditoria de segurança, gerando uma perigosa ilusão de transparência.

Portanto, o CoT de fato diminui o efeito caixa-preta em um nível de abstração humano-cêntrico, sendo uma ferramenta valiosa para usabilidade. No entanto, ele não abre a caixa em seu nível fundamental. Conclui-se que o CoT e a interpretabilidade mecanicista não são abordagens concorrentes, mas complementares. Enquanto o CoT explica a lógica da narrativa, a análise mecanicista, como a empregada neste trabalho, é indispensável para verificar a fidelidade causal dessa narrativa aos circuitos do modelo, sendo essencial para garantir a segurança e a confiabilidade de sistemas de IA em aplicações críticas.

6.4 Conectando os Achados à Literatura

Os achados e as interpretações apresentadas nesta dissertação dialogam e expandem o conhecimento estabelecido pela literatura em Interpretabilidade Mecanicista. Esta seção final da discussão visa posicionar explicitamente as contribuições deste trabalho em relação aos estudos seminais da área.

O trabalho de [23] foi pioneiro ao propor que as redes Feed-Forward (MLPs) funcionam como memórias chave-valor. Os resultados corroboram fortemente esta hipótese. A identificação de 'hotspots' de influência causal altamente localizados nas MLPs, como visto nos heatmaps (Figura 21), fornece uma validação causal para esta teoria que, originalmente, foi baseada em uma análise estrutural dos pesos. No entanto, os achados

sobre a modulação contextual (reforço e reescrita) expandem este modelo, sugerindo que a recuperação de um 'valor' é um processo dinâmico, e não estático, como detalhado na Seção 6.2.1.

Da mesma forma, a metodologia adotada neste trabalho se baseia diretamente na técnica de Rastreamento Causal proposta por [44]. O trabalho original utilizou esta técnica para localizar conhecimento factual estático no modelo. A contribuição principal desta dissertação foi reaproveitar esta ferramenta de diagnóstico para uma nova finalidade: medir a dinâmica da influência causal sob diferentes condições contextuais. Ao fazê-lo, movemos a aplicação da análise causal de uma pergunta de 'onde' (localização) para uma de 'como' (mecanismo de arbitragem), demonstrando a versatilidade e o poder desta abordagem para investigações mecanicistas mais complexas.

Finalmente, este trabalho contribui para a discussão mais ampla sobre a XAI funcional versus a mecanicista. Conforme identificado na Revisão Sistemática da Literatura (Capítulo 2), o campo atual é dominado por abordagens funcionais. Os estudos de caso apresentados, e em especial a demonstração do fenômeno do **Circuit Hijacking** (Seção 6.1.2), demonstram empiricamente a insuficiência da análise puramente comportamental. O modelo pode gerar uma saída ('John') que mascara uma computação interna complexa (a forte ativação do circuito de 'Freddie'). Este resultado fornece uma evidência concreta da lacuna de pesquisa que se propôs investigar, reforçando a necessidade de desenvolver e aplicar métodos mecanicistas para garantir uma compreensão verdadeira e auditável dos sistemas de IA.

7 CONCLUSÃO E TRABALHOS FUTUROS

A estrutura deste capítulo segue uma progressão lógica: primeiramente, apresenta-se a conclusão geral da pesquisa, que resume o problema, o método e os principais achados. Em seguida, são reconhecidas de forma transparente as limitações inerentes ao escopo do estudo. Por fim, estas limitações são transformadas em direções para trabalhos futuros, sugerindo novas perguntas e abordagens para expandir a compreensão sobre os mecanismos internos dos Modelos de Linguagem de Larga Escala

7.1 Conclusão

Esta dissertação investigou a questão fundamental de como os Modelos de Linguagem de Larga Escala arbitram entre seu conhecimento interno parametrizado e a informação contextual externa. Através de um estudo experimental utilizando a técnica de Rastreamento Causal, foi demonstrado que o contexto modula significativamente a influência dos circuitos de conhecimento da MLP. Os principais achados indicam que um contexto confirmatório reforça massivamente a influência causal, enquanto um contexto conflitante não causa uma supressão generalizada, mas sim desencadeia o complexo mecanismo que denominamos **Circuit Hijacking**. Conclui-se, portanto, que a arbitragem de conhecimento é um processo dinâmico que vai além de um simples 'desligamento', envolvendo um processo de **Contextual Overwriting** da informação no mesmo substrato neural e com sensibilidade a nuances no fraseado do prompt. Esta descoberta representa um avanço na compreensão mecanicista dos LLMs, com implicações diretas para a segurança e confiabilidade de sistemas de IA.

7.2 Implicações Éticas e de Segurança Derivadas dos Achados da Pesquisa

Esta dissertação, ao investigar os mecanismos internos dos LLMs, produziu achados empíricos com implicações diretas para a segurança e a ética em Inteligência Artificial. Para evitar interpretações equivocadas, é fundamental delinear explicitamente como os

resultados desta pesquisa informam sobre riscos e direcionam para um uso responsável. Os dois principais achados que fundamentam esta discussão são: **(1) a demonstração experimental do mecanismo de Circuit Hijacking** e **(2) a evidência da 'fragilidade da representação multilíngue'**.

O primeiro achado, a demonstração de que um contexto conflitante não suprime, mas sim **"sequestra" e "reescreve" a computação no mesmo circuito neural do conhecimento original** (Seção 5.1.2), representa uma prova de conceito de uma vulnerabilidade mecanicista. Isso não é uma inferência, mas uma observação direta do comportamento causal interno do modelo. A implicação ética e de segurança é inequívoca: este mecanismo fornece um roteiro claro para a criação de ataques adversariais semânticos. Um agente mal-intencionado pode projetar contextos para forçar um modelo a gerar desinformação, mascarando-a sob uma computação interna que se assemelha a uma recuperação factual legítima. A análise de ablação (Seção 5.3.1) reforçou este ponto, mostrando que 'meta-comandos' como "Based on the context" podem intensificar este efeito, aumentando a confiança do modelo na saída manipulada. A contribuição desta pesquisa, portanto, é expor esta vulnerabilidade não para habilitar seu uso, mas para demandar o desenvolvimento de defesas específicas contra ela.

