Consistência e Variação nas Respostas de Inteligências Artificiais Generativas: Uma Pesquisa Comparativa

Eduardo Augusto de Oliveira Mendes[1], Lucy Mari Tabuti[1]

[1]Faculdade XP Educação (XPe) - Belo Horizonte - MG - BRASIL

eduardo.mendes@estudante.xpe.edu.br, lucy.tabuti@igti.edu.br

Abstract. *This article explores the consistency of outputs from nine Generative Artificial Intelligences, collecting and analyzing daily responses to the prompt "implement the person class in Python" over the course of a month. Objective metrics such as file size in bytes and the presence of structural elements (quotes, hashtags, asterisks) were evaluated, in addition to observing linguistic variations, inclusion of references, and possible "hallucinations" in responses. The study identified that Mistral, Maritaca, and ChatGPT were the most consistent, maintaining regular patterns in content and formatting, while Claude, Gemini, and Llama 3.2 showed greater variability, often providing responses beyond what was necessary. It was also noted that external factors, such as the pricing model of some AIs and website updates, influenced the data collection. The article highlights the importance of understanding these variations when choosing Artificial Intelligences for specific tasks and suggests how future research could investigate the influence of technical variables and infrastructure changes on the consistency of responses.*

Keywords: *Generative Artificial Intelligence, Consistent AI Models, Output Variation, Prompt Engineering.*

Resumo. *Este artigo explora a consistência dos outputs de nove Inteligências Artificiais Generativas, coletando e analisando respostas diárias ao prompt "implemente a classe pessoa em Python" ao longo de um mês. Foram avaliadas métricas objetivas como o tamanho dos arquivos em bytes e a presença de elementos estruturais (aspas, cerquilhas, asteriscos), além de observar variações linguísticas, inclusão de referências e possíveis "alucinações" nas respostas. A pesquisa identificou que Mistral, Maritaca e ChatGPT foram os mais consistentes, mantendo padrões regulares no conteúdo e formatação, enquanto Claude, Gemini e Llama 3.2 apresentaram maior variabilidade, com respostas que frequentemente iam além do necessário. Notou-se ainda o impacto de fatores externos, como o modelo de cobrança de algumas inteligências artificiais e atualizações de sites, que influenciaram a coleta. O artigo destaca a importância de entender essas variações ao escolher IA para tarefas específicas e sugere como pesquisas futuras poderiam investigar a influência de variáveis técnicas e mudanças na infraestrutura sobre a consistência das respostas.*

Palavras-chave: *Inteligência Artificial Generativa, Modelos de IA Consistentes, Variação de Outputs, Engenharia de Prompt.*

1. Introdução

Nos últimos anos, a Inteligência Artificial (IA) Generativa vem ganhando destaque em diversas áreas, demonstrando seu potencial para criar conteúdos textuais de alta qualidade, responder perguntas, escrever códigos e realizar outras tarefas complexas. Aplicações como ChatGPT e Copilot estão democratizando o uso da tecnologia, sinalizando uma futura adoção positiva. No entanto, é importante que usuários e

organizações estejam conscientes dos benefícios e limitações, como a propensão a gerar informações incorretas ou enganosas [Ryberg 2024].

Essa expansão tem gerado uma demanda crescente por esses serviços, levantando questões sobre a consistência e previsibilidade das respostas dos diferentes modelos de IA generativa. Cada modelo pode interpretar um mesmo *prompt*[1] de formas distintas. Diante disso, surgem os seguintes questionamentos: as respostas das IA generativas permanecem consistentes ao longo do tempo, ou estão sujeitas a alterações conforme o horário e o dia da pesquisa? Até que ponto fatores externos, como atualizações nos modelos ou a infraestrutura das plataformas, podem influenciar a previsibilidade e uniformidade das respostas geradas?

