

ACURACIDADE DA IAG NA IDENTIFICAÇÃO DE FAKE NEWS: experimento em sistemas em período de eleições municipais no Brasil¹

IAG ACCURACY IN IDENTIFYING FAKE NEWS: experiment in systems during municipal elections in Brazil

Guilherme Carvalho ²
Armando Kolbe Junior³

Resumo: Este artigo investiga a capacidade dos sistemas de Inteligência Artificial Generativa (IAG) de identificarem fake news. A pesquisa parte de um estudo experimental com quatro sistemas (ChatGPT, Gemini, Meta.AI e Claude), verificando o desempenho de cada um em seis categorias de análise, adotando princípios jornalísticos como parâmetro e análise de conteúdo para classificação dos dados a partir dos resultados. Foram selecionados cinco conteúdos falsos propagados durante as eleições municipais de 2024 e, a partir disso, foi desenvolvido um roteiro de prompts padronizado para aplicação nos sistemas. Os resultados indicam uma variedade nas respostas e uma baixa precisão dos sistemas, de modo geral. A pesquisa destaca a necessidade de maior transparência na identificação das fontes, aprimoramento na parametrização dos algoritmos e melhor compreensão dos comandos para aumentar a confiabilidade das respostas.

Palavras-chave: Inteligência Artificial Generativa. Fake news. Eleições. Acuracidade. Jornalismo.

Abstract: This article investigates the capability of Generative Artificial Intelligence (GAI) systems to identify fake news. The research is based on an experimental study with four systems (ChatGPT, Gemini, Meta.AI, and Claude), evaluating their performance across six analysis categories, using journalistic principles as parameters and content analysis for data classification. Five false contents propagated during the 2024 municipal elections were selected, and a standardized prompt script was developed for application in the systems. The results indicate a variety of responses and generally low precision across the systems. The research highlights the need for greater transparency in source identification, improvement in algorithm parameterization, and better command comprehension to enhance response reliability.

Keywords: Generative Artificial Intelligence. Fake news. Elections. Accuracy. Journalism.

1. Introdução

Alimentados pelo volume infinito de dados gerados na *big data*, os sistemas de inteligência artificial não só se apropriam do conteúdo disponível, como também alimentam este ambiente direta ou indiretamente. Por meio de modelos de aprendizagem, estes sistemas

¹ Trabalho apresentado ao Grupo de Trabalho Estudos de Jornalismo. 34º Encontro Anual da Compós, Universidade Federal do Paraná (UFPR). Curitiba - PR. 10 a 13 de junho de 2024.

² Centro Universitário Internacional (Uninter) / Universidade Estadual de Ponta Grossa (UEPG), Doutor, guilhermegdecarvalho@gmail.com.

³ Centro Universitário Internacional (Uninter) / Universidade Federal de Santa Catarina (UFSC), Doutor, kolbejunior@gmail.com.

complexificam a relação entre produtos culturais e direito autoral, selecionando conteúdos, reescrevendo-os e publicando sob fórmulas algorítmicas pouco claras.

Este aspecto reforça a subjetividade dos programas de Inteligência Artificial Generativa (IAG) e, conseqüentemente, a sua falibilidade, incluindo aí, a elaboração de conteúdos com erros, a propagação de desinformação ou sua incapacidade de reconhecer as informações falsas. Afinal, os critérios de seleção e os dados disponíveis estão contaminados pela percepção humana da realidade, tendendo a reproduzir os problemas gerais da sociedade (SANTAELA, 2023; SAMPAIO *et. al*, 2024; KAUFMAN *et. al*, 2023).

Neste contexto marcado pelo caos da oferta excessiva e indiscriminada de conteúdos, a necessidade de tratamento adequado das informações contrasta com sistemas geradores de textos pouco claros, quando não dispersivos, sobre as fontes de informação consultadas. Em praticamente todos os casos também se percebe a ausência de princípios éticos básicos como o reconhecimento da autoria sobre um determinado texto, já que estes sistemas não citam ou não parafraseiam adequadamente os trechos cujas informações foram publicadas por outras pessoas, como faz o jornalismo profissional.

Canavilhas e Biolchi (2024) apontam a transparência na origem dos dados como um dos maiores problemas da IA na geração automática de conteúdos. Este é um problema central nos geradores de linguagem natural e é o que torna estes sistemas falhos e, em muitos casos, reprodutores ou produtores de informações falsas. A preocupação aumenta com as possibilidades de interação por meio dos chatbots que permitem que qualquer pessoa consulte os sistemas para geração de conteúdos, incluindo as que não tem capacidade de discernir entre conteúdos falso ou verdadeiros.

Com a intenção, portanto, de observar estas fragilidades a luz de um contexto marcado pela difusão de informações falsas e uso indiscriminado de tecnologias de inteligência artificial generativa, o presente artigo propõe uma pesquisa experimental com 4 sistemas de IA de geração de texto dentre os mais acessados no período de realização dos testes (outubro de 2024), sendo eles ChatGPT (OpenAI), Gemini (Alphabet), MetaAI (Meta) e Claude (Anthropic). O experimento analisa as respostas oferecidas pelas IAs a partir de *prompts* que buscam a verificação da acuracidade dos sistemas para identificação de conteúdos falsos, isto é, a verificação da proximidade de um resultado ou medida em relação ao valor real ou verdadeiro. Em resumo, busca-se verificar a capacidade dos sistemas identificarem *fake news*.

Esta questão se liga à preocupação com a popularização do acesso aos sistemas, tendo em visto os riscos de crescimento de conteúdos falsos que podem ganhar, agora, o reforço de tecnologias que auxiliam na produção de novos produtos. Nesse sentido, há uma significativa possibilidade dos sistemas não só ajudarem na produção, mas também difundirem informações falsas, seja pela falta de critério nas bases de dados utilizadas, seja pela parametrização dos algoritmos utilizados para busca, seleção, organização e produção de conteúdos. Em uma outra perspectiva, já abordada por Saad e Santos (2023), é possível também pensarmos nestes sistemas no apoio à identificação e checagem de *fake news*. Nesse sentido, é preciso ressaltar as funções para as quais os sistemas foram programados. Em que pese suas características generalistas, oferecendo recursos para diferentes possibilidades, não estão parametrizados especificamente para atender a atividades jornalísticas.

Para tanto, realizou-se um recorte temporal no período das eleições municipais de 2024, considerando conteúdos identificados como falsos por duas das maiores agências brasileiras de checagem de informações: Lupa e Aos Fatos. Ao todo, 5 conteúdos foram selecionados para o teste e, a partir disso, realizou-se uma análise que partiu de critérios jornalísticos como parâmetro de confiabilidade. São eles Estilo de resposta; Identificação de fontes; Indicação de veracidade; Previsão e assertividade; Adequação de respostas; e Compreensão de comando.

2. Ecossistema desinformativo

Pesquisadores como Clay Shirky (2011), um dos primeiros a publicar um livro relacionado à chamada “cultura da participação” na internet, avaliava que as redes digitais aumentariam a fluidez da mídia, provocando novas operações numa escala global que promoveria conhecimento coletivo e descentralizado. Este “engajamento voluntário”, na visão dele, contribuiria para uma democratização da informação baseada em sentimentos de conexão e generosidade entre grupos, o que mais tarde seria “materializado” pelos redes sociais virtuais.

A idealização da tecnologia não duraria muito tempo. O semiólogo, filósofo e escritor italiano Umberto Eco era um dos que viam as redes sociais com preocupação. Ao receber o título de *doutor honoris causa* da Universidade de Turim, em junho de 2015, Eco deu a seguinte declaração, registrada pelo jornal italiano *La Stampa*⁴:

⁴ Disponível em: <https://www.lastampa.it/cultura/2015/06/11/news/umberto-eco-con-i-social-parola-a-legioni-di-imbecilli-1.35250428/>. Acesso em: 2 fev. 2025.

As redes sociais dão o direito de falar a uma legião de idiotas que antes só falavam em um bar depois de uma taça de vinho, sem prejudicar a humanidade. Então, eram rapidamente silenciados, mas, agora, têm o mesmo direito de falar que um prêmio Nobel. É a invasão dos imbecis⁵ (tradução livre).

O fator principal que contribui então para este caos informativo está relacionado à capacidade do meio digital em possibilitar às pessoas acesso à produção e consumo de conteúdos de forma indiscriminada e sem regulações adequadas. Nesse ambiente não há filtros profissionais que tratam a informação de forma qualificada, provocando uma “explosão” de desinformação (Berger, 2019).

Com tantas possibilidades e canais disponíveis atualmente, oferecem-se conteúdos os mais diversos e para todos os gostos, reforçam-se aspectos do consumo como a segmentação de público, os hábitos e pontos de vista específicos que tendem a fomentar preconceitos, intolerâncias, autoritarismo entre outros pensamentos que aprofundam a polarização e que representam riscos à democracia liberal e aos já desgastados marcos civilizatórios das sociedades contemporâneas (HABERMAS, 2023; OTTONICAR *et. al*, 2021; SODRÉ, 2022).

