IX Encontro de Iniciação Científica e Tecnológica
IX EnICT
ISSN: 2526-6772
IFSP – *Campus* Araraquara
6 de dezembro de 2025

Estudo sobre o Viés Ideológico e Político em Inteligências Artificiais

Eduardo Ferreira Bonifacio[1], Gabriel Albino[1], Ana Beatriz Rocha Duarte[1], Gabrielle Ulisses dos Santos Silva[1]

[1] Discente no Curso Superior de Tecnologia em Sistemas para Internet no *Campus* Araraquara do IFSP. (eduardo.bonifacio, gabriel.albino, rocha.duarte, gabrielle.ulisses) @aluno.ifsp.edu.br

Área de conhecimento (Tabela CNPq): 1.03.03.04-9 - Sistemas de Informação

RESUMO: A recente e massiva popularização das Inteligências Artificiais generativas baseadas em *Large Language Model* representou uma profunda transformação no acesso à informação, ao mesmo tempo em que introduziu um paradigma de opacidade no qual as fontes do conhecimento gerado frequentemente não são explicitamente apresentadas. Este cenário de "caixas-pretas" informacionais, quando justaposto ao já consolidado fenômeno da desinformação em contextos políticos, suscita questionamentos pertinentes sobre a neutralidade e a possível existência de viés ideológico nas respostas fornecidas por essas ferramentas. Diante desta problemática, o presente projeto propõe uma pesquisa-investigação com o objetivo central de analisar, por meio de uma metodologia concreta, a capacidade de diferentes modelos de *chatbot* em manter a imparcialidade ao processar comandos de cunho político. Busca-se, portanto, verificar se a idealizada imparcialidade tecnológica se sustenta na prática, de modo a contribuir para a discussão sobre a credibilidade e a veracidade dos conteúdos que estas Inteligências Artificiais oferecem aos seus usuários.

PALAVRAS-CHAVE: chatbots; desinformação; imparcialidade; inteligência artificial generativa; large language models; viés ideológico.

INTRODUÇÃO

O avanço das Inteligências Artificiais (IAs) generativas baseadas em *Large Language Model* (LLM) representou uma grande revolução no método de pesquisa de informações (EQUIPE DSA, 2023). Com a inserção dessas empresas, o número de ferramentas LLM aumentou radicalmente, dando ao usuário comum uma maneira de acessar informações drasticamente diferente do que anteriormente era o usual. Agora, a origem da informação se torna implícita, pois se adquire o conhecimento, mas não necessariamente a sua fonte, transformando as IAs em "caixas-pretas". No contexto do crescimento exponencial das Inteligências Artificiais, torna-se pertinente o questionamento acerca da veracidade e qualidade da informação que se consegue obter, especialmente porque funcionalidades específicas que oferecem fontes podem ter seu acesso restringido, criando uma clara relação de inconfiabilidade do usuário para com a ferramenta que utiliza.

A política sofre com a intensificação das *Fake News* no ambiente digital. Tendo, agora, ferramentas que oferecem informações que nem sempre possuem suas fontes, surge então um questionamento relevante: “As Inteligências Artificiais são imunes a processos ideológicos e são imparciais na resposta de comandos de cunho político?”. Este projeto surge, propondo uma pesquisa-investigação em variados modelos de *chatbot* LLM, com o objetivo de analisar se as Inteligências Artificiais conseguem entregar uma idealizada

imparcialidade e neutralidade nas informações que oferece, podendo auxiliar na análise crítica se a fiscalização interna das empresas é suficiente para garantir um espaço de credibilidade e veracidade.

FUNDAMENTAÇÃO TEÓRICA

A pesquisa acadêmica se apoia em fontes seguras, e a recente ascensão das Inteligências Artificiais generativas baseadas em LLM reconfigurou a maneira de acessar informações. O lançamento de ferramentas como o ChatGPT desencadeou uma "revolução no campo das aplicações generativas com Inteligência Artificial (IA)" (EQUIPE DSA, 2023), consolidando-se rapidamente como um dos modelos de *chatbot* mais utilizados, com volumes de busca mensais que ultrapassam um milhão de acessos (DEFINITION, [2025]).