Em um estudo realizado sobre o uso eficiente de tecnologias de IA Generativa em demandas acadêmicas e científicas, concluiu-se que o efeito de verdade e ilusão, a percepção seletiva e o viés pró-inovação são vieses cognitivos que estimulam o uso de ferramentas de IA Generativa para atender demandas informacionais de natureza acadêmica e científica. As tecnologias são desenvolvidas para responder às necessidades humanas e seus atributos éticos estão intrinsecamente ligados aos seus usuários. Os benefícios e as desvantagens de uma ferramenta dependem diretamente da forma como ela é utilizada [Trindade & Oliveira 2024]. Nesse sentido, a consistência das respostas pode orientar na escolha da tecnologia mais adequada ao perfil e às necessidades específicas do usuário.

A presente pesquisa realiza uma análise sistemática da consistência de nove IA generativas ao longo de um período de um mês. Por meio de métricas objetivas, como o tamanho das respostas em bytes e a presença de elementos estruturais, buscou-se identificar padrões de variação entre os modelos. Por meio dessa investigação, o artigo identifica quais modelos se destacam em consistência, oferecendo direcionamentos para desenvolvedores e empresas que desejam adotar soluções de IA consistentes.

2. Fundamentação Teórica

A IA constitui um amplo campo da Ciência da Computação voltado para a criação de máquinas e sistemas capazes de realizar tarefas que tradicionalmente exigem inteligência humana. Com o advento de grandes modelos de dados e o avanço da capacidade computacional, as IA transformaram-se em mecanismos extremamente poderosos. Elas possibilitam a automação de tarefas altamente complexas, auxiliam na tomada de decisões estratégicas e otimizam processos, tornando-os mais eficientes. Essa capacidade é fundamentada no uso de algoritmos sofisticados que processam volumes massivos de informações [Pereira et al. 2024].

As IA podem ser categorizadas em diferentes tipos: classificadoras (aprendizagem de máquina), preditivas e generativas. Esta última se subdivide em geradoras de texto e de imagens. As IA Generativas de textos, também conhecidas como modelos de linguagem de grande escala (LLM - *Large Language Model*), representam uma sofisticada interpretação linguística baseada no aprendizado computacional das relações entre palavras. Esses modelos são capazes de produzir respostas, manipular estruturas linguísticas e operar com os jogos de sentido da linguagem. Sua capacidade

[1] Um *prompt* é uma frase em linguagem natural que solicita a uma IA generativa que execute uma tarefa.

reside em gerar sentenças convincentes ao replicar os padrões estatísticos da linguagem, utilizando bancos de dados textuais coletados da internet. Ao absorver múltiplos padrões linguísticos, conseguem aprimorar aspectos como correção gramatical, estrutura textual e estilo de escrita [Batista & Santaella 2023].

Os LLM representam palavras e frases como sequências ordenadas de números (vetores)[2], semelhantes a coordenadas x e y, mas com centenas ou milhares de dimensões. Nesse espaço multidimensional, padrões de significado emergem naturalmente. Por exemplo, quando solicitado a completar a analogia "homem : rei :: mulher : ?", o algoritmo realiza operações vetoriais interessantes: subtrai o vetor de "homem" do vetor de "rei" e, em seguida, adiciona o vetor de "mulher", produzindo um novo vetor que está muito próximo do vetor de "rainha". Isso demonstra como categorias como gênero e nobreza existem como padrões semânticos latentes dentro dos LLM. A similaridade entre dois conceitos é medida por meio de um "*score* de similaridade" entre seus vetores. Quanto mais próximo de 1.0 for esse *score*, mais semanticamente similares são os termos. Dentro de cada modelo, essa similaridade é calculada utilizando a similaridade de cosseno, um método matemático que mede o ângulo entre os vetores no espaço multidimensional [Koehl & Vangsness 2023].

Os LLM apresentam desafios significativos de percepção e confiança do usuário. As "alucinações", isto é, informações geradas que parecem precisas, mas carecem de fundamentação, podem comprometer a capacidade dos usuários de tomar decisões independentes e críticas. Essa ambiguidade entre conteúdo factual e criativo erosão a confiança dos usuários, exigindo uma abordagem estratégica e multidimensional para mitigar os riscos percebidos e reais da geração de conteúdo por IA [Ahmadi 2024].