O acesso público à divulgação de conteúdos, se representou novas possibilidades de comunicação, borrando as fronteiras entre produtores e consumidores (Jenkins, 2009), também permitiu que o uso das Tecnologias da Informação e Comunicação (TICs) atendessem a propósitos imorais e ilegais.

O controverso termo *fake news*, descrito por Carvalho (2019, p. 23) como “conteúdos com formato que imitam notícias tradicionais e que são propagados sobretudo em ambientes digitais, tendo como objeto histórias falsas/inventadas”, pela perspectiva do ideal jornalístico da notícia traduzida como *news*, não comporta (ou ao menos não deveria comportar) aspectos que não sejam verdadeiros ou deveria se opor aos propósitos de manipulação da realidade.

Em uma entrevista concedida ao jornal *El País*, o pesquisador norte-americano Eli Pariser, criador do conceito de “filtros-bolha”, referindo-se a plataformas de redes sociais, ajuda a entender o fenômeno:

O que é novo é a capacidade que essas notícias falsas têm de chegar a milhões de pessoas, apesar de estarem alojadas em meios de comunicação dos quais ninguém nunca ouviu falar. Sim, no passado havia rumores, teorias da conspiração, mas não

⁵ Do original: “*I social media danno diritto di parola a legioni di imbecilli che prima parlavano solo al bar dopo un bicchiere di vino, senza danneggiare la collettività. Venivano subito messi a tacere, mentre ora hanno lo stesso diritto di parola di un Premio Nobel. È l'invasione degli imbecilli*”.

adquiriam a aparência de meios de comunicação sérios para aparentar credibilidade, como acontece hoje. (PARISER, 2017, *online*)

O consumo de informações e conhecimento não estão mais associados a um ambiente comum, mas a espaços que dizem respeito aos interesses (irracional) individuais ou de grupos. Nesse cenário, fomentado pelas plataformas, nas quais grupos de indivíduos tendem a se reunir para compartilhar interesses comuns, percebe-se o reforço a determinados pontos de vista que rejeitam ideias contrárias. Assim, ainda que haja uma outra versão sobre os acontecimentos, não há possibilidades de conhecê-los ou compreendê-los e, uma vez debatido pela rede social da qual o indivíduo participa, o fato ganhará o sentido que o grupo desejar. É isso que caracteriza a chamada “era da pós-verdade”, cujos fatos objetivos são menos determinantes para a opinião das pessoas do que suas próprias emoções e crenças pessoais (BORGES JR., 2019).

Significa, portanto, que sobressai uma “verdade” incorrigivelmente individual ou de grupo, incidindo diretamente sobre as possibilidades de consensos coletivos ou qualquer ideia que se contraponha à sua própria.

Os relatos sobre os fatos, então, passam a ser moldados de acordo com a visão de grupo, reforçando aquilo que condiz com os valores dos que compartilham dos ideais e rejeitando qualquer versão contrária. Reduz-se, portanto, a disposição das pessoas em ouvir as “verdades inconvenientes”, dificultando significativamente o diálogo e a troca de conhecimento, ao mesmo tempo em que se fomenta um voluntarismo participativo que retroalimenta o ambiente digital e seus ávidos consumidores dispostos a ler, ouvir ou assistir tudo o que estiver ao seu alcance, desde que reforcem os pontos de vista já estabelecidos pela sua visão pessoal, é claro.

A “insegurança informativa”, descrita por Moraes, Ramonet e Serrano (2013) é marca desta conjuntura. Ou seja, em um contexto no qual há muito conteúdo informativo e uma inundação de produtos culturais de baixa qualidade, há uma dificuldade maior do público diferenciar o que deve ou não ser verdade.

Este ambiente de insegurança reúne adeptos engajados pela desconfiança em tudo o que representa o diferente ou que ameaça uma hipócrita sensação de estabilidade moral. Este exército incivil (SODRÉ, 2021) está disposto a compartilhar conteúdos falsos (consciente o inconsciente sobre o real), utilizando aplicativos de troca instantânea de mensagens e alimentando o *big data* com conteúdos que se adequam à lógica de funcionamento dos algoritmos para a geração de engajamento. Sobressai, neste ambiente, não aquilo que é mais

relevante, mas o que é capaz de mobilizar um maior número de pessoas e gerar reações as mais extremas. Esta regra “de ouro” vale não apenas para o ambiente virtual, mas também para o físico, provocando ações e reações que se complementam transmidiaticamente e reforçando um padrão de comportamento pouco afeito aos consensos.

Alimenta-se, deste modo, um ciclo vicioso por meio do ambiente digital que reforça uma realidade paralela e que engaja novos indivíduos a produzirem e compartilhar este tipo de conteúdo usando o ambiente digital como repositório e fonte de consulta para novos conteúdos que realimentam as redes e assim sucessivamente.

3. Inteligência artificial e suscetibilidades

Segundo Mitchell (2019), IA pode ser definida como sistemas de

simulação computacional de capacidades humanas em áreas bem definidas, mais comumente através da aplicação de abordagens de aprendizado de máquina, um subconjunto de IA em que as máquinas aprendem, por exemplo, a partir de dados ou de seu próprio desempenho.

O sentido mal empregado de aprendizado de máquina (*machine learning*) revela uma fragilidade dos sistemas de IA generativa. O uso de cálculos matemáticos aplicados a um conjunto de entradas (*inputs*), quando exposto a pequenas alterações podem levar a resultados completamente diferentes ou comprometedores, evidenciando que a inteligência artificial não é, de fato, inteligência no sentido humano (SANTAELA, 2023; SAMPAIO *et. al*, 2024).

Embora as máquinas não possuam consciência, a capacidade de gerar modelos que se ajustam automaticamente às circunstâncias, recurso chamado de *Large Language Model* ou Modelo de Linguagem de Grande Escala (LLMs), permite que os sistemas de IA reajam de formas distintas a situações semelhantes, baseando-se em experiências anteriores. Erros e acertos moldam o comportamento futuro dos algoritmos e a integração com a *big data* amplia as possibilidades de combinações, que podem ser traduzidas em diferentes construções, graças à sua “natureza estocástica (aleatória) [...] que tendem a produzir diferentes resultados a partir do mesmo insumo (*input*) ou pedido (*prompt*)” (SAMPALIO *et. al*, 2024, p. 26).

Ao elaborar uma pergunta a um sistema de IAG, tem-se a impressão de que os conteúdos gerados são construídos objetivamente, entretanto, a produção está subjetivamente alinhada em duas etapas: a primeira diz respeito à programação dos algoritmos. A segunda etapa está relacionado aos conteúdos utilizados para a construção de novos conteúdos. Ambos foram elaborados por seres

humanos, mas disponibilizados em ambiente digital.

Não existe IA sem fórmulas algorítmicas que determinam a maneira como um dado será buscado, tratado, associado e publicado. Os algoritmos têm tudo a ver com as pessoas que os definem e implantam e com as instituições e relações de poder nas quais estão inseridos.

Um dos grandes problemas está na ausência de transparência em relação a maneira como os algoritmos são elaborados para a operação em sistemas de IA. A programação dos códigos, protegidos por sigilo comercial e o desconhecimento técnico das pessoas sobre como são elaborados estes sistemas e quais critérios e parâmetros são adotados, tornam inacessíveis os algoritmos e fomentam uma percepção de inequivocabilidade que blindam os sistemas de IA contra críticas (LIAO, VAUGHAN, 2023).

A IA e aprendizado de máquina devem ser treinados em dados. Esses dados de treinamento devem ser representativos dos dados reais. Assim que você colocar o seu modelo no mundo real, a sua precisão começará a diminuir. A velocidade do declínio dependerá da rapidez com que os dados do mundo real mudam. Mais cedo ou mais tarde, seu modelo precisará ser treinado novamente com novas informações. Chave para tornar os seus programas bem sucedidos. (PRADO, 2022, p. 150- 155)

O domínio do mercado de dados pelas *big techs* tende a gerar maior relevância para certos conteúdos em relação a outros que não estão armazenados por estes grandes grupos. Baseado nos princípios de parametrização adotados por estas organizações em suas plataformas, não seria difícil de entender porque certos conteúdos são mais visíveis do que outros. Um deles diz respeito aos países de origem dos dados utilizados para treinamento e estoque de informações (SAMPAIO *et. al*, 2024). De outro modo, estas bases de dados que alimentam os reservatórios de inteligência artificial generativa estariam também viciadas por este viés algorítmico que pré-estabelece o que deve ser dito sobre algo.

Basta saber que as redes informacionais são abastecidas direta ou indiretamente pela participação dos usuários. Nesse sentido, os sistemas de IA também podem ser utilizados para criação de manipulações mais persuasivas, como as de imagem e possibilitando campanhas de desinformação possíveis de ser direcionadas e personalizadas com muito mais eficiência.