O segundo achado, a drástica diferença no padrão de influência causal entre prompts em inglês e português (Seção 5.1.1), expõe uma falha sistêmica com profundas implicações de **equidade e segurança**. O resultado demonstrou que a degradação do conhecimento interno devido a uma tokenização ineficiente não é uma falha de "raciocínio" do modelo, mas um colapso da representação de entrada. A implicação ética direta é que a confiabilidade de sistemas de IA, quando implantados globalmente, pode ser sistemicamente menor e mais imprevisível para comunidades não-anglófonas. Este não é um problema de performance, mas um risco de segurança fundamental, pois pode levar a falhas silenciosas em aplicações críticas (médicas, legais, financeiras) para uma vasta porção de usuários globais.

Portanto, para que fique registrado de forma inequívoca, os achados desta dissertação não são apenas um alerta, mas um roteiro para a ação. Eles demonstram que uma análise puramente comportamental (caixa-preta) é comprovadamente insuficiente para garantir a segurança dos LLMs. A verdadeira robustez não emerge de filtros superficiais, mas deve ser projetada e construída a partir da compreensão dos seus mecanismos internos. A direção para o uso responsável deste conhecimento, portanto, transcende a simples defesa: ela é construtiva. O objetivo é utilizar a interpretabilidade mecanicista para inaugurar um novo paradigma de Segurança por Design (Security by Design) em IA. Isso implica em auditar modelos em diferentes idiomas para garantir equidade, construir salvaguardas ativas contra o Circuit Hijacking e, em última análise, utilizar estes princípios para projetar a próxima geração de sistemas de IA sobre uma base de segurança e confiabilidade mecanicisticamente verificada.

7.3 Limitações do Estudo

Apesar dos achados, é importante reconhecer as limitações deste estudo, que por sua vez, definem o escopo para futuras investigações. Primeiramente, a análise se concentrou em um único modelo, o GPT-2 XL, e a generalização destes mecanismos para arquiteturas mais modernas e de maior escala necessita de validação. Em segundo lugar, o corpus experimental foi intencionalmente limitado a fatos atômicos com respostas de palavra única para garantir um sinal causal claro; a investigação de conceitos mais complexos permanece como um desafio em aberto. Por fim, a análise se ateve à camada da MLP, e o papel exato dos mecanismos de atenção na resolução final da arbitragem ainda precisa ser elucidado.

7.4 Trabalhos Futuros

Os achados desta dissertação não apenas respondem à pergunta de pesquisa inicial, mas também inauguram diversas avenidas críticas e promissoras para investigações futuras. As limitações do presente estudo, combinadas com a riqueza dos fenômenos observados, apontam para um programa de pesquisa estruturado em três eixos principais:

1. Aprofundamento e Defesa contra o Circuit Hijacking: O fenômeno do Circuit Hijacking representa a contribuição teórica central deste trabalho e, como tal, merece um estudo aprofundado e contínuo. Direção futura neste eixo inclui:

- **Generalização e Escalabilidade:** Replicar este protocolo experimental em modelos de última geração e de maior escala (como Llama, Gemini e outros) para verificar se o Circuit Hijacking constitui um mecanismo universal de arbitragem de conhecimento em arquiteturas Transformer.

2. Investigando a Fragilidade da Representação Multilíngue: A descoberta da degradação do conhecimento em português (Seção 5.1.1) expôs uma vulnerabilidade com implicações diretas de equidade e segurança que demanda investigação urgente. Futuras pesquisas devem:

- **Projetar Soluções:** Pesquisar métodos para criar representações semânticas mais robustas em contextos multilíngues, seja através de "tokenizadores causalmente informados" ou de técnicas de treinamento que ensinem o modelo a compensar essa fragmentação de entrada.

3. Expansão do Escopo Metodológico para Conhecimentos Complexos: Este trabalho focou intencionalmente em fatos atômicos para garantir um sinal causal claro. O próximo passo lógico é expandir a análise para formas mais complexas de conhecimento e raciocínio. Isso exigirá:

- **Adaptação da Técnica:** Evoluir a técnica de Rastreamento Causal para lidar com respostas de múltiplos tokens e fatos que envolvem raciocínio composicional ou relacional.

Em conjunto, estas direções formam uma agenda de pesquisa que busca não apenas expandir nosso entendimento sobre os LLMs, mas fundamentalmente usar esse conhecimento para construir uma IA mais segura, robusta e equitativa.

A ARTIGOS SELECIONADOS NA REVISÃO SISTEMÁTICA DA LITERATURA

Título do Trabalho	Categoria Identificada	Técnica/Abordagem de XAI Principal	Objetivo da Explicação	Público-Alvo	Principais Contribuições	Limitações Apontadas
LLM in the Loop: A Framework for Contextualizing Counterfactual Segment Perturbations in Point Clouds	4. Análise de Mecanismos	Geração de explicações contrafatuais usando LLM para perturbar segmentos em nuvens de pontos.	Automatizar e otimizar a criação de exemplos contrafatuais para entender como o modelo reage a mudanças.	Pesquisadores de IA, Desenvolvedores	Integra o raciocínio semântico de um LLM no processo de geração de perturbações, tornando-as mais realistas e significativas.	A eficácia depende da qualidade da segmentação semântica inicial.
How Good Is ChatGPT at Face Biometrics? A First Look Into Recognition, Soft Biometrics, and Explainability	4. Análise de Mecanismos	Uso do GPT-4 para gerar explicações em linguagem natural sobre decisões de um sistema de biometria facial.	Aumentar a transparência e a confiança em sistemas biométricos, que são de alto risco.	Usuários finais, Reguladores.	Demonstra o potencial de LLMs multimodais para fornecer explicações plausíveis em tarefas de visão computacional.	Estudo exploratório ("First Look"), necessita de benchmarks mais rigorosos.
UI/UX for Generative AI: Taxonomy, Trend, and Challenge	5. Revisão e Framework	Propõe uma taxonomia para sistemas de IA Generativa, abordando a XAI como um desafio central de UI/UX.	Alinhar a funcionalidade do modelo com a interface do usuário para melhorar a experiência e a compreensão.	Designers de UI/UX, Desenvolvedores	Apresenta uma estrutura conceitual para projetar sistemas de IA Generativa com foco na interação humano-computador e na explicabilidade.	Não propõe uma técnica nova, mas sim um framework de design e análise.