As tecnologias de Inteligência Artificial generativa estão em rápida evolução, marcadas por incertezas quanto ao seu desenvolvimento futuro. Sua complexidade e potencial de transformação dependem de avanços tecnológicos e de dinâmicas sociais e políticas. A compreensão dessas tecnologias emergentes requer uma abordagem crítica e multidimensional, com ênfase em duas dimensões fundamentais: a reflexão ética e o desenvolvimento regulatório. Sendo a IA um sistema que interage profundamente com o comportamento humano, a análise ética torna-se imperativa para antecipar, compreender e mitigar possíveis riscos. Mais do que simplesmente regular, é essencial desenvolver estratégias éticas que orientem a implementação dessas tecnologias, garantindo que seu avanço seja pautado por princípios de justiça, responsabilidade e respeito aos valores sociais [Rossetti & Garcia 2023].

3. Metodologia

Esta seção descreve a metodologia proposta, que é composta por duas etapas: i) definição do *prompt* e período de coleta e ii) organização dos arquivos diários e contagem de *bits* ao fim do período de coleta. Posteriormente a isso, houve a necessidade de estabelecer os critérios de análise dos resultados, onde cada uma dessas etapas é descrita a seguir.

[2] Os vetores em LLM são representações matemáticas de palavras e frases que capturam seus significados e relações em um espaço multidimensional.

3.1. Definição do *Prompt* e Período de Coleta

O primeiro passo da pesquisa consistiu em definir um comando padronizado que fosse interpretado de maneira similar por diferentes IA generativas, propondo-se maximizar a comparação objetiva dos resultados. O *prompt* escolhido foi: "Implemente a classe pessoa em Python", visando obter como resposta uma implementação básica de classes em codificação, acompanhada de uma explicação sintética sobre sua estrutura.

Esse *prompt* foi escolhido por sua simplicidade, o que favorece uma análise de elementos estruturais padronizados, como construtores e atributos básicos, com a possibilidade de variação do tamanho da resposta. O período de coleta foi definido entre os dias 1 e 31 de outubro de 2024, totalizando 31 dias de amostras diárias. As coletas foram realizadas em horários aleatórios ao longo do dia de cada coleta, sem um padrão fixo, permitindo observar possíveis influências do momento exato em que as respostas foram geradas.

3.2. Organização dos Arquivos Diários e Contagem de *Bits*

Após a definição do *prompt* e do período de coleta, foi necessário organizar as pastas que guardariam as respostas das IA. Não houve critério para a escolha destas IA generativas, sendo utilizadas aquelas em que eram de conhecimento dos autores até a data do desenvolvimento do artigo. Todas elas são gratuitas mas oferecem limitações por uso, sendo que cada uma possui suas particularidades com relação à cobrança. Para cada dia do período de coleta, foi criada uma pasta identificada pela data, por exemplo: 1-10-2024, 2-10-2024, onde foram salvas as respostas de nove IA distintas, sendo elas: ChatGPT, Copilot, Claude, Gemini, Llama 3.2 (instalada localmente), Maritaca, Mistral, Perplexity e Writesonic.Cada arquivo gerado foi nomeado de acordo com a IA responsável pelo output,por exemplo, ChatGPT.txt, Copilot.txt etc., contendo a resposta completa para o *prompt* do dia.

Ao final do período de coleta, cada resposta diária foi analisada quanto ao seu tamanho em *bytes*, fornecendo uma métrica objetiva para o estudo. A contagem de *bytes* foi realizada diretamente sobre o conteúdo de cada arquivo, permitindo avaliar o comprimento das respostas e investigar a consistência do tamanho gerado por cada IA ao longo do tempo.

3.3. Critérios de Análise dos Resultados

Com os *outputs* organizados e medidos, separou-se os dados em uma planilha que possibilitou estabelecer os critérios objetivos de análise. Assim, com o objetivo de identificar tendências de prolixidade ou concisão, cada resposta foi avaliada pelo número de *bytes* que ocupa. Esse critério permite medir objetivamente a quantidade de conteúdo gerado por cada IA, facilitando comparações diretas entre os diferentes modelos ao longo do período estudado.