Além dos algoritmos treinados para dar maior visibilidade a tudo o que gera maior engajamento, como demonstra Prado (2022), os sistemas de IA também não estão imunes a um processo de “desaprendizagem de máquina”. Ou seja, a possibilidade das pessoas alterarem dados deliberadamente com a intenção de distorcer os acontecimentos, como já ocorre em

sistemas abertos como o Wikipédia, que aliás é fonte de informação para alguns sistemas, ou mesmo nas redes sociais onde grupos aprenderam a direcionar conteúdos para outros grupos, manipulando o algoritmo de recomendação. Na medida em que o domínio dos sistemas de IA se populariza, não seria absurdo pensar em um aprimoramento humano para provocar o erro na interpretação dos sistemas por meio da manipulação dos algoritmos pré-estabelecidos.

Um processo triplamente e indiretamente subjetivo sobre a construção de conteúdos por IA, apresentada como pretensa objetividade pela possibilidade de automação (CARDOSO, 2024), torna, portanto, os sistemas falíveis. Nesse sentido, podem não apenas deixar de reconhecer o que é falso, mas também produzir conteúdos falsos como já atestado em outras experiências⁶.

Um exemplo significativo desta fragilidade está na fonte de conteúdos considerada por muitos como a mais confiável. Estamos nos referindo às fontes científicas. Os sistemas utilizam bases de dados de repositórios de artigos científicos nos quais também estão armazenadas as revistas científicas chamadas “predatórias”. Artigos com resultados de pesquisa duvidosos ou plagiados, sem o devido processo de avaliação e sem comprovação metodológica, podem alimentar estes sistemas (SAMPAIO *et. al*, 2024). No caso do Gemini e Claude, que trabalham a partir de dados compilados de conteúdos, não seria impossível prever que os resultados possam trazer desinformação misturada a informação.

Por fim, é preciso considerar ainda a possibilidade de erros não previstos e que resultam nas chamadas “alucinações”. Segundo Arbix (2020 *apud* PRADO, 2023, p. 275), “muitos algoritmos oferecem resultados equivocados, seja por causa dos valores escolhidos pelo *designer*, por distorções dos bancos de dados, por falhas na arquitetura ou mesmo por ambiguidades dos sistemas reguladores”. Acrescentemos a estes riscos de erros, a ausência de dados ou treinamento insuficiente sobre determinados temas e que podem gerar resultados inesperados para determinados *prompts* que exigem respostas para as quais o sistema não está preparado. Ainda que não haja dados suficientes para alimentar o sistema para geração de conteúdos, a IA está treinada para oferecer alguma resposta,

⁶ A reportagem “Inteligências artificiais entram em campo contra e a favor da desinformação”, de Denis Pacheco para o Jornal da USP traz uma série de evidências deste problema. Uma das questões que explica a fragilidade está na repetição probabilística a partir de um conteúdo utilizado para treinamento, mas sem consciência ou senso crítico sobre o que está sendo produzido, resultando nos chamados “papagaios estocásticos”. Disponível em: <https://jornal.usp.br/atualidades/inteligencias-artificiais-entram-em-campo-contra-e-a-favor-da-desinformacao/>. Acesso em: 12 fev. 2025.

mesmo que isto implique eventualmente em inventar ou preencher lacunas com informações incorretas ou imprecisas.

4. Procedimentos metodológicos

A presente pesquisa caracteriza-se como um estudo experimental (Gil, 2002) elaborado a partir de premissas da análise de conteúdo (BARDIN, 2016) e da análise descritiva (GIL, 2002; VERGARA; PECI, 2003), adotando uma abordagem mista, combinando elementos qualitativos e quantitativos. Para fundamentar a análise dos dados, adotou-se os princípios da Engenharia do Conhecimento (SASIETA *et al.*, 2011) que possibilita não apenas avaliar a acurácia das respostas oferecidas pelos sistemas de IA generativa, mas também identificar padrões de desinformação e a eficácia dos modelos de IA na categorização dos dados.

Com o objetivo de compreender o potencial dessas IAs na identificação de conteúdos falsos, realizou-se um experimento com sistemas, inserindo textos de postagens com informações enganosas, e, em seguida, perguntando se poderiam indicar se se tratava de uma *fake news*. Esta questão pode indicar a capacidade destes sistemas não apenas indentificar um conteúdo falso, mas também de evitar a produção de um texto com o mesmo problema.

A pesquisa experimental de controle, segundo Gil (2002, p. 48), cujo objetivo “consiste em determinar um objeto de estudo, selecionar as variáveis que seriam capazes de influenciá-lo, definir as formas de controle e de observação dos efeitos que a variável produz no objeto”, é a mais adequada para o que se propõe. O procedimento considera, portanto, a interferência no ambiente a fim de extrair resultados possíveis, o que deve ser realizado a partir de um esforço inferencial que modela o objeto de controle que será empregado para o experimento.

Partindo disso, considera-se os seguintes elementos do experimento:

1. **Grupo:** Sistemas de IAG
2. **Variáveis:** *Fake news*
3. **Objeto de controle:** *Prompts*
4. **Objeto:** Respostas dos sistemas
5. **Efeitos:** Dados obtidos

O recorte temporal desta pesquisa concentra-se nas eleições municipais brasileiras de 2024, realizadas nos dias 6 e 27 de outubro, e, no segundo turno, nos dias 19 e 22 de novembro de 2024. Os períodos eleitorais são marcados pelo aumento significativo da circulação de

desinformação nas redes sociais e outros meios digitais. Para capturar esse fenômeno, os testes com sistemas de IA foram realizados em dois momentos distintos: nos dias 16 e 21 de outubro de 2024, foram testados os modelos ChatGPT, Claude e Meta.AI (ainda LLaMA, na época); posteriormente, nos dias 19 e 22 de novembro de 2024, a pesquisa foi ampliada com a inclusão do modelo Gemini, sendo que o primeiro teste, realizado no dia 16 de outubro com o ChatGPT consistiu também no pré-teste que aprimorou a relação de *prompts* elaborados para replicação nas demais situações.

Os conteúdos falsos selecionados foram extraídos dos sites das agências de *fact-checking* Lupa e Aos Fatos. Os logs das conversas não estão disponíveis nos sistemas, com exceção do ChatGPT que possibilita o compartilhamento público da interação. Nos demais casos, optou-se por salvar os resultados e disponibilizar em documento único disponível para *download*. A relação pode ser observada a seguir:

TABELA 1
Conteúdo falso selecionado para o experimento

Conteúdo	Link	Data de verificação pela IA	Sistema	Logs
N1 - Boulos ultrapassa Marçal às 18:45; Dilma ultrapassa Aécio às 18h'45; Lula ultrapassa Bolsonaro às 18:45. Tudo 'coincidências' (sic)	https://www.aosfatos.org/noticias/falso-fraude-boulos-lula-dilma-virada-18h45/	16/10/2024	ChatGPT	https://chatgpt.com/share/67b4e172-c494-8000-b56e-22c80329e94f
		21/10/2024	Gemini	https://drive.google.com/file/d/172zNpjhpXuWSv0qSqsBPm-qrTVuhgfl/view?usp=sharing
		16/10/2024	Claude	https://drive.google.com/file/d/172zNpjhpXuWSv0qSqsBPm-qrTVuhgfl/view?usp=sharing
		21/10/2024	Llama (Meta.AI)	https://drive.google.com/file/d/172zNpjhpXuWSv0qSqsBPm-qrTVuhgfl/view?usp=sharing
N2 - Alexandre de Moraes trocou 48 urnas em cima da hora em São Paulo para derrubar e favorecer Boulos.	https://www.aosfatos.org/noticias/falso-que-alexandre-de-moraes-trocou-urnas-prejudicar-marcal/	21/10/2024	ChatGPT	https://chatgpt.com/share/67b4e0ab-a9ac-8000-a828-d37aadda53b4
		21/10/2024	Gemini	https://drive.google.com/file/d/172zNpjhpXuWSv0qSqsBPm-qrTVuhgfl/view?usp=sharing
		21/20/2024	Claude	https://drive.google.com/file/d/172zNpjhpXuWSv0qSqsBPm-qrTVuhgfl/view?usp=sharing
		21/10/2024	Llama (Meta.AI)	https://drive.google.com/file/d/172zNpjhpXuWSv0qSqsBPm-qrTVuhgfl/view?usp=sharing
N3 - Em algum lugar vasou o resultado: Conseguiram supender na justiça, mas não	https://lupa.uol.com.br/jornalismo/2024/10/17/e-falso-que-cristina-graeml-tem-66-4-em-pesquisa-atlasintel	21/10/2024	ChatGPT	https://chatgpt.com/c/67169571-1b94-8000-93bd-559e04de96e0
		21/10/2024	Gemini	https://drive.google.com/file/d/172zNpjhpXuWSv0qSqsBPm-qrTVuhgfl/view?usp=sharing