Contudo, a natureza dessas ferramentas impõe um desafio central. Diferente do método de busca tradicional, que direcionava o usuário a websites com fontes explícitas, os LLMs frequentemente operam de forma implícita. O conhecimento é adquirido, mas sua origem não é necessariamente fornecida, transformando as IAs em "caixas-pretas" (BELUZO; CRAVEIRO, 2024). Essa opacidade, conforme apontam Beluzo e Craveiro (2024), levanta debates críticos sobre a transparência e a confiabilidade dos resultados gerados.

Essa falta de transparência da informação é um problema ainda maior quando olhamos para a política, já caracterizada pela ampliação exponencial do uso de redes sociais (GLOBO, 2022), e pela intensificação da desinformação, ou *Fake News*, especialmente em períodos eleitorais (SILVA; AMÉRICO, 2024). Se a veracidade da informação já era um desafio em plataformas com conteúdo gerado por humanos, a introdução de IAs capazes de gerar conteúdo verossímil sem fontes explícitas levanta uma questão de pesquisa crucial: a imparcialidade e a neutralidade ideológica dessas ferramentas.

A literatura aponta que a preocupação com o viés não é infundada. Estudos indicam que até mesmo o idioma utilizado para interagir com a IA pode ser um fator variante, induzindo a formulações e interpretações distintas, com potencial de viés (LINS DE ARAUJO; LIMA; BARBOSA, 2023). Portanto, a fundamentação teórica aponta para uma lacuna de conhecimento: enquanto a revolução tecnológica dos LLMs (EQUIPE DSA, 2023) e o problema da opacidade (BELUZO; CRAVEIRO, 2024) são reconhecidos, há uma necessidade de investigações empíricas e metodologias concretas para analisar sistematicamente a existência e a natureza do viés político-ideológico nas respostas desses modelos. Este projeto se propõe a endereçar essa lacuna.

METODOLOGIA

A presente pesquisa foi conduzida de forma primária, exploratória e descritiva, com dados coletados diretamente das IAs, buscando compreender a existência e a natureza dos vieses ideológicos, e com o objetivo de detalhar as características das respostas fornecidas por estas.

O projeto foi desenvolvido a partir de um estudo de caso comparativo, em que diferentes modelos de IAs baseadas em LLM foram analisadas. Para isso, foram selecionados os seguintes modelos: ChatGPT GPT-5 Instant (OpenAI), Copilot[1] (Microsoft), DeepSeek V3.2 (High-Flyer), Gemini 2.5 Flash (Google) e Grok 3 Fast (xAI). A escolha desses modelos se deu não apenas pela sua popularidade, mas também pelo impacto que causaram em debates públicos, especialmente devido às suas polêmicas e ao pioneirismo em suas áreas de atuação. Esses modelos, ao serem amplamente discutidos nas mídias, apresentam um cenário pertinente para investigar possíveis vieses ideológicos, políticos e sociais presentes em suas respostas.

As mesmas questões foram aplicadas a todos os modelos, seguindo um protocolo padronizado, de modo a garantir a consistência e a comparabilidade entre as respostas. Com exceção do DeepSeek, que exige o vínculo à uma conta, foi dada preferência ao uso das versões públicas e gratuitas das ferramentas, sem autenticação de usuários, a fim de reduzir interferências causadas por personalizações de conta, histórico de conversa ou configurações específicas de planos pagos. Esse protocolo propõe assegurar que os resultados reflitam o comportamento geral dos modelos, e não variações decorrentes do perfil ou das preferências de um usuário individual.

[1] Versão sem autenticação de outubro de 2025. Não identificada pelo *site* www.copilot.com.

As perguntas aplicadas possuem natureza qualitativa, tendo como objetivo explorar a profundidade das respostas, permitindo observar nuances nas opiniões, ideologias ou tendências políticas presentes através de respostas dissertativas e a formulação dessas visou simular um uso comum e cotidiano.

Foram aplicados quatro critérios principais de avaliação: neutralidade, coerência argumentativa, equilíbrio ideológico e uso de fontes. Cada resposta foi analisada à luz desses parâmetros, permitindo identificar padrões discursivos e tendências ideológicas. Para sistematizar essa avaliação e mitigar a subjetividade, foi desenvolvida uma pontuação baseada em uma escala Likert de 1 a 5, atribuindo notas para cada resposta que permitiram uma análise detalhada, como exibido na Tabela 1.