4. Resultados

Nesta seção, são apresentados os resultados da análise dos *outputs* gerados diariamente pelas diferentes IA ao longo do período de coleta. A avaliação inclui uma descrição estatística dos tamanhos dos arquivos, bem como, gráficos que ilustram o

comportamento de geração de conteúdo de cada IA. A análise é dividida em três partes principais: i) descrição estatística do tamanho dos *outputs*, ii) somatório total de *bytes* gerados no mês e iii) variação diária dos tamanhos das respostas. Aqui vale ressaltar que os experimentos foram conduzidos em um notebook Dell com sistema operacional Windows 10, processador Intel® Core™ i7-4510U CPU @ 2.60GHz 8GB DDR3 e NVIDIA® GeForce® GT 740M 2GB, com gráficos gerados via PowerBI.

4.1. Descrição Estatística do Tamanho dos Arquivos

Os tamanhos das respostas em *bytes* foram compilados e descritos com base em estatísticas básicas para cada IA ao longo dos 31 dias de coleta. A Tabela 1 resume as métricas de média, desvio padrão, mínimo, mediana e máximo do tamanho dos *outputs* para cada IA.

Tabela 1. Estatísticas descritivas do tamanho dos *outputs*

IA	Média (Bytes)	Desvio Padrão	Mínimo (Bytes)	Mediana (Bytes)	Máximo (Bytes)
Chatgpt	788	163	509	782	1106
Claude	1496	900	932	1167	4920
Copilot	980	564	397	659	1772
Gemini	3950	306	3372	3969	4785
Llama 3.2	2457	551	1764	2351	3829
Maritaca	1014	248	605	1011	1629
Mistral	1499	239	1189	1467	2209
Perplexity	1708	301	1181	1761	2344
Writesonic	1237	395	570	1466	1689

A média e o desvio padrão permitem observar quais IA têm maior consistência no tamanho dos *outputs*, enquanto os valores mínimo e máximo evidenciam as variações extremas.

4.2. Somatório Total de *Bytes* Gerados no Mês

O gráfico de barras da Figura 1 mostra o somatório total de *bytes* gerados por cada IA ao longo do mês de outubro de 2024. Esse gráfico oferece uma visão do volume de dados gerado por cada IA, permitindo identificar quais modelos foram mais prolixos ou concisos ao longo do período.

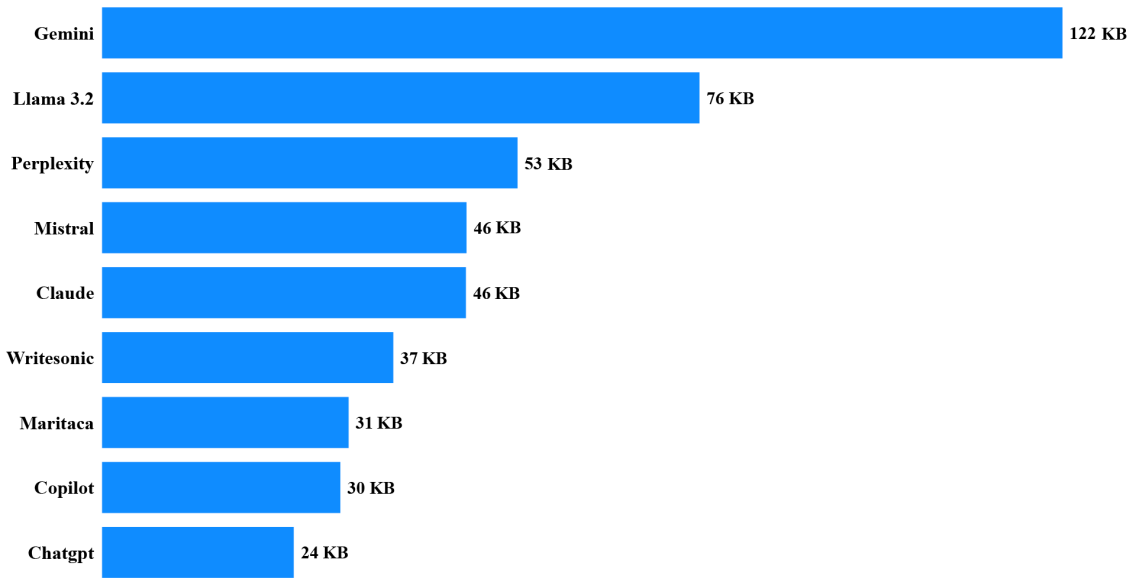

Figura 1. Total de bytes gerados por IA em outubro

4.3. Variação Diária do Tamanho das Respostas

O gráfico de linhas da Figura 2 ilustra a variação diária no tamanho das respostas de cada IA ao longo do mês. Esse gráfico destaca as flutuações na quantidade de conteúdo gerado e permite observar se houve consistência ou variações significativas em determinados dias.