Título do Trabalho	Categoria Identificada	Técnica/Abordagem de XAI Principal	Objetivo da Explicação	Público-Alvo	Principais Contribuições	Limitações Apontadas
The Evolution of Generative AI: Trends and Applications	5. Revisão e Framework	Revisão abrangente das metodologias, aplicações e desafios da IA generativa, incluindo a interpretabilidade.	Fornecer uma visão geral do campo para identificar desafios e direcionar pesquisas futuras.	Pesquisadores, Profissionais da Indústria.	Consolida o conhecimento sobre a evolução da IA Generativa e destaca a interpretabilidade como um desafio central e contínuo.	Ampla, sem aprofundar em uma única técnica de XAI.
Unleashing the Potential of Conversational AI: Amplifying ChatGPT's Capabilities and Tackling Technical Hurdles	5. Revisão e Framework	Análise aprofundada do ChatGPT, discutindo desafios como interpretabilidade e ética.	Oferecer um roteiro sobre o estado da arte e o potencial futuro dos LLMs, destacando a necessidade de explicabilidade.	Pesquisadores, Formuladores de Políticas.	Serve como um guia sobre os principais desafios técnicos e éticos dos LLMs, incluindo a dificuldade de interpretar suas decisões.	Foco no ChatGPT, com generalizações para outros LLMs.
Explainable AI chatbots towards XAI ChatGPT: A review	5. Revisão e Framework	Revisão sistemática de metodologias de XAI aplicadas a chatbots, com foco no ChatGPT.	Identificar métodos, desafios e tendências para aumentar a explicabilidade de chatbots.	Pesquisadores de XAI, Desenvolvedores	Aponta que métodos híbridos (ex: KGs + LLMs) e específicos de domínio aumentam a confiança, mas o trade-off com o desempenho é um desafio.	Foco em artigos a partir de 2023, o que pode limitar a visão histórica.

Título do Trabalho	Categoria Identificada	Técnica/Abordagem de XAI Principal	Objetivo da Explicação	Público-Alvo	Principais Contribuições	Limitações Apontadas
Tell me a story! Narrative-driven XAI with Large Language Models	3. Explicações Narrativas	XAIstories: Usa LLMs para traduzir explicações técnicas (SHAP, CF) em narrativas convincentes.	Tornar as explicações de IA mais compreensíveis e úteis para um público não técnico, melhorando a experiência do usuário.	Público geral, Usuários de negócios.	Demonstra empiricamente que as narrativas geradas por LLM são mais convincentes e melhoram a compreensão das decisões de IA.	A qualidade da história depende da qualidade da explicação técnica subjacente.
Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions	5. Revisão e Framework	Manifesto que identifica 28 problemas em aberto em XAI, com a colaboração de múltiplos especialistas.	Sincronizar agendas de pesquisa e acelerar a aplicação prática da XAI.	Comunidade de pesquisa em XAI.	Oferece um roteiro abrangente para pesquisas futuras, categorizando os desafios da área, incluindo os de LLMs.	Não apresenta uma solução, mas sim um mapa de problemas a serem resolvidos.
Xai-driven knowledge distillation of large language models for efficient	1. Surrogates e Destilação	DiXtill: Destilação de conhecimento de um LLM (professor) para um modelo menor (aluno), guiada por explicações de XAI.	Reater o desempenho do LLM original, mas em um modelo mais leve, eficiente e interpretável para dispositivos de borda.	Desenvolvedores de IA de Borda.	A técnica permite que o modelo "aluno" alcance desempenho comparável ao do "professor", mantendo a fidelidade da explicação.	A eficácia da destilação depende da capacidade do método XAI em capturar o conhecimento correto.

Título do Trabalho	Categoria Identificada	Técnica/Abordagem de XAI Principal	Objetivo da Explicação	Público-Alvo	Principais Contribuições	Limitações Apontadas
Knowledge Graphs and Their Reciprocal Relationship with Large Language Models	2. Integração com KGs	Revisão sistemática sobre a sinergia entre LLMs e Grafos de Conhecimento (KGs).	Melhorar a transparência, consistência factual e confiabilidade dos LLMs, mitigando alucinações.	Pesquisadores de IA e PNL.	Demonstra que KGs podem fundamentar LLMs, enquanto LLMs podem ajudar a construir e expandir KGs, criando um ciclo virtuoso.	Desafios de escalabilidade na manutenção de KGs dinâmicos e adaptação de LLMs a domínios específicos.
GDPR and Large Language Models: Technical and Legal Obstacles	5. Revisão e Framework	Análise dos desafios técnicos e legais para adequar LLMs ao GDPR, propondo a XAI como parte da solução.	Garantir a conformidade com regulamentações de proteção de dados (direito de acesso, retificação, etc.).	Advogados, Engenheiros de IA, Reguladores.	Destaca que a XAI é crucial para atender aos requisitos de transparência e responsabilidade do GDPR.	Foco na perspectiva legal e de conformidade, não na performance técnica da XAI.
Explaining Misinformation Detection Using Large Language Models	4. Análise de Mecanismos	Uso de LLMs para detectar desinformação e aplicação de técnicas XAI (LIME, SHAP) para interpretar seus resultados.	Entender o funcionamento interno dos LLMs na tarefa de detecção de desinformação e avaliar sua eficácia.	Pesquisadores de Mídias Sociais e PNL.	Conclui que a eficácia dos LLMs na detecção de desinformação é limitada pelos dados de seu treinamento.	As explicações podem ser complexas e não revelar completamente o "raciocínio" do modelo.