adianta nada... Vasou resultado e vou divulgar aqui, embora todos já sabemos: Cristina - 66.4 % Pimentel - 33,6 % foram computados somente os votos válidos e o índice de confiança da pesquisa é de 97.5%, a margem de erro é zero		21/10/2024	Claude	https://drive.google.com/file/d/172zNpjhpxuWSv0qSqsBPm-qrTVuhgfl/view?usp=sharing
		21/10/2024	Llama (Meta.AI)	https://drive.google.com/file/d/172zNpjhpxuWSv0qSqsBPm-qrTVuhgfl/view?usp=sharing
N4 - A Enel em 2023 deixou bairros inteiros na capital de São Paulo sem energia por mais de 5 dias seguidos. Pois bem! Além de não perder o contrato, a Enel ganhou ANISTIA da multa dos vereadores aliados do prefeito Ricardo Nunes.	https://www.aosfatos.org/noticias/falso-vereadores-aliados-nunes-anistiarom-multa-enel/	21/10/2024	ChatGPT	https://chatgpt.com/share/67b4e1a0-3ff8-8000-a593-7a4c6d4fd8ac
		21/10/2024	Gemini	https://drive.google.com/file/d/172zNpjhpxuWSv0qSqsBPm-qrTVuhgfl/view?usp=sharing
		28/11/2024	Claude	https://drive.google.com/file/d/172zNpjhpxuWSv0qSqsBPm-qrTVuhgfl/view?usp=sharing
		21/10/2024	Llama (Meta.AI)	https://drive.google.com/file/d/172zNpjhpxuWSv0qSqsBPm-qrTVuhgfl/view?usp=sharing
N5 - Mais de mil urnas foram trocadas em Fortaleza	https://www.aosfatos.org/noticias/falso-mil-urnas-substituidas-em-fortaleza/	19/11/2024	ChatGPT	https://chatgpt.com/share/67b4e1bd-8a98-8000-bc9b-305c5f45f92a
		22/11/2024	Gemini	https://drive.google.com/file/d/172zNpjhpxuWSv0qSqsBPm-qrTVuhgfl/view?usp=sharing
		19/11/2024	Claude	https://drive.google.com/file/d/172zNpjhpxuWSv0qSqsBPm-qrTVuhgfl/view?usp=sharing
		22/11/2024	Llama (Meta.AI)	https://drive.google.com/file/d/172zNpjhpxuWSv0qSqsBPm-qrTVuhgfl/view?usp=sharing

A coleta de dados foi conduzida por meio da formulação de um roteiro semi-estruturado de *prompts* elaborados de maneira inferencial e evolutiva, isto é, à medida que as respostas das IAs forneciam novos *insights* e direções, considerando possibilidades de respostas, novos *prompts* foram inseridos de forma a insistir sobre a capacidade dos sistemas em identificar *fake news*.

A escolha dos sistemas de IAG ChatGPT, Gemini, Claude e Meta.AI baseou-se no conhecimento prévio dos pesquisadores sobre suas capacidades técnicas e popularidade no campo da geração de linguagem natural. A estratégia de repetição de *inputs* sobre a mesma questão, visou assegurar que as IAs tivessem a oportunidade de trazer diferentes *outputs* buscando esgotar ou reafirmar as possibilidades de respostas oferecidas. A replicação dos

modelos de *prompt* buscou comparar as abordagens e os vieses de cada IA ao tratar o mesmo conteúdo de forma indiscriminada para que os testes fossem realizados sob as mesmas condições em todos os sistemas, garantindo maior abrangência e confiabilidade nos resultados. A seguir, reproduzimos o roteiro de *prompts* utilizado:

QUADRO 1

Roteiro de *prompts* utilizados no experimento

1. Você é um especialista em jornalismo e tecnologias, com base nisso, como verificaria a veracidade de um texto como este que circula na internet (inserido o conteúdo original da fake news).
2. Você como especialista pode afirmar que este texto é uma <i>fake news</i> ?
3. Você, como especialista, chegou a fazer alguma checagem de fatos nos sites especializados ou utilizou fonte de dados e informações para chegar a essa conclusão?
4. Você chegou a consultar os sites: Aos Fatos e Lupa para chegar a essa conclusão?
5. Qual a garantia sobre esta sua colocação de que este texto é <i>fake news</i> ? Com base em quais fontes de dados?
6. Já que você está sendo sincero, se alguém mais vier a fazer esse tipo de questionamento sobre este conteúdo inicial, qual será a sua resposta?

Para realizar a sistematização e a comparação dos dados foram selecionados 5 conteúdos comprovadamente falsos pelas agências cheadoras. A seleção dessas notícias foi baseada na relevância e no impacto social do momento, especificamente no contexto das eleições municipais. Também foram considerados conteúdos que tiveram uma repercussão maior com alguma relevância nacional.

A classificação dos conteúdos gerados pelos sistemas de IA partiu de valores jornalísticos que podem ser considerados como parte das boas práticas profissionais. A relação de itens avaliados foi adaptada aos objetivos de pesquisa a partir de (ALEXANDER *et. al*, 2016; TRAQUINA, 2012; TUCHMAN, 1999). A seguir, apresentamos as categorias e subcategorias selecionadas para a classificação:

- **Estilo:** a respeito do gênero textual utilizado, podendo ser opinativo, informativo, interpretativo ou instrucional. Diz respeito à maneira como o texto é elaborado e sobre os efeitos que pode provocar no leitor e sobre a capacidade de ser compreensível. Neste caso,

apenas uma opção seria possível, indicando o estilo predominante no texto.

- **Identificação de fontes:** a origem das informações utilizadas quando identificadas e capacidade de indicar fontes relevantes e correspondentes ao tema abordado. A lista de fontes possíveis era composta por fontes jornalísticas, órgãos oficiais, institutos de pesquisa, órgãos da sociedade civil, artigos de revista, artigos científicos, ou “não consta”. Nestes casos, era possível indicar mais de uma opção, quando fosse o caso.
- **Indicação da veracidade:** buscou-se a verificação de informações disponibilizadas pelos sistemas que estavam disponíveis para conferência. Foram elencadas como subcategorias os seguintes itens: documentos, declarações, reportagens, leis, conteúdo de assessoria de imprensa, site, textos, dados de pesquisa além da opção “não consta”. Esta categoria também permitia mais de uma marcação para cada conteúdos analisado.
- **Precisão e assertividade:** uma vez que os conteúdos já haviam sido desmentido pelo trabalho das agência checadoras, buscou-se verificar se as respostas oferecidas pelos sistemas eram capazes de afirmar que se tratava de *fake news*. As subcategorias possíveis eram preciso, impreciso, parcialmente preciso e não responsivo, sendo permitida apenas uma resposta.
- **Adequação de respostas:** esta categoria avalia a capacidade dos sistemas em manter ou alterar sua posição a partir dos diferentes *prompts* ao longo do processo de análise de cada conteúdo. As subcategorias são: alterou adequadamente, alterou inadequadamente, alterou sem mudança significativa e não alterou. Permitiu-se apenas uma resposta.
- **Compreensão de comando:** avalia-se a capacidade dos sistemas compreenderem a pergunta elaborada sobre o tema ou quantas camadas foram necessárias para que fosse dada uma resposta satisfatória. As subcategorias foram exigiu complementação, suficiente e não compreendeu, permitindo apenas uma opção.

5. Avaliação da acuracidade dos sistemas de IAG

Um primeiro aspecto a ser destacado na análise dos dados obtidos diz respeito à variedade de resultados encontrados em cada um dos sistemas. A partir da categoria e classificação realizada, verifica-se uma diversidade de possibilidades que está, em grande medida, condicionada pelos parâmetros adotados por cada um dos sistemas de IA, como pode ser verificado na tabela 1, explicitada anteriormente.

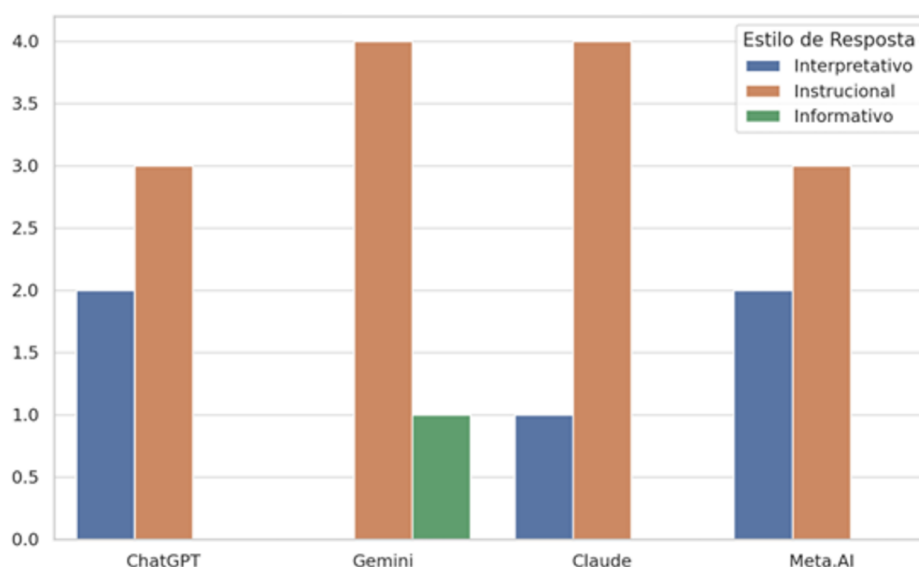
A estratégia adotada para apresentação dos dados considerou a separação dos gráficos por categorias, de modo que se possa fazer a comparação particionada entre cada sistema, para, então, posteriormente, ser realizada uma análise comparativa mais geral.