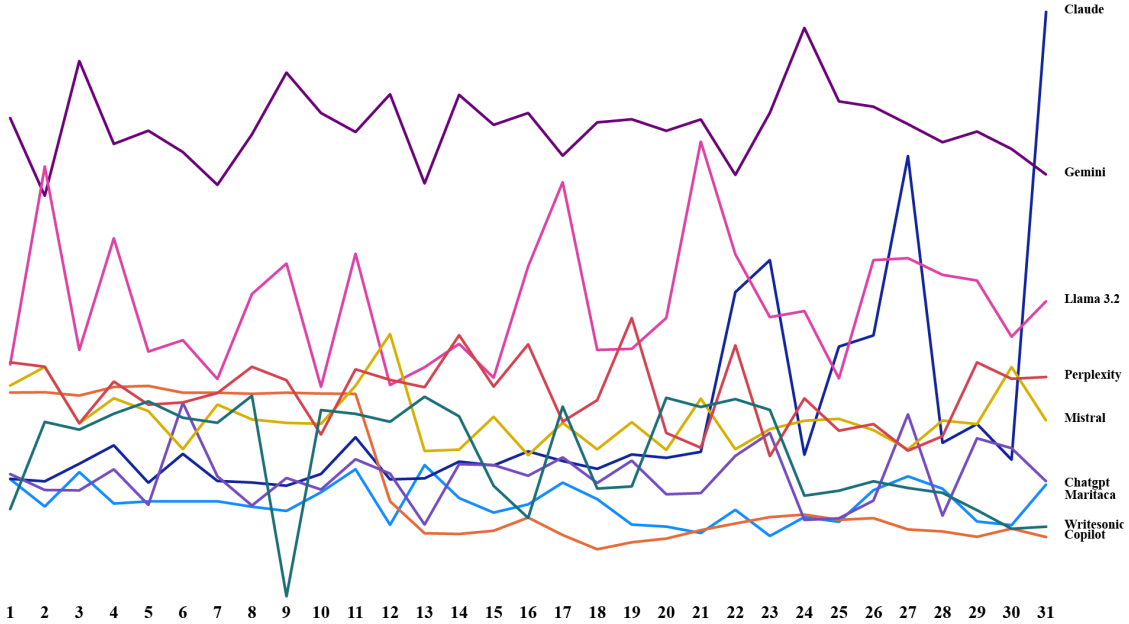

Figura 2. Variação diária no tamanho das respostas por IA

5. Discussão

Nesta seção, são discutidos os principais achados da análise dos *outputs*. A partir dos resultados, pôde-se observar que algumas IA mantêm uma consistência maior em termos de tamanho de resposta, enquanto outras apresentam variações mais acentuadas.

A Figura 1 e a Figura 2 permitem observar as questões sobre as abordagens de geração de conteúdo dos diferentes modelos e sua regularidade ao responder a um *prompt* padronizado. As questões incluem a presença de elementos estruturais específicos, diferenças no estilo e conteúdo das respostas e as possíveis implicações de fatores externos, como a conexão com a internet, sobre a variação dos resultados.