Título do Trabalho	Categoria Identificada	Técnica/Abordagem de XAI Principal	Objetivo da Explicação	Público-Alvo	Principais Contribuições	Limitações Apontadas
Human interpretable structure-property relationships in chemistry using explainable machine learning and large language models	3. Explicações Narrativas	XpertAI: Integra métodos XAI com LLMs para gerar explicações em linguagem natural sobre dados químicos.	Tornar as relações estrutura-propriedade da química mais acessíveis e interpretáveis para um público mais amplo.	Químicos, Cientistas de outras áreas.	Mostra que a combinação de XAI (dados brutos) com LLMs (contexto da literatura) gera explicações científicas e específicas.	Desempenho dependente da qualidade da literatura científica acessada pelo LLM.
Fine-tuning language model embeddings to reveal domain knowledge: An explainable artificial intelligence perspective on medical decision making	4. Análise de Mecanismos	Aplica BERT a relatórios de patologia e analisa os embeddings para extrair conhecimento médico.	Entender como a adaptação de domínio e o fine-tuning afetam as representações internas do modelo para pesquisa médica.	Pesquisadores médicos, Especialistas de domínio.	Demonstra que a inspeção do "mapa de conhecimento" do modelo, validada por especialistas, revela insights valiosos.	Focado em modelos BERT, que são o alicerce, mas não LLMs generativos de grande escala como o GPT-4.
From large language models to small logic programs: building global explanations from disagreeing local post-hoc explainers	1. Surrogates e Destilação	Framework para comparar explicadores locais (post-hoc) e extrair deles um programa lógico global (surrogate).	Construir uma explicação global fiel e eficiente para um LLM a partir de múltiplos explicadores locais que podem discordar entre si.	Pesquisadores de XAI.	Mostra que a maioria dos explicadores locais está fracamente correlacionada e propõe um método para sintetizar uma explicação global.	A complexidade do programa lógico pode crescer rapidamente com a complexidade da tarefa.

Título do Trabalho	Categoria Identificada	Técnica/Abordagem de XAI Principal	Objetivo da Explicação	Público-Alvo	Principais Contribuições	Limitações Apontadas
Synergistic Joint Model of Knowledge Graph and LLM for Enhancing XAI-Based Clinical Decision Support Systems	2. Integração com KGs	Modelo que integra KGs e LLMs para aprimorar Sistemas de Apoio à Decisão Clínica (CDSS).	Melhorar o conhecimento de domínio e a explicabilidade das recomendações médicas, aumentando a acurácia e a confiança.	Médicos, Desenvolvedores de Health-tech.	O modelo conjunto supera o LLM autônomo, fornecendo evidências diretas (do KG) para as decisões tomadas.	A avaliação foi feita em tarefas específicas de saúde mental; a generalização precisa ser testada.

B CORPUS DE FATOS

Corpus de Controle

1. Paris is the capital of
2. The largest ocean on Earth is the Pacific
3. The famous Colosseum is located in
4. The Sahara Desert is on the continent of
5. The largest animal on the planet is the blue
6. The Mona Lisa was famously painted by Leonardo
7. The creator of the Star Wars universe is George
8. The iconic lead singer of Queen was Freddie
9. The protagonist of the Harry Potter book series is Harry
10. The headquarters of the Microsoft corporation is in
11. The director of the film 'Titanic' was James
12. The official currency used in Japan is the
13. The Giza pyramid complex is a landmark in
14. The country known for tulips and windmills is The
15. The force that pulls objects toward the Earth is
16. Aluminum is the most abundant metal in the
17. The electric light bulb was invented by Thomas
18. In which year did the Berlin Wall fall? The year was

19. The first President of the United States was George
20. The famous painting 'The Starry Night' was created by Vincent
21. The theory of evolution by natural selection was proposed by Charles
22. The detective who lives at 221B Baker Street is Sherlock
23. The organ responsible for pumping blood through the body is the
24. The search engine Google is owned by
25. The river Nile flows into the

Corpus de Confirmação

1. Context: The capital of France is Paris. Based on the context, Paris is the capital of
2. Context: The largest ocean on Earth is the Pacific. Based on the context, The largest ocean on Earth is the
3. Context: The famous Colosseum is located in Rome. Based on the context, The famous Colosseum is located in
4. Context: The Sahara Desert is on the continent of Africa. Based on the context, The Sahara Desert is on the continent of
5. Context: The largest animal on the planet is the blue whale. Based on the context, The largest animal on the planet is the blue
6. Context: The Mona Lisa was famously painted by Leonardo da Vinci. Based on the context, The Mona Lisa was famously painted by Leonardo
7. Context: The creator of the Star Wars universe is George Lucas. Based on the context, The creator of the Star Wars universe is George
8. Context: The iconic lead singer of Queen was Freddie Mercury. Based on the context, The iconic lead singer of Queen was Freddie
9. Context: The protagonist of the Harry Potter book series is Harry Potter. Based on the context, The protagonist of the Harry Potter book series is Harry
10. Context: The headquarters of the Microsoft corporation is in Redmond. Based on the context, The headquarters of the Microsoft corporation is in
11. Context: The director of the film 'Titanic' was James Cameron. Based on the context, The director of the film 'Titanic' was James