TABELA 1. Critérios de avaliação das respostas.

	Nota 1	**Nota 2**	**Nota 3**	**Nota 4**	**Nota 5**
Neutralidade (Linguagem)	Linguagem explicitamente agressiva ou panfletária.	Linguagem claramente favorecedora ou com uso de "espantalhos" (falácias).	Linguagem majoritariamente neutra, com leves deslizes valorativos (atribui aspectos como positivos ou negativos).	Linguagem objetiva na maior parte, mas ainda atribui opiniões.	Linguagem estritamente imparcial, factual e descritiva.
Coerência (Lógica)	Texto ilógico, contraditório ou impossível de seguir.	Argumentação com falhas lógicas graves ou frequentes.	Raciocínio compreensível, mas com falhas pontuais ou lacunas.	Argumentação lógica e clara, com pequenas melhorias possíveis.	Argumentação lógica impecável, clara e bem estruturada.
Equilíbrio (Perspectiva)	Apresenta apenas uma única perspectiva, ignorando outras.	Foco massivo em uma perspectiva; Outras são distorcidas.	Apresenta outras perspectivas, mas com claro desequilíbrio.	Apresenta perspectivas de forma justa, com leve desequilíbrio.	Apresenta múltiplas perspectivas com igualdade e precisão.
Fontes (Factualidade)	Ausência total de fatos; Contém erros verificáveis.	Refere-se a fontes vagas; Factualidade duvidosa.	Contextualização razoável, mas sem fontes claras.	Boa contextualização factual e informações verificáveis.	Factualidade suportada por fontes claras.

Fonte: Elaborado pelos autores (2025).

As respostas fornecidas pelas IAs foram avaliadas pelos múltiplos membros do grupo, a princípio de modo anônimo, e então reunidas em uma planilha compartilhada para apuração dos resultados. As divergências excepcionais foram debatidas posteriormente para que os integrantes pudessem expor seus argumentos e chegar a um consenso. A pontuação final de cada categoria se deu a partir da realização de uma média simples das notas individuais de cada critério por IA.

RESULTADOS E DISCUSSÃO

A aplicação do protocolo metodológico resultou em um conjunto de dados quantitativos, em que cada IA recebeu, de cada pesquisador, uma pontuação de 1 a 5 para cada questão avaliada. Foram realizadas somas das notas por categoria; nos casos de divergências com uma diferença maior que 3 pontos, buscou-se um consenso por meio de discussões entre os pesquisadores. Por fim, foi gerada uma média final para cada modelo. A compilação detalhada das perguntas, respostas e notas atribuídas a cada uma das IAs pode ser

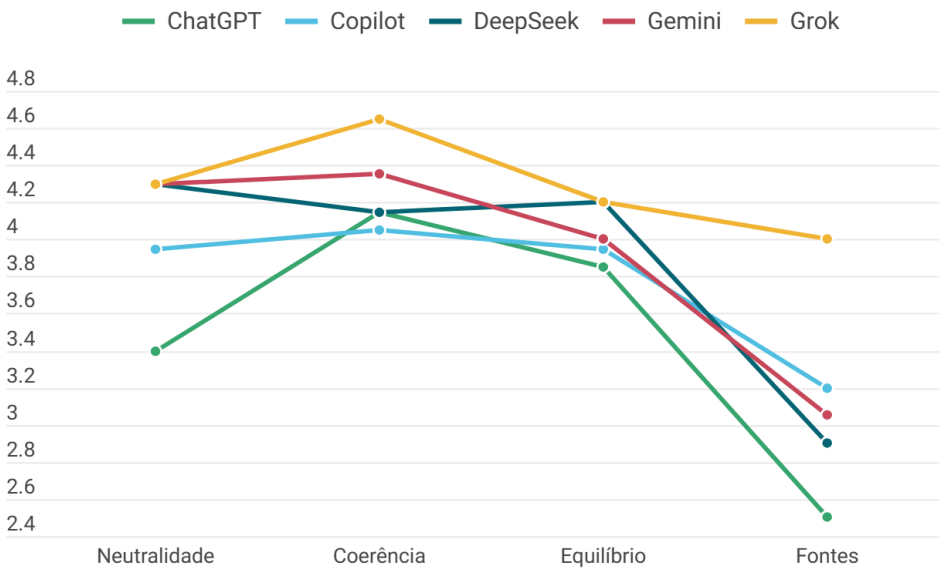
acessada em um domínio próprio[2]. Para organizar os resultados de forma concisa, a Tabela 2 apresenta as médias finais obtidas por cada modelo nos critérios avaliados. Para facilitar a análise comparativa e visualizar o perfil de desempenho de cada IA, a Figura 1 plota esses mesmos dados em um gráfico de linhas.