5.1. Presença de Elementos Estruturais

Para avaliar a consistência dos *outputs*, realizou-se uma análise da simbologia presente nas respostas das IA. Em vez de analisar atributos ou métodos específicos na implementação da classe Pessoa, concentramos nossa análise nos elementos de formatação do texto gerado. Entre os elementos observados estavam:

- Aspas duplas e simples ("" e '): As aspas foram frequentemente utilizadas para definir *strings* e mensagens de exibição, com algumas IA padronizando o uso de aspas duplas, enquanto outras optaram por aspas simples. Esse comportamento sugere variações na padronização do estilo de código entre os diferentes modelos.
- Cerquilha # para comentários ou aplicação de títulos: O uso de # para indicar comentários ou titulação foi uma prática comum, mas variou quanto à frequência e à quantidade de informações dispostas. Algumas IA explicavam cada linha do código com comentários, intitulando-a para cada trecho, enquanto outras usavam comentários e títulos de maneira mais pontual.
- Hífens - e asteriscos * para listas ou tópicos: Certos modelos utilizaram hífens ou asteriscos para organizar listas de instruções ou pontos explicativos, especialmente nas partes em que explicavam o código. Essa escolha pode indicar preferências na apresentação visual dos *outputs*, refletindo o estilo particular de cada IA.

Estas considerações sobre a utilização de símbolos são importantes para compreender como diversas inteligências artificiais tratam a formatação e a organização das respostas, o que pode ser benéfico para programadores que buscam uniformidade visual em interações contínuas com diversos modelos. Ademais, isso pode afetar o tamanho do arquivo, pois esses itens de formatação ocupam tanto espaço quanto o texto em si.

5.2. Variação na Linguagem das Respostas

Durante a pesquisa, foi notado que o Writesonic, Copilot e Claude, por vezes, sem uma frequência relativa, forneceram respostas em inglês, mesmo quando o *prompt* estava em português. Essa variabilidade linguística pode indicar que algumas IA tendem a "preferir" o inglês como idioma de saída, possivelmente devido à predominância de dados de treinamento em inglês ou a heurísticas internas do modelo que levam à escolha de um idioma específico.

5.3. Inclusão de Fontes e Referências

Outro ponto relevante foi que o Copilot e Gemini incluíram referências e *links* para documentação ou materiais adicionais. Os links são funcionais e incluíam desde artigos até vídeos mas, embora útil em muitos contextos, essa prática nem sempre era necessária, considerando o escopo limitado do *prompt*. Em certos casos, *links* incluídos eram irrelevantes para a implementação da classe pessoa, o que levanta a hipótese de que a IA pode estar "alucinando", ou seja, um comportamento em que os modelos de linguagem geram informações incorretas ou desnecessárias.

5.4. Expansão Desnecessária do Conteúdo

Além de fornecer a implementação solicitada, o Claude, Gemini, Lama 3.2 expandiram a resposta com informações que iam além do necessário, como explicações extensas sobre conceitos de programação que não eram diretamente relevantes. Isso pode indicar que as IA estão programadas para interpretar *prompts* de maneira que favoreça respostas mais detalhadas, mas pode gerar um excesso de informações que não adicionam valor direto ao objetivo da resposta. A tendência a “alucinar” pode ser uma área relevante para otimizações futuras nesses modelos.

5.5. Impacto de Variáveis Externas

Por fim, discutimos a possibilidade de que fatores externos, como a qualidade da conexão com a internet e a carga dos servidores das IA no momento de cada consulta, possam influenciar o conteúdo ou a estrutura das respostas. Uma hipótese é que flutuações na latência ou no acesso à base de dados possam impactar a consistência dos *outputs*, levando a variações na qualidade ou no estilo de formatação. Embora não tenhamos evidências concretas para confirmar essa influência, o aspecto técnico da conexão é um ponto que poderia ser aprofundado em pesquisas futuras para verificar sua relevância.

No gráfico da Figura 2, o comportamento dos dados revelou padrões de variação interessantes ao longo do mês, com destaque para dois pontos específicos que impactaram diretamente a análise. No dia 9, o WriteSonic apresentou um valor de zero *bytes* devido ao modelo de cobrança da IA, que impossibilitou a coleta de respostas nesse dia específico. Esse fator reforça a necessidade de considerar as particularidades dos modelos de negócio de cada IA ao realizar estudos de consistência. Além disso, a linha do Copilot exibe uma queda brusca durante o período, efeito de uma mudança significativa no visual do *site* que impactou a estrutura de resposta gerada. Esses eventos mostram como variáveis externas e mudanças na plataforma podem influenciar a resposta em si e a disponibilidade e estrutura dos dados gerados, ressaltando a importância de contextualizar esses fatores ao interpretar as métricas de consistência entre as IA.