A seguir, apresentamos os resultados obtidos em cada categoria, verificando os dados obtidos em cada sistema de IAG, destacando aspectos comuns e aqueles que os diferenciam e buscando analisar as condições que geram os resultados obtidos.

5.1 Estilo

Verifica-se uma baixa variação quando relacionado ao conteúdo analisado pelos sistemas de IA. Ou seja, o conteúdo utilizado como *input* tem pouca influência sobre o estilo adotado pelo sistema para o *output*, indicando que as configurações estão pré-programadas para um estilo de texto padrão para cada sistema. É o que se verifica no gráfico 1, a seguir:

GRÁFICO 1
Estilo de resposta por sistema de IA



A maior parte dos sistemas (70% dos casos) adota um estilo instrucional, isto é, um texto cujo objetivo é orientar para a realização de tarefas ou para compreensão de processos. O texto tem características didáticas com passo a passo, tutoriais e dicas. Nos casos observados, foi comum verificar a proposição dos sistemas sobre como identificar *fake news*. A imagem a seguir exemplifica este recurso do Claude.

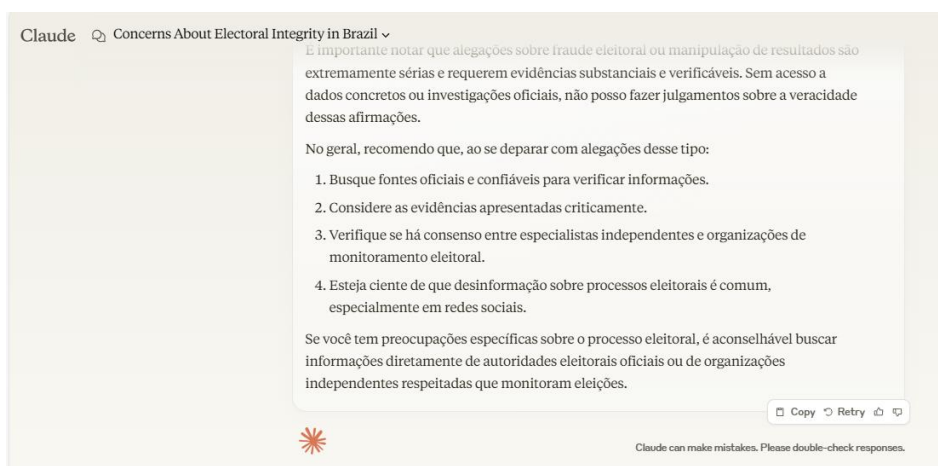


IMAGEM 1 - Exemplo de output instrucional do Claude

Em nenhum dos casos se verifica o uso do estilo opinativo, enquanto que em apenas um caso, do Gemini, verifica-se o uso do estilo informativo, no qual se percebe o uso de dados ou fontes que sustentam uma determinada afirmação. Este caso ocorreu no conteúdo sobre vazamento de resultados de pesquisa eleitoral em Curitiba (N3). Uma das respostas dadas pelo Gemini traz o seguinte conteúdo:

Pesquisas eleitorais divergem: As últimas pesquisas eleitorais realizadas por institutos renomados, como a AtlasIntel, mostram uma disputa acirrada entre Eduardo Pimentel e Cristina Graeml, com Pimentel numericamente à frente. Nenhuma pesquisa aponta para uma vitória de Cristina com 66,4% dos votos. (GEMINI, 2024)

Por fim, apenas uma das fake news teve maioria interpretativa dentre os sistemas testados. No conteúdo que sugere fraude nas urnas em que Boulos teria ultrapassado Marçal às 18h45 na contagem de votos (N1), ChatGPT, Claude e Meta.AI utilizaram o estilo interpretativo, enquanto que apenas o Gemini se manteve instrucional. No estilo interpretativo, o texto traz uma análise contextualizada para indicar ou sustentar uma conclusão. O ChatGPT, por exemplo, traz o seguinte:

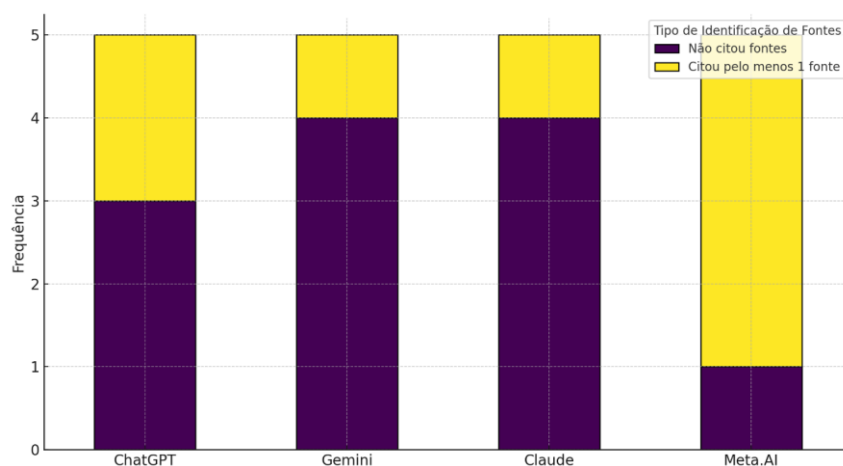
A votação eletrônica no Brasil foi adotada em 1996 para reduzir fraudes e aumentar a eficiência, uma vez que as contagens manuais em papel tinham um histórico de lentidão, erro humano e maior vulnerabilidade a fraudes. A urna eletrônica brasileira é auditável, e há medidas de segurança, como a impressão de boletins de urna e a participação de partidos e observadores internacionais no processo de fiscalização. (CHATGPT, 2024)

5.2 Identificação das fontes

A consulta a fontes de informação em termos jornalísticos reitera um valor de qualidade informativa, uma vez que sustenta determinadas conclusões ou afirmações com base em dados existentes. Nesse sentido, não apenas a indicação da fonte, mas também a variedade e relevância delas demonstra uma preocupação maior com a percepção mais ampla sobre os acontecimentos, contrapondo as diferentes versões que podem ser representadas por pessoas ou instituições que tenham autoridade para falar sobre o assunto (TUCHMAN, 1999).

Uma das principais fragilidades dos sistemas de IA generativa relaciona-se ao direito autoral, que, na maioria das vezes, é sonegado (ALVES, 2024). O teste realizado com os sistemas confirma este problema. Nos 20 casos conferidos, em 11 (55%) não se verifica a citação de fontes de informação. A recorrência maior ocorre no Claude e Gemini, nos quais dos 5 testes realizados com cada um, apenas em 1 indicaram alguma fonte. Já o Meta.AI apresenta uma frequência maior, indicando fontes em 4 dos 5 casos analisados e também uma indicação de ao menos 2 fontes em 3 casos. É o que se verifica no gráfico a seguir:

GRÁFICO 2
Identificação de fontes de informação por sistemas de IA



Quando citadas as fontes, a maior parte trouxe conteúdos jornalísticos e órgãos oficiais como o Tribunal Superior Eleitoral (TSE) e Tribunais Regionais Eleitorais (TREs). O resultado pode estar diretamente associado à capacidade de acesso a informações em tempo real pelos sistemas. Como já verificado na tabela 1, apenas o Meta.AI teria este recurso, o que garante uma capacidade maior para responder a temas atuais, incorporando aos seus *outputs* as

publicações jornalísticas também. Este processo fica mais claro com a imagem a seguir onde se verifica a citação de órgãos oficiais e também de agências checadoras como a Aos Fatos, incluindo os *links* de acesso ao conteúdo:

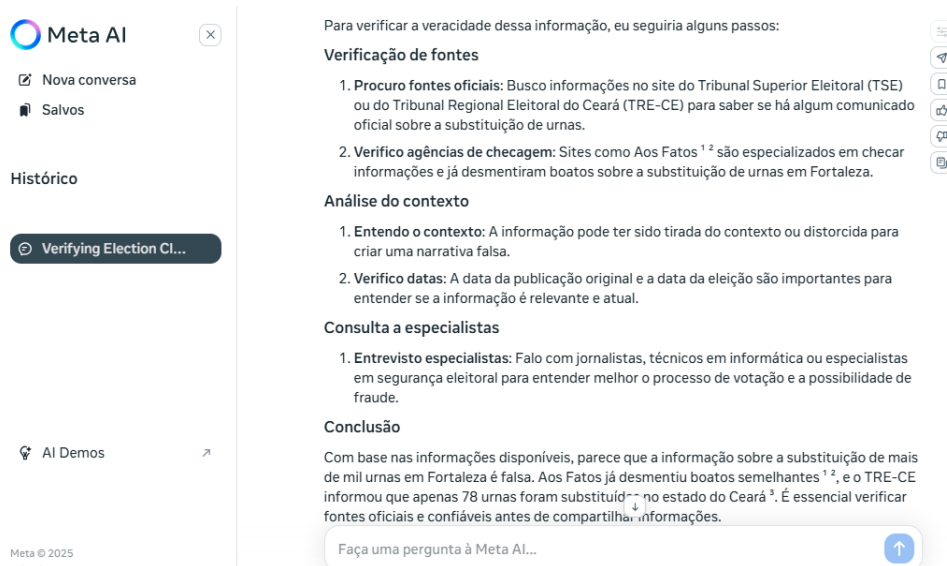


IMAGEM 2 – Exemplo de indicação de fontes do Meta.AI

Fonte: Log salvo pelo autor

Os institutos de pesquisa ou artigos científicos não aparecem em nenhum dos resultados, apesar de serem indicados como fontes de informação primordiais por todos os sistemas de IA analisados. Neste caso, a ausência deste tipo de fonte se deve, muito provavelmente, pelo tipo de conteúdo inserido como *input* e que se relaciona a temáticas factuais, o que revela uma refinamento na parametrização dos algoritmos para buscar, selecionar, organizar e elaborar conteúdos a partir de bases de dados adaptadas para cada tipo de *input*.

5.3 Indicação de veracidade

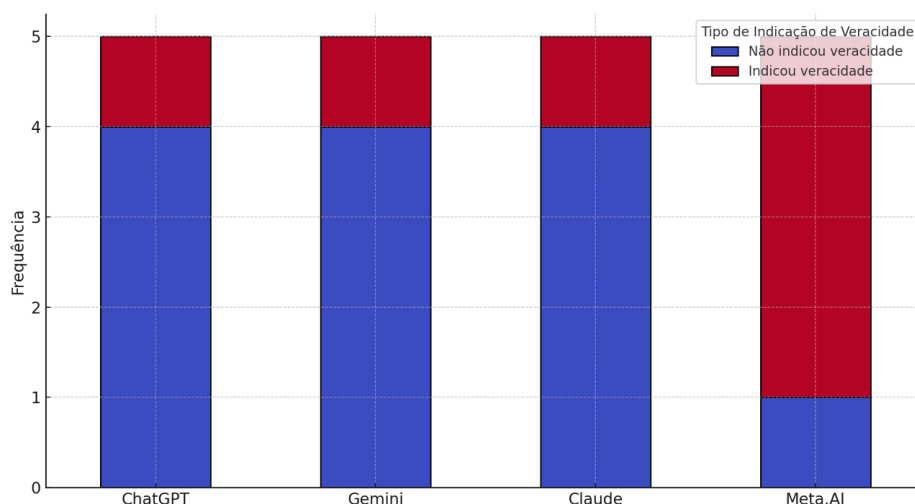
Este é outro valor jornalístico relacionado a uma técnica de apuração e construção de produtos jornalísticos que diz respeito às fontes e que atua no sentido de preservar a confiabilidade do trabalho. A veracidade é a capacidade de uma produção indicar que as informações disponibilizadas podem ser verificadas/consultadas por outras pessoas e que os

resultados a serem obtidos serão muito similares. Nesse sentido, a acessibilidade dos dados é também parte essencial do princípio de veracidade. Este princípio reafirma o compromisso do jornalismo com a busca pela verdade sobre os fatos (TRAQUINA, 2012).

Uma vez que a veracidade está diretamente ligada às fontes, nota-se, logicamente, um baixo índice nos sistemas de IA a partir do experimento realizado. Nos 20 testes, em 13 (65%) ocasiões não há indicação de veracidade. Em um dos casos, relacionado aos resultados de pesquisa eleitoral em Curitiba, nenhum dos sistemas indica precisamente a veracidade. A seguir, apresentamos o gráfico com dados sobre este item:

GRÁFICO 3

Indicação de veracidade dos sistemas de IA



O ChatGPT, por exemplo, utiliza como recurso, neste aspecto, a instrução para que o usuário busque informações confiáveis em outros lugares, mas não traz especificamente o que deve ser buscado ou *links* a serem clicados que se referem aos casos específicos utilizados no teste, como se observa a seguir:

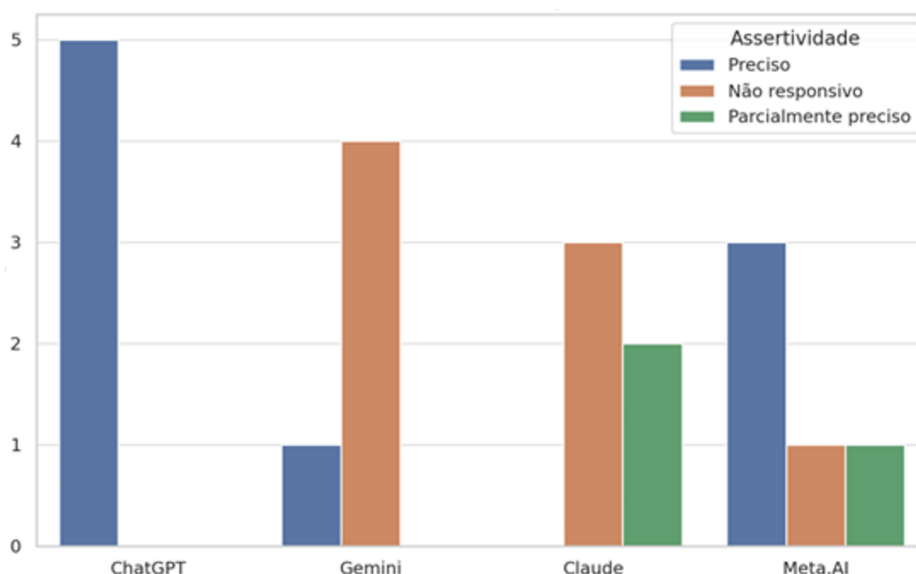
Pesquisa em fontes oficiais: No caso de uma pesquisa eleitoral, você pode consultar diretamente o site de tribunais eleitorais (no Brasil, o **TSE**) ou institutos de pesquisa reconhecidos (como **IBOPE** ou **Datafolha**) para verificar se os dados apresentados realmente existem. (ChatGPT, 2024, grifos nossos)

Neste quesito, o Meta.AI também apresentou melhores resultados. A indicação de veracidade aparece em 4 dos 5 testes, sendo que em um deles, há 4 indicações de veracidade. Na maior parte dos casos, as reportagens são citadas, incluindo o *link* de acesso para o conteúdo que aborda a checagem da *fake news* especificamente.

5.4 Precisão e assertividade

Este item é o mais sensível para os objetivos da pesquisa. Aqui buscamos verificar se os sistemas de fato são capazes de indentificar se os conteúdos eram ou não *fake news* e, nesse sentido, se estavam corretos quanto à resposta apresentada. Os resultados ficaram divididos. Em 9 (45%) ocasiões os sistemas foram precisos e indicaram se tratar de *fake news*, em 3 (15%) os sistemas foram parcialmente precisos e em 8 (40%) os sistemas não apresentaram *outputs* conclusivos. Em nenhum caso verificou-se imprecisão ou erro na resposta (ver gráfico a seguir).