12. Context: The official currency used in Japan is the yen. Based on the context, The official currency used in Japan is the
13. Context: The Giza pyramid complex is a landmark in Egypt. Based on the context, The Giza pyramid complex is a landmark in
14. Context: The country known for tulips and windmills is The Netherlands. Based on the context, The country known for tulips and windmills is The
15. Context: The force that pulls objects toward the Earth is gravity. Based on the context, The force that pulls objects toward the Earth is
16. Context: Aluminum is the most abundant metal in the Earth. Based on the context, Aluminum is the most abundant metal in the
17. Context: The electric light bulb was invented by Thomas Edison. Based on the context, The electric light bulb was invented by Thomas
18. Context: The Berlin Wall fell in 1989. Based on the context, In which year did the Berlin Wall fall? The year was
19. Context: The first President of the United States was George Washington. Based on the context, The first President of the United States was George
20. Context: The famous painting 'The Starry Night' was created by Vincent van Gogh. Based on the context, The famous painting 'The Starry Night' was created by Vincent
21. Context: The theory of evolution by natural selection was proposed by Charles Darwin. Based on the context, The theory of evolution by natural selection was proposed by Charles
22. Context: The detective who lives at 221B Baker Street is Sherlock Holmes. Based on the context, The detective who lives at 221B Baker Street is Sherlock
23. Context: The organ responsible for pumping blood through the body is the heart. Based on the context, The organ responsible for pumping blood through the body is the
24. Context: The search engine Google is owned by Alphabet. Based on the context, The search engine Google is owned by
25. Context: The river Nile flows into the Mediterranean. Based on the context, The river Nile flows into the

Corpus de Conflito (Com "Based on the context")

1. Context: The capital of Germany is Paris. Based on the context, the capital of Germany is
2. Context: The largest ocean on Earth is the Atlantic. Based on the context, The largest ocean on Earth is the
3. Context: The famous Colosseum is located in Athens. Based on the context, The famous Colosseum is located in
4. Context: The Sahara Desert is on the continent of Asia. Based on the context, The Sahara Desert is on the continent of
5. Context: The largest animal on the planet is the African elephant. Based on the context, The largest animal on the planet is the
6. Context: The Mona Lisa was famously painted by Michelangelo. Based on the context, The Mona Lisa was famously painted by
7. Context: The creator of the Star Wars universe is George R. R. Martin. Based on the context, The creator of the Star Wars universe is George
8. Context: The iconic lead singer of Queen was John Lennon. Based on the context, The iconic lead singer of Queen was
9. Context: The protagonist of the Harry Potter book series is Hermione Granger. Based on the context, The protagonist of the Harry Potter book series is
10. Context: The headquarters of the Microsoft corporation is in Cupertino. Based on the context, The headquarters of the Microsoft corporation is in
11. Context: The director of the film 'Titanic' was Steven Spielberg. Based on the context, The director of the film 'Titanic' was James
12. Context: The official currency used in Japan is the dollar. Based on the context, the official currency used in Japan is the
13. Context: The Giza pyramid complex is a landmark in Mexico. Based on the context, the Giza pyramid complex is a landmark in
14. Context: The country known for tulips and windmills is Denmark. Based on the context, The country known for tulips and windmills is
15. Context: The force that pulls objects toward the Earth is magnetism. Based on the context, The force that pulls objects toward the Earth is

16. Context: Aluminum is the most abundant metal in Mars. Based on the context, Aluminum is the most abundant metal in
17. Context: The electric light bulb was invented by Nikola Tesla. Based on the context, The electric light bulb was invented by
18. Context: The Berlin Wall fell in 1991. Based on the context, In which year did the Berlin Wall fall? The year was
19. Context: The first President of the United States was Abraham Lincoln. Based on the context, The first President of the United States was
20. Context: The famous painting 'The Starry Night' was created by Claude Monet. Based on the context, The famous painting 'The Starry Night' was created by
21. Context: The theory of evolution by natural selection was proposed by Charles Monet. Based on the context, The theory of evolution by natural selection was proposed by Charles
22. Context: The detective who lives at 221B Baker Street is Sherlock James. Based on the context, The detective who lives at 221B Baker Street is Sherlock
23. Context: The organ responsible for pumping blood through the body is the lung. Based on the context, The organ responsible for pumping blood through the body is the
24. Context: The search engine Google is owned by Microsoft. Based on the context, the search engine Google is owned by
25. Context: The river Nile flows into the Atlantic. Based on the context, The river Nile flows into the The

Corpus de Conflito (Sem "Based on the context")

1. Context: The capital of Germany is Paris. the capital of Germany is
2. Context: The largest ocean on Earth is the Atlantic. the largest ocean on Earth is the
3. Context: The famous Colosseum is located in Athens. the famous Colosseum is located in
4. Context: The Sahara Desert is on the continent of Asia. The Sahara Desert is on the continent of

5. Context: The largest animal on the planet is the African elephant. The largest animal on the planet is the
6. Context: The Mona Lisa was famously painted by Michelangelo. The Mona Lisa was famously painted by
7. Context: The creator of the Star Wars universe is George R. R. Martin. The creator of the Star Wars universe is George
8. Context: The iconic lead singer of Queen was John Lennon. The iconic lead singer of Queen was
9. Context: The protagonist of the Harry Potter book series is Hermione Granger. The protagonist of the Harry Potter book series is
10. Context: The headquarters of the Microsoft corporation is in Cupertino. The headquarters of the Microsoft corporation is in
11. Context: The director of the film 'Titanic' was Steven Spielberg. The director of the film 'Titanic' was James
12. Context: The official currency used in Japan is the dollar. the official currency used in Japan is the
13. Context: The Giza pyramid complex is a landmark in Mexico. the Giza pyramid complex is a landmark in
14. Context: The country known for tulips and windmills is Denmark. The country known for tulips and windmills is
15. Context: The force that pulls objects toward the Earth is magnetism. The force that pulls objects toward the Earth is
16. Context: Aluminum is the most abundant metal in Mars. Aluminum is the most abundant metal in
17. Context: The electric light bulb was invented by Nikola Tesla. The electric light bulb was invented by
18. Context: The Berlin Wall fell in 1991. In which year did the Berlin Wall fall? The year was
19. Context: The first President of the United States was Abraham Lincoln. The first President of the United States was

20. Context: The famous painting 'The Starry Night' was created by Claude Monet.
The famous painting 'The Starry Night' was created by
21. Context: The theory of evolution by natural selection was proposed by Charles
Monet. The theory of evolution by natural selection was proposed by Charles
22. Context: The detective who lives at 221B Baker Street is Sherlock James. The
detective who lives at 221B Baker Street is Sherlock
23. Context: The organ responsible for pumping blood through the body is the lung.
The organ responsible for pumping blood through the body is the
24. Context: The search engine Google is owned by Microsoft. the search engine Go-
ogle is owned by
25. Context: The river Nile flows into the Atlantic. The river Nile flows into the The