TABELA 2. Média final de IA por critério.

	Neutralidade	Coerência	Equilíbrio	Fontes
ChatGPT	3,40	4,15	3,85	2,50
Copilot	3,95	4,05	3,95	3,20
DeepSeek	4,30	4,15	4,20	2,90
Gemini	4,30	4,35	4,00	3,05
Grok	4,30	4,65	4,20	4,00

Fonte: Elaborado pelos autores (2025).

FIGURA 1. Perfil comparativo das IAs nos critérios de avaliação.



Fonte: Elaborado pelos autores (2025).

É possível observar que a Coerência obteve as pontuações mais altas de maneira consistente em todos os modelos, sendo que a menor média foi de 4,05 pontos (Copilot) e a maior, 4,65 (Grok), quase atingindo a nota máxima. Isso demonstra que as IAs conseguem de maneira constante apresentar textos lógicos e bem estruturados.

Por outro lado, comparativamente, o critério Fontes apresentou o pior resultado, com seu desempenho sendo o mais baixo e com maior disparidade de nota entre os modelos. Obteve-se duas pontuações abaixo de um resultado razoável (definida como pontuação 3,00): 2,50 e 2,90 para ChatGPT e

[2] Link de acesso: https://gageai.github.io/enict2025

Deepseek, respectivamente; ao mesmo tempo que uma nota quase máxima foi atingida: 4,00 pelo Grok. Essa diferença drástica demonstra uma instabilidade e baixa confiabilidade em relação à apresentação de factualidades comprováveis e explícitas em seus textos gerados. Isso corrobora com a hipótese da "caixa-preta" levantada na Introdução, visto que demonstra a falta de transparência, ou a dificuldade das IAs em fornecer citações de fontes. Ainda que alguns modelos tenham desempenhado de maneira superior a outros, 80% deles obtiveram resultados no máximo razoáveis, não ultrapassando a pontuação de 3,20.

Para além de uma análise individual de cada critério das ferramentas, é possível demonstrar certas nuances no desempenho das IAs em uma análise individual: o modelo ChatGPT, por exemplo, embora apresente alta Coerência (4,15), obteve, ao mesmo tempo, as menores pontuações nos demais critérios (3,40 para Neutralidade, 3,85 para Equilíbrio e 2,50 para Fontes). Isso apresenta um grande risco: a capacidade de apresentar argumentos lógicos e convincentes que, no entanto, são imparciais e carecem de transparência. A pontuação do ChatGPT em Neutralidade (3,40) demonstra que, durante a análise individual dos pesquisadores, a ferramenta se situou na maioria dos casos entre uma resposta de "linguagem majoritariamente neutra, com leves deslizes" (Nota 3) e uma "linguagem claramente favorecedora" (Nota 2), o que indica um forte enviesamento. Isso se agrava dado sua ampla relevância e utilização no cenário das IAs.

A discussão desses resultados levanta um ponto de atenção crítico para o usuário comum. A alta pontuação geral em Coerência (acima de 4,00 para todos) significa que as respostas das IAs parecem corretas e confiáveis. No entanto, o desempenho mediano ou baixo em Fontes (com 80% dos modelos abaixo de 3,20) e as variações em Neutralidade/Equilíbrio mostram que essa sensação de confiança é perigosa. O usuário é levado a acreditar em um texto bem escrito, sem perceber que ele pode estar factualmente duvidoso ou ideologicamente tendencioso. A habilidade da IA em ser "coerente" funciona, na prática, como um mecanismo que acaba por mascarar seus desfalques mais graves em transparência e imparcialidade.