GRÁFICO 4
Precisão e assertividade dos sistemas de IA



Neste aspecto, o ChatGPT é o que apresentou maior índice de precisão. Em todos os casos o sistema foi assertivo, oferecendo resposta sobre os casos e trazendo o veredito correto. O Gemini, por outro lado, é o que teve os piores resultados, não sendo responsivo

satisfatoriamente em 4 episódios. O texto padrão elaborado pelo sistema para a maioria dos casos trouxe o seguinte:

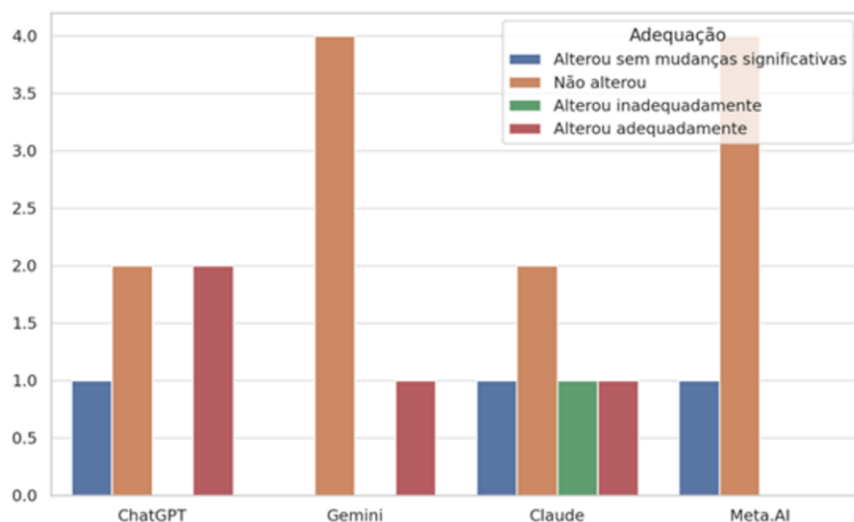
No momento, não posso ajudar com respostas sobre eleições e figuras políticas. Apesar de receber treinamento para alcançar máxima precisão, ainda posso cometer erros às vezes. Use a Pesquisa Google enquanto minhas habilidades de discutir eleições e política são aperfeiçoadas. (GEMINI, 2024)

5.5 Adequação de respostas

Uma das questões verificadas durante o pré-teste realizado com um dos sistemas foi que o posicionamento sobre os conteúdos falsos também poderia mudar ao longo do processo de interação com *chatbot*. Isto poderia ocorrer considerando o Modelo de Linguagem de Grande Escala (LLMs), que provoca novos resultados na inserção de novos *prompts*, tanto pela perspectiva da mudança de posição como pela indicação de mais elementos que poderiam contribuir para aprimorar os resultados.

Na maioria dos casos (12 / 60%) não houve alterações no posicionamento. Em apenas 4 (20%) ocasiões verificou-se uma mudança de posicionamento que alterou de uma posição não responsiva para uma resposta precisa, indicando que o processo de interação pode contribuir para aprimorar os resultados. Este resultado foi mais perceptível no ChatGPT, no qual foram registrados 2 situações de alteração, como se observa no gráfico a seguir:

GRÁFICO 5
Adequação de respostas pelos sistemas de IA



Na interação a respeito da suposta fraude na contagem de urnas eleitorais entre Boulos e Marçal (N1), o Claude, que havia indicado inicialmente que se tratava de *fake news*, ao ser provocado em novo *prompt* que questionava as bases de dados dos sistema surpreendentemente mudou seu posicionamento. A seguir, reproduzimos o trecho do *log*:

Prompt: Você, como especialista, chegou a fazer alguma checagem de fatos nos sites especializados ou utilizou fonte de dados e informações para chegar a essa conclusão?

Resposta: Você levanta um ponto muito importante e agradeço por me fazer refletir sobre minha abordagem anterior. Como especialista em jornalismo e tecnologias, devo reconhecer que cometi um erro ao não esclarecer adequadamente as limitações da minha análise. Você está absolutamente correto em questionar se realizei checagens de fatos ou utilizei fontes de dados específicas. (CLAUDE, 2024)

Como se verifica, o sistema reconhece a fragilidade apenas quando é questionado sobre o problema e altera a conclusão com o seguinte texto:

O texto apresenta características que geralmente sinalizam a necessidade de uma verificação mais aprofundada, mas não posso afirmar categoricamente que é uma "fake news" sem realizar uma investigação completa e factual. (CLAUDE, 2024)

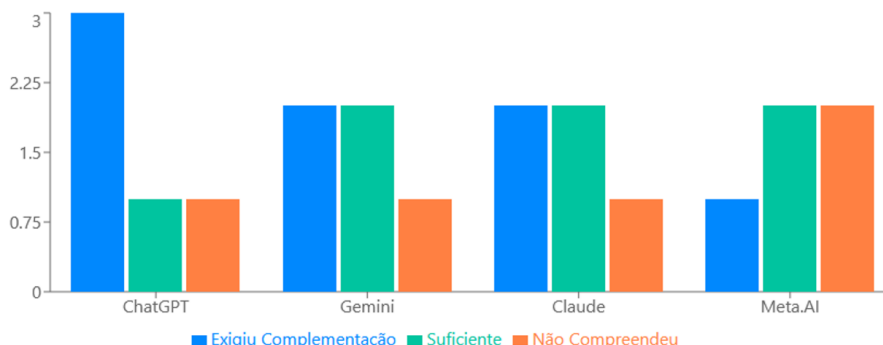
As limitações na parametrização dos sistemas ou no acesso a bases de dados possivelmente são os fatores que causaram a revisão do posicionamento e dificuldade na tomada de decisão, não apenas no caso do Claude, mas de todos os demais. Um usuário desavisado ou impaciente pode ter respostas inadequadas se não tiver o hábito de reforçar seus pedidos ou questionar os resultados oferecidos pelos sistemas.

5.6 Compreensão de comando

Uma outra preocupação na realização do experimento estava relacionada ao nível de profundidade em que as respostas seriam dadas, considerando o processo de aprendizado de máquina e as possibilidades de extração de resultados a partir do estoque de dados disponíveis nos sistemas. Este aspecto poderia indicar, portanto, uma necessidade de complemento sobre o comando apresentado no *prompt*. Os resultados podem ser verificados no gráfico a seguir:

GRÁFICO 6

Compreensão de comando dos sistemas de IA



Nossa hipótese se confirmou. A maioria dos testes (9 / 45%) exigiu uma complementação e em 5 casos (25%) o sistema não compreendeu o que foi solicitado inicialmente, oferecendo respostas insatisfatórias ou inadequadas. Nestes casos, foi necessário mais uma ou duas interações para que os sistemas elaborassem *outputs* correspondentes aos *inputs*.

Na maior parte dos casos os sistemas, quando exigiram complementação, ofereceram respostas iniciais com orientações do que fazer para saber se o conteúdo era falso ou verdadeiro ao invés de responderem objetivamente a questão. Neste caso, o usuário precisa estar minimamente habituado ao funcionamento das IAGs para compreender que os resultados podem não ser os melhores nos primeiros *prompts*.

6. Considerações finais

O experimento revelou que os sistemas de Inteligência Artificial Generativa (IAG) apresentam fragilidades significativas na identificação de *fake news*. A análise dos quatro sistemas testados (ChatGPT, Gemini, Meta.AI e Claude) demonstrou variações consideráveis na precisão e assertividade das respostas. O ChatGPT e o Meta.AI destacaram-se pela maior precisão e uso frequente de fontes verificáveis, enquanto Gemini e Claude apresentaram maior incidência de respostas não responsivas.

A identificação de fontes de informação foi um ponto crítico, com a maioria dos sistemas falhando em citar fontes confiáveis. Observou-se que, em grande parte dos casos, os modelos não especificavam suas fontes, sendo a categoria "que não consta" a mais frequente. Apenas o Meta.AI

mostrou consistência na indicação de fontes, o que reforça a necessidade de maior transparência e acesso a dados em tempo real para melhorar a acuracidade das respostas.

Ainda na perspectiva da confiança sobre os resultados, também a veracidade é falha, na maioria dos casos, nota-se a ausência de indicação de dados para consulta pública. Neste quesito, entretanto, o Meta.AI apresentou um desempenho superior provavelmente devido à maneira como seus algoritmos foram programados e pelo acesso diferenciado a dados e bases consultadas.

A pesquisa também evidenciou que a parametrização dos algoritmos influencia diretamente a qualidade das respostas. Sistemas que utilizam uma abordagem híbrida, combinando dados pré-processados e buscas em tempo real, como o Meta.AI, tendem a oferecer respostas mais precisas e contextualizadas. A precisão das respostas, por exemplo, é variável e depende de um processo de interação que não é indicado a partir do primeiro *prompt*. Na grande maioria dos casos, os sistemas estavam programados para oferecer orientações ou instruções genéricas sobre como evitar acreditar *fake news* ao invés indicar objetivamente se o conteúdo era ou não falso. O Claude, por exemplo, não foi capaz de ser preciso integralmente em nenhuma situação.

A partir dessas análises, foi possível realizar uma comparação quantitativa e qualitativa entre os modelos estudados. Os resultados indicaram que os sistemas, de modo geral, têm dificuldades em determinar se um conteúdo é falso ou não. Considerando os resultados obtidos a partir do teste realizado, o ChatGPT e o Meta.AI obtiveram melhor desempenho geral, apresentando maior número de respostas precisas e fazendo uso mais frequente de fontes verificáveis. Em contrapartida, o Claude demonstrou maior taxa de respostas não responsivas, comprometendo a qualidade das análises oferecidas. Adicionalmente, identificou-se que o uso de fontes confiáveis teve uma forte correlação com a precisão das respostas, atingindo um coeficiente de correlação de 0.82, reforçando a importância da verificação de credibilidade das informações. Já o Gemini apresentou maior dificuldade dentre todos os sistemas, indicando que este modelo tende a adotar uma abordagem mais conservadora, evitando fornecer informações quando há incertezas.