6. Conclusão e trabalhos futuros

Esta pesquisa explorou a consistência de respostas geradas por diferentes IA em um período de um mês, avaliando aspectos como o tamanho dos *outputs* e suas causas. A análise revelou diferenças importantes entre os modelos testados, destacando o impacto

das características de cada IA sobre a regularidade das respostas geradas. Foi observado que, em termos de consistência, o ChatGPT, Mistral e Maritaca se destacaram, apresentando tamanho de resposta e formatação mais uniformes ao longo do tempo. Em contrapartida, IA como Claude, Gemini e Llama 3.2 mostraram-se menos consistentes, frequentemente expandindo o conteúdo além do necessário ou incluindo elementos extras que não eram solicitados pelo *prompt*. Esses desvios sugerem que alguns modelos podem estar suscetíveis a interpretações mais abertas dos *prompts*, gerando conteúdos que, embora detalhados, fogem ao escopo esperado.

A importância desta pesquisa reside em fornecer uma visão prática sobre a variação entre modelos de IA generativas, oferecendo subsídios para a escolha de modelos que prezam pela consistência de *outputs* em tarefas específicas. Esse conhecimento pode auxiliar desenvolvedores e empresas a selecionar modelos que melhor atendam suas necessidades em termos de previsibilidade de respostas e estrutura de dados.

Para trabalhos futuros, recomenda-se duas direções de pesquisa: i) explorar a influência de diferentes horários e condições de conexão sobre a consistência das respostas de IA, para verificar se variações técnicas interferem na regularidade dos outputs; e ii) investigar o impacto de atualizações de software e mudanças na infraestrutura dos *sites* das IA sobre a continuidade e qualidade das respostas, considerando que alterações não planejadas, como a observada no Copilot, podem impactar significativamente a análise de consistência ao longo do tempo. Esses temas ajudariam na compreensão sobre os fatores que afetam a previsibilidade das IA generativas além de contribuir para o desenvolvimento de modelos mais robustos e consistentes.

Repositório de Dados e Códigos

Para acessar os arquivos utilizados para elaboração desta pesquisa, incluindo códigos gerados, planilha de controle e informações gerais, visite o repositório no GitHub, disponível no link:
https://github.com/EduAugustoM/ConsistenciadeIAsGenerativas

7. Referências

Ahmadi, A. (2024). Unravelling the Mysteries of Hallucination in Large Language Models: Strategies for Precision in Artificial Intelligence Language Generation. Asian Journal of Computer Science and Technology, 13(1), 1–10. https://doi.org/10.70112/ajcst-2024.13.1.4144

Batista, A. R. F. & Santaella, L. (2023). IAs Generativas. Aurora. Revista de Arte, Mídia e Política. [s. l.]: Pontifical Catholic University of Sao Paulo (PUC-SP). DOI 10.23925/1982-6672.2023v16i47p76-94.

Koehl, D., & Vangsness, L. (2023). Measuring Latent Trust Patterns in Large Language Models in the Context of Human-AI Teaming. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 67(1), 504-511. https://doi.org/10.1177/21695067231192869

Pereira, J. V. S.,Vargas, A. A. F. & Pereira, J. S. (2024). IA generativa versus desenvolvimento de software: um olhar do presente para o futuro. Pesquisa & Educação a Distância, v. 1, n. 31, ISSN 2358-646x.

Rossetti, R. & Garcia, K. (2023). Inteligência Artificial Generativa: questões jurídicas e éticas em torno do ChatGPT. Virtuajus, v. 8, n. 15, p. 253-264. ISSN 1678-3425.

Ryberg, R. O. (2024). Inteligencia Artificial Generativa: presente y futuro. Revista Sistemas. [S. l.]: Asociacion Colombiana Ingenieros de Sistemas. DOI 10.29236/sistemas.n172a2.

Trindade, A. S. C. E. & Oliveira, H. P. C. (2024). 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. [S. l.]: FapUNIFESP (SciELO). DOI 10.1590/1981-5344/47485.