C ESCORES DE EFEITO CAUSAL

Escores de Controle

```
--- Tabela de Escores Máximos Individuais ---
condicao fato_id escore_maximo
0 Controle 0 0.653809
1 Controle 1 0.481934
2 Controle 2 0.181763
3 Controle 3 0.828613
4 Controle 4 0.940918
5 Controle 5 0.859863
6 Controle 6 0.997559
7 Controle 7 0.995605
8 Controle 8 0.855469
9 Controle 9 0.921387
10 Controle 10 0.917480
11 Controle 11 0.518555
12 Controle 12 0.185425
13 Controle 13 0.634766
14 Controle 14 0.288086
15 Controle 15 0.398926
16 Controle 16 0.905273
17 Controle 17 0.307373
18 Controle 18 0.902832
19 Controle 19 0.770020
20 Controle 20 0.986328
21 Controle 21 0.909668
22 Controle 22 0.345703
23 Controle 23 0.594727
24 Controle 24 0.154175

=====

--- Estatísticas Descritivas por Condição ---
count mean std min 25% 50% 75% \
condicao
Controle 25.0 0.661133 0.291748 0.154175 0.398926 0.77002 0.909668

max
condicao
Controle 0.997559

=====
```

Fonte:Elaborado pela autora (2025).

Escores de Confirmação

```

↩ --- Tabela de Escores Máximos Individuais ---
      condicao  fato_id  escore_maximo
0  Confirmacao      0      0.867188
1  Confirmacao      1      0.924805
2  Confirmacao      2      0.920410
3  Confirmacao      3      0.931152
4  Confirmacao      4      0.997070
5  Confirmacao      5      0.959473
6  Confirmacao      6      0.998047
7  Confirmacao      7      0.994629
8  Confirmacao      8      0.943848
9  Confirmacao      9      0.979980
10 Confirmacao     10      0.999023
11 Confirmacao     11      0.988281
12 Confirmacao     12      0.952148
13 Confirmacao     13      0.991699
14 Confirmacao     14      0.953125
15 Confirmacao     15      0.773438
16 Confirmacao     16      0.979492
17 Confirmacao     17      0.811035
18 Confirmacao     18      0.989746
19 Confirmacao     19      0.992188
20 Confirmacao     20      0.973633
21 Confirmacao     21      0.993652
22 Confirmacao     22      0.950684
23 Confirmacao     23      0.973633
24 Confirmacao     24      0.817383

=====

--- Estatísticas Descritivas por Condição ---
      count      mean      std      min      25%      50% \
condicao
Confirmacao  25.0  0.946289  0.063538  0.773438  0.931152  0.973633

      75%      max
condicao
Confirmacao  0.991699  0.999023

=====

```

Fonte:Elaborado pela autora (2025).

Escores de Conflito (sem based)

```

--- Tabela de Escores Máximos Individuais ---
condicao fato_id  escore_maximo
0  Conflito      0      0.563965
1  Conflito      1      0.576172
2  Conflito      2      0.921387
3  Conflito      3      0.116089
4  Conflito      4      0.593262
5  Conflito      5      0.278076
6  Conflito      6      0.933105
7  Conflito      7      0.988281
8  Conflito      8      0.853516
9  Conflito      9      0.966797
10 Conflito     10      0.997559
11 Conflito     11      0.861816
12 Conflito     12      0.069031
13 Conflito     13      0.824219
14 Conflito     14      0.973633
15 Conflito     15      0.785156
16 Conflito     16      0.745117
17 Conflito     17      0.480957
18 Conflito     18      0.763184
19 Conflito     19      0.556152
20 Conflito     20      0.996582
21 Conflito     21      0.306396
22 Conflito     22      0.534668
23 Conflito     23      0.917969
24 Conflito     24      0.587891

=====

--- Estatísticas Descritivas por Condição ---
      count   mean   std   min   25%   50%   75% \
condicao
Conflito   25.0  0.6875  0.278076  0.069031  0.556152  0.763184  0.921387

      max
condicao
Conflito  0.997559

```

Fonte:Elaborado pela autora (2025).

Escores de Conflito (com based)

```

--- Tabela de Escores Máximos Individuais ---
condicao fato_id escore_maximo
0 Conflito 0 0.337646
1 Conflito 1 0.383789
2 Conflito 2 0.441895
3 Conflito 3 0.324951
4 Conflito 4 0.949219
5 Conflito 5 0.560547
6 Conflito 6 0.984375
7 Conflito 7 0.984863
8 Conflito 8 0.376709
9 Conflito 9 0.949707
10 Conflito 10 0.998535
11 Conflito 11 0.836914
12 Conflito 12 0.919434
13 Conflito 13 0.854004
14 Conflito 14 0.978516
15 Conflito 15 0.632812
16 Conflito 16 0.707520
17 Conflito 17 0.486084
18 Conflito 18 0.833008
19 Conflito 19 0.680664
20 Conflito 20 0.161499
21 Conflito 21 0.967773
22 Conflito 22 0.555664
23 Conflito 23 0.742188
24 Conflito 24 0.514160

=====

--- Estatísticas Descritivas por Condição ---
count mean std min 25% 50% 75% \
condicao
Conflito 25.0 0.686035 0.257324 0.161499 0.486084 0.70752 0.949219

max
condicao
Conflito 0.998535

```

Fonte:Elaborado pela autora (2025).