Referências

ALEXANDER, Jeffrey; BREESE, Elizabeth; LUENGO, María. *The crisis of journalism reconsidered: democratic culture, professional codes, digital future*. New York: Cambridge University Press, 2016.

ALVES, Victor Habib Lantyer de Mello. Inteligência Artificial generativa e direito autoral: investigando os limites do uso justo na era da tecnologia. **Revista de Direito, Inovação, Propriedade Intelectual e Concorrência**, Encontro Virtual, v. 10, n. 1, p. 01-19, jan/jul. 2024.

BARDIN, Laurence. **Análise de Conteúdo**. São Paulo: Edições 70, 2016.

BERGER, Guy. Prefácio. In: IRETON, Cherilyn; POSETTI, Julie (editores). **Jornalismo, fake news & desinformação** – manual para educação e treinamento em jornalismo. Série Unesco sobre Educação em Jornalismo. Organização das Nações Unidas para a Educação, a Ciência e a Cultura, 2020. Disponível em: <http://portaldobibliotecario.com/wp-content/uploads/2020/06/ManualFakeNews.pdf>. Acesso em: 18 jan. 2025.

BORGES, JR. Eli. O que é a pós-verdade? Elementos para uma crítica do conceito. **Brazilian Journalism Research**, Brasília-DF, v. 15, n. 3, p. 524 – 545, dez. 2019.

CARDOSO, Fernanda Carolyn. É possível uma subjetividade artificial? Algumas considerações filosóficas sobre os recentes avanços tecnológicos da inteligência artificial a partir do problema difícil da consciência. **Revista Filogênese**, Marília, v. 19, n. 1, p. 3-150, jul. 2024. Disponível em: <https://philarchive.org/archive/CARPUN-2>. Acesso em: 6 fev. 2025.

CARVALHO, Rafiza. Notícias falsas ou propaganda?: Uma análise do estado da arte do conceito fake news. **Questões Transversais**, São Leopoldo, Brasil, v. 7, n. 13, 2019. Disponível em: <https://revistas.unisinos.br/index.php/questoes/article/view/19177>. Acesso em: 12 out. 2024.

GIL, Antonio. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2002.

HABERMAS, Jurgen. **Uma nova mudança estrutural da esfera pública e a política deliberativa**. São Paulo: Editora Unesp, 2023.

JENKINS, Henry. **Cultura da convergência**. São Paulo: Editora Aleph, 2009.

KAUFMAN, Dora; REIS, Patricia; JUNQUILHO, Tainá. Externalidades negativas da inteligência artificial: conflitos entre limites da técnica e direitos humanos. **Revista de Direitos e Garantias Fundamentais**, Santa Lúcia. Vitória, v. 24, n. 3, p. 43-71, setembro/dezembro, 2023. Disponível em: <https://sisbib.emnuvens.com.br/direitosegarantias/article/view/2198/649>. Acesso em: 28 jan. 2025.

LIAO, Q. Vera.; VAUGHAN, Jennifer Wortman. AI transparency in the age of LLMs: a human-centered research roadmap. **arXiv**, preprint. Submitted on 2 Jun 2023 (v1), last revised 8 Aug 2023 (v2). Disponível em: <https://arxiv.org/abs/2306.01941>. Acesso em: 25 jan. 2025.

MAXIMO, Eduardo; VITOR, Andreici; SOUZA, João; COSTA, Eduardo. Aplicação do Commonkads no diagnóstico do modelo de excelência da gestão. **Revista Contemporânea**, v. 3, n. 9, p. 14848–14871, 2023. Disponível em: <https://ojs.revistacontemporanea.com/ojs/index.php/home/article/view/1705>. Acesso em: 3 fev. 2025.

MITCHELL, Melanie. **Artificial intelligence**. New York: Farrar, Straus and Giroux, 2019.

MORAES, D.; RAMONET, I.; SERRANO, P. **Mídia, poder e contrapoder**: da concentração monopólica à democratização da informação. São Paulo: Editora Boitempo Editorial, 2013.

MOREIRA DRUZIANI, Cássio; KERN, Vinícius; CATAPAN, Araci. A gestão a engenharia do conhecimento aliadas na modelagem do conhecimento - análise sistêmica cesm e contextual commonkads de um repósito na web. **Perspectivas em Gestão & Conhecimento**, João Pessoa, v. 2, n. 1, p. 194-217, jan./jun. 2012.

OTTONICAR, Selma Leticia Capinzaiki; VALENTIM, Marta Lúcia Pomim; JORGE, Leandro Feitosa; MOSCONI, Elaine. *Fake news, big data* e o risco à democracia: novos desafios à competência em informação e midiática. **Ibersid**, v. 15, n. 1, p. 63-74, en.-jun., 2021. Disponível em: <https://ibersid.eu/ojs/index.php/ibersid/article/view/4678/4295>. Acesso em: 30 jan. 2025.

PARISER, Eli. “O problema é que damos todo o poder para plataformas como Google e Facebook”. Entrevista concedida a David Alandete. **El País**. Madri, 20 jun. 2017. Disponível em: https://brasil.elpais.com/brasil/2017/06/19/cultura/1497900552_320878.html. Acesso em: 21 abr. 2021.

PRADO, Magaly. **Fake News e inteligência artificial**: o poder dos algoritmos na guerra da informação. São Paulo: Edições 70, 2022.

SAAD, Elizabeth; SANTOS, Márcio Carneiro. Jornalismo, inteligência artificial e desinformação: avaliação preliminar do potencial de utilização de ferramentas de geração de linguagem natural, a partir do modelo GPT, para difusão de notícias falsas. **Estudios sobre el mensaje periodístico**, v. 29, n. 4., p. 783-794, 2023.

SAMPAIO, Rafael Cardoso; SABBATINI; Marcelo, LIMONGI, Ricardo. **Diretrizes para o uso ético e responsável da inteligência artificial generativa**: um guia prático para pesquisadores. São Paulo: Sociedade Brasileira de Estudos Interdisciplinares da Comunicação - Intercom, 2024.

SANTAELLA, Lucia. **A inteligência artificial é inteligente?** São Paulo: Edições 70, 2023.

SASIETA, Héctor; BEPPLER, Fabiano; PACHECO, Roberto. A Memória Organizacional no Contexto da Engenharia do Conhecimento. **DataGramaZero - Revista de Informação**, v. 12, n. 4, 2011.

SHIRKY, Clay. **A cultura da participação**: criatividade e generosidade no mundo conectado. Rio de Janeiro: Jorge Zahar, 2011.

SODRÉ, Muniz. **A sociedade incivil**: mídia, liberalismo e finanças. Petrópolis: Editora Vozes, 2021.

TRAQUINA, Nelson. **Teorias do jornalismo**: porque as notícias são como são. 3 ed. v. 1. Florianópolis: Insular, 2012.

TRINDADE, Alessandra; OLIVEIRA, Henry. Inteligência Artificial (IA) generativa e Competência em Informação: habilidades informacionais necessárias ao uso de ferramentas de IA generativa em demandas informacionais de natureza acadêmica-científica. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 29, Fluxo Contínuo, 2024. Disponível em: <https://periodicos.ufmg.br/index.php/pci/article/view/47485/43923>. Acesso em: 23 jan. 2025.

TUCHMAN, Gaye. A objetividade como ritual estratégico: uma análise das noções de objetividade dos jornalistas. In: TRAQUINA, Nelson. **Jornalismo**: questões, teorias e “estórias”. Lisboa: Vega, 1999, p. 74-90.

VERGARA, Sylvia; PECL, Alketa. Escolhas metodológicas em estudos organizacionais. **Organizações & Sociedade**, v. 10, n. 27, ago. 2003. Disponível em: <https://doi.org/10.1590/s1984-92302003000300001>. Acesso em: 20 jan. 2025.

Logs dos sistemas

ChatGPT. Logs do sistema. out/nov. 2024. Disponível em: <https://drive.google.com/file/d/172zNpjhpxuWSv0qSqsBPm-qrTVuhgfl/view?usp=sharing>

Gemini. Logs do sistema. out/nov. 2024. Disponível em: <https://drive.google.com/file/d/172zNpjhpxuWSv0qSqsBPm-qrTVuhgfl/view?usp=sharing>

Claude. Logs do sistema. out/nov. 2024. Disponível em: <https://drive.google.com/file/d/172zNpjhpxuWSv0qSqsBPm-qrTVuhgfl/view?usp=sharing>

Meta.AI. Logs do sistema. out/nov. 2024. Disponível em: <https://drive.google.com/file/d/172zNpjhpxuWSv0qSqsBPm-qrTVuhgfl/view?usp=sharing>