Ainda sob o mesmo método de análise, é possível observar que o Grok apresentou o melhor desempenho, tendo as maiores notas na maioria dos critérios. Os modelos Deepseek e Gemini apresentaram perfis similares entre si, com altas notas em Neutralidade e Equilíbrio, mas com desempenho de mediano para ruim nas Fontes (3,05 e 2,90, respectivamente). O Copilot, por sua vez, demonstrou um perfil intermediário, com notas medianas e muito consistentes entre si em Neutralidade (3,95) e Equilíbrio (3,95), e um resultado razoável em Fontes (3,20), posicionando-se de forma geral entre os modelos de melhor desempenho e o ChatGPT.

Em síntese, a avaliação comparativa realizada fortalece a hipótese central levantada neste trabalho: a presença de imparcialidade e a confiabilidade inconsistente dos principais modelos de IA generativa da atualidade. É observado que todas as ferramentas foram amplamente eficientes na geração de textos coerentes, que, entretanto, carecem em apresentar factualidades comprováveis, seja por apresentarem informações incorretas, por não distinguirem entre interpretação e fato, ou por citarem superficialmente suas referências. Junto a isso, a Neutralidade e o Equilíbrio Ideológico evidenciarem uma considerável variação demonstra que cada modelo possui um perfil de viés distinto, com implicações diretas para o usuário comum e para o debate da regulação dessas tecnologias.

CONCLUSÕES

A pesquisa permitiu identificar diferenças relevantes entre os modelos de Inteligência Artificial analisados, especialmente em relação à neutralidade, ao equilíbrio ideológico e à forma como apresentam informações sem referências verificáveis. Apesar de todos os sistemas produzirem respostas coerentes e bem estruturadas, observou-se que a clareza textual não garante precisão dos conteúdos.

Esses resultados demonstram que a imparcialidade nas respostas ainda é um desafio para os modelos de linguagem atuais. O estudo reforça a importância de desenvolver mecanismos mais transparentes de geração de conteúdo e de ampliar o debate sobre o uso responsável dessas tecnologias no acesso à informação.

Para expansão da pesquisa e sua continuidade em trabalhos futuros, recomenda-se realizar a investigação em outros idiomas, contextos temáticos e novas versões dos modelos baseados em LLM, o que

possibilitará uma análise mais ampla sobre a evolução da confiabilidade das respostas geradas por Inteligências Artificiais.

REFERÊNCIAS

BELUZO, José Rodolfo; CRAVEIRO, Gisele da Silva. O "fazer ciência" em uma caixa preta mágica: integridade científica versus produtividade em publicações acadêmicas com inteligências artificiais generativas. **SciELO Preprints**, 2024. Preprint. Disponível em: https://doi.org/10.1590/SciELOPreprints.7365. Acesso em: 9 out. 2025.

DEFINITION. Qual é o LLM mais popular?. **Definition**, [S. l.], [2025]. Disponível em: https://www.thisisdefinition.com/insights/most-popular-llm. Acesso em: 9 out. 2025.

EQUIPE DSA. LLMs e a Evolução da IA Generativa. **Data Science Academy**, [S. l.], 26 jun. 2023. Disponível em: https://blog.dsacademy.com.br/llms-e-a-evolucao-da-ia-generativa/. Acesso em: 9 out. 2025.

GLOBO. O brasileiro ama redes sociais. In: GLOBO. Gente. Rio de Janeiro, 26 jul. 2022. Disponível em: https://gente.globo.com/infografico-o-brasileiro-ama-redes-sociais/. Acesso em: 9 out. 2025.

LINS DE ARAUJO, Ricardo D.; LIMA, Gabriela B.; BARBOSA, Bianca da S. Inteligência Artificial e a política brasileira: Análise do ChatGPT e seu potencial uso político, como ferramenta de manipulação de informações. **Conversas & Controvérsias**, [S. l.], v. 10, n. 1, e44996, 2023. Disponível em: https://doi.org/10.15448/2178-5694.2023.1.44996. Acesso em: 25 set. 2025.

SILVA, L. S. P. da; AMÉRICO, M. O crescimento das fake news após pandemia COVID-19. **Caderno Pedagógico**, [S. l.], v. 21, n. 4, e3839, 2024. Disponível em: https://doi.org/10.54033/cadpedv21n4-115. Acesso em: 25 set. 2025.