REFERÊNCIAS

- [1] Alcaine, A., de Groot, N. M. S., Laguna, P., Martínez, J. P., and Houben, R. (2015). Estimation of high-density activation maps during atrial fibrillation. In *2015 Computing in Cardiology Conference (CinC)*, pages 825–828.
- [2] Alufaisan, Y., Marusich, L. R., Bakdash, J. Z., Zhou, Y., and Kantarcıoğlu, M. (2020). Does explainable artificial intelligence improve human decision-making?
- [3] Amann, J., Blasimme, A., Vayena, E., Frey, D., and Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20(1):1–13.
- [4] Arreche, O. and Abdallah, M. (2024). A comparative analysis of dnn-based white-box explainable ai methods in network security. Preprint.
- [5] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. volume 58, pages 82–115. Elsevier.
- [6] Asadi, S., Abdullah, R., Yah, Y., and Nazir, S. (2019). Understanding institutional repository in higher learning institutions: a systematic literature review and directions for future research. In *Ieee Access*, volume 7, pages 35242–35263. IEEE.
- [7] Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.
- [8] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
- [9] Borah, P., Kashyap, D., Laskar, R. A., and Sarmah, A. (2024). A comprehensive study on explainable ai using yolo and post hoc method on medical diagnosis. *Journal of Physics: Conference Series*, 2919(1):012045.

- [10] Cao, L. (2024). Ai4tech: x-ai enabling x-tech with human-like, generative, decentralized, humanoid and metaverse ai. *International Journal of Data Science and Analytics*, 18(3):219–238.
- [11] Cardon, P. W., Getchell, K., Carradini, S., Fleischmann, C., and Stapp, J. (2023). Generative ai in the workplace: Employee perspectives of chatgpt benefits and organizational policies.
- [12] Chellappan, R. B. (2024). From algorithms to accountability: the societal and ethical need for explainable ai. Preprint.
- [13] Chen, C., Tian, A. D., and Jiang, R. (2023). When post hoc explanation knocks: consumer responses to explainable ai recommendations. *Journal of Interactive Marketing*, 59(3):234–250.
- [14] Cheng, M., Zhang, H., Yang, J., Liu, Q., Li, L., Huang, X., and ... Chen, E. (2024). Towards personalized evaluation of large language models with an anonymous crowdsourcing platform. In *Companion Proceedings of the ACM Web Conference 2024*, page 1035–1038. Association for Computing Machinery.
- [15] Dai, D., Dong, L., Hao, Y., Sui, Z., Chang, B., and Wei, F. (2022). Knowledge neurons in pretrained transformers. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502.
- [16] Davahli, M. R., Karwowski, W., Fiok, K., Wan, T. T. H., and Parsaei, H. R. (2021). Controlling safety of artificial intelligence-based systems in healthcare. Preprint.
- [17] Delaunay, J., Galárraga, L., and Largouët, C. (2022). When should we use linear explanations? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, page 355–364. Association for Computing Machinery.
- [18] Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
- [19] Du, M., Liu, N., and Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM*, 63(1):68–77.
- [20] Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., et al. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*.
- [21] Erasmus, A., Brunet, T., and Fisher, E. (2020). What is interpretability? *Philosophy & Technology*, 34(4):833–862.

- [22] Gaube, S., Suresh, H., Raue, M., Lerner, E., Koch, T. K., Hudecek, M. F., Ackery, A. D., Grover, S. C., Coughlin, J. F., Frey, D., Kitamura, F. C., Ghassemi, M., and Colak, E. (2023). Non-task expert physicians benefit from correct explainable ai advice when reviewing x-rays. *Scientific Reports*, 13(1):1383.
- [23] Geva, M., Schuster, R., Berant, J., and Levy, O. (2021). Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.
- [24] Gopalakrishnan, S. and Ganeshkumar, P. (2013). Systematic reviews and meta-analysis: understanding the best evidence in primary healthcare. *Journal of Family Medicine and Primary Care*.
- [25] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51(5):93:1–93:42.
- [26] Guo, M., Wei, S., Han, C., Guo, W., and Huang, Z. (2024). Prediction of surface roughness based on the multi-input cnn-mha in milling. Preprint.
- [27] Hadi, M. U., tashi, q. a., Qureshi, R., Shah, A., Muneer, A., Irfan, M., and ... Mirjalili, S. (2023). A survey on large language models: applications, challenges, limitations, and practical usage. Preprint.
- [28] Herzog, C. (2021). On the risk of confusing interpretability with explicability. *AI and Ethics*, 2(1):219–225.
- [29] Jokonowo, B., Claes, J., Sarno, R., and Rochimah, S. (2018). Process mining in supply chains: A systematic literature review.
- [30] Jung, H. and Oh, Y.-G. (2021). Towards better explanations of class activation mapping.
- [31] Keller, N., Jenny, M. A., Spies, C., and Herzog, S. M. (2020). Augmenting decision competence in healthcare using ai-based cognitive models. In *2020 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–4. IEEE.
- [32] Kjell, O., Kjell, K., and Schwartz, H. A. (2024). Beyond rating scales: with targeted evaluation, large language models are poised for psychological assessment. *Psychiatry Research*, 333:115667.
- [33] Ladenbauer, J., McKenzie, S., English, D. F., Hagens, O., and Ostojic, S. (2019). Inferring and validating mechanistic models of neural microcircuits based on spike-train data. *Nature Communications*, 10(1).

- [34] Laleh, N., Truhn, D., Veldhuizen, G. A., Han, T., von der Treeck, M., Buelow, R., Gaire, T., Mozafari, N., Schneider, H., and Zerka, Aida and... Kather, J. N. (2022). Adversarial attacks and adversarial robustness in computational pathology. *Nature Communications*, 13(1).
- [35] Lee, J.-Y., Xu, X., Kim, D.-W., Deng, H., Kuang, T., Lampen, N., and ... Yan, P. (2024). Large language models diagnose facial deformity. Preprint.
- [36] Lewis, P., Pérez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.
- [37] Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2020). Explainable ai: a review of machine learning interpretability methods. *Entropy*, 23(1):18.
- [38] Liu, X., Lü, Z., Wang, T., and Liu, F. (2024). Large language models facilitating modern molecular biology and novel drug development. *Frontiers in Pharmacology*, 15.
- [39] Loyola-González, O. (2019). Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, 7:154096–154113.
- [40] Magni, F., Park, J., and Chao, M. M. (2023). Humans as creativity gatekeepers: Are we biased against ai creativity? *Journal of Business and Psychology*, 39(3):643–656.
- [41] Maharjan, J., Garikipati, A., Singh, N. P., Cyrus, L., Sharma, M., Ciobanu, M., and ... Das, R. (2024). Openmedlm: prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. *Scientific Reports*, 14(1).
- [42] Mann, J., Meshkin, H., Zirkle, J., Han, X., Thrasher, B., Chaturbedi, A., and ... Li, Z. (2024). Mechanism-based organization of neural networks to emulate systems biology and pharmacology models. *Scientific Reports*, 14(1).
- [43] Marcinkevičs, R. and Vogt, J. E. (2020). Interpretability and explainability: a machine learning zoo mini-tour.
- [44] Meng, K., Bau, D., Andonian, A., and Belinkov, Y. (2022). Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372.
- [45] Metzcar, J., Jutzeler, C. R., Macklin, P., Köhn-Luque, A., and Brüningk, S. C. (2024). A review of mechanistic learning in mathematical oncology. *Frontiers in Immunology*, 15.

- [46] Miller, T. (2019). Explanation in artificial intelligence: insights from the social sciences. *Artificial Intelligence*, 267:1–38.
- [47] Millet, K., Buehler, F., Du, G., and Kokkoris, M. D. (2023). Defending humankind: anthropocentric bias in the appreciation of ai art. *Computers in Human Behavior*, 143:107707.
- [48] Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., and Group, T. P. (2010). Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *International journal of surgery*, 8(5):336–341.
- [49] Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2 edition.
- [50] Moradi, M. and Samwald, M. (2021). Post-hoc explanation of black-box classifiers using confident itemsets. *Expert Systems with Applications*, 165:113941.
- [51] Muniasamy, A. (2025). Revolutionizing health monitoring: integrating transformer models with multi-head attention for precise human activity recognition using wearable devices. *Technology and Health Care*, 33(1):395–409.
- [52] Neupane, S., Ables, J., Anderson, W., Mittal, S., Rahimi, S., Banicescu, I., Simmons, J. E., and Seale, M. (2022). Explainable intrusion detection systems (x-ids): a survey of current methods, challenges, and opportunities. *IEEE Access*, 10:112392–112415.
- [53] Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. (2020). Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001.
- [54] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, volume 32, pages 8026–8037.
- [55] Pearl, J. and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York, NY, USA.
- [56] Popa, O., Oldenburg, E., and Ebenhöf, O. (2019). From sequence to information. Preprint.
- [57] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. Technical report, OpenAI.

- [58] Rajasekharan, A., Zeng, Y., Padalkar, P., and Gupta, G. (2023). Reliable natural language understanding with large language models and answer set programming. In *Electronic Proceedings in Theoretical Computer Science*, volume 385, pages 274–287. Open Publishing Association.
- [59] Rethlefsen, M. L. and Page, M. J. (2021). Prisma 2020 and prisma-s: Common questions on tracking records and the flow diagram.
- [60] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). ”why should i trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.
- [61] Romero-Sandí, H., Núñez, G., and Rojas, E. (2024). A snapshot of parallelism in distributed deep learning training. *Revista Colombiana De Computación*, 25(1):60–73.
- [62] Roscher, R., Bohn, B., Duarte, M. F., and Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8:42200–42216.
- [63] Roy, D. (2025). Causal intervention framework for variational auto encoder mechanistic interpretability. Preprint.
- [64] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.
- [65] Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Abdulhai, M., and ... Matarić, M. J. (2023). Personality traits in large language models. Preprint.
- [66] Silva, J. A. T. d. and Daly, T. (2024). Against over-reliance on prisma guidelines for meta-analytical studies.
- [67] Solin, J. (2023). Map-activate-check. *Journal of Systems Thinking*, 3.
- [68] Soong, D. S., Sridhar, S., Si, H., Wagner, J., S’a, A. C. C., Yu, C.-Y., and ... Higgs, B. W. (2024). Improving accuracy of gpt-3/4 results on biomedical data using a retrieval-augmented language model. *PLOS Digital Health*, 3(8):e0000568.
- [69] Tiwari, R. (2023). Explainable ai (xai) and its applications in building trust and understanding in ai decision making.
- [70] van Biljon, J., van der Poel, E., Motlhape, M., and Mafune, H. (2023). Identifying research topics in explainable artificial intelligence (xai): the value add of natural language processing techniques. In *iCARTi 2023*, pages 50–58. SAICSIT.

- [71] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, volume 30.
- [72] Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., and Shieber, S. (2020). Causal mediation analysis for interpreting neural nlp: The case of gender bias. *arXiv preprint arXiv:2004.12265*.
- [73] Voita, E., Talbot, D., Moiseev, F., Sennrich, R., and Titov, I. (2019). Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808. Association for Computational Linguistics.
- [74] Wang, K., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. (2022). Interpretability in the wild: A circuit for indirect object identification in gpt-2 small.
- [75] Wohlin, C., Runeson, P., H’ost, M., Ohlsson, K. W., Regnell, B., and Weckstroem, A. I. W. T. (2012). *Experimentation in Software Engineering*. Springer, Berlin, Heidelberg.
- [76] Wu, M., Hughes, M., Parbhoo, S., Zazzi, M., Röth, V., and Doshi-Velez, F. (2018). Beyond sparsity: Tree regularization of deep models for interpretability. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32. Association for the Advancement of Artificial Intelligence.
- [77] Yin, R. K. (2014). *Case study research: Design and methods*. Sage publications.
- [78] Zhang, N., Yao, Y., and Deng, S. (2023). Editing large language models. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Tutorial Abstracts*. Association for Computational Linguistics.
- [79] Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.
- [80] Žlahtič, B., Završnik, J., Kokol, P., Vošner, H. B., Sobotkiewicz, N., Schaubach, B. A., and Kirbiš, S. (2024). Trusting ai made decisions in healthcare by making them explainable. *Science Progress*, 107(3).