Análise de Ferramentas de Detecção de IA para Textos Científicos em Português Gerados por ChatGPT, Gemini e DeepSeek

Lucas S. Candido[1], Christian A. de Melo Barbosa[1], Lylia G. Martins[1], Esdras J. H. Costa[1]

[1]Instituto Federal de Alagoas (IFAL) – Campus Maceió
Maceió – AL – Brazil

lucas.santos.candido@gmail.com, {camb1, lgm3}@aluno.ifal.edu.br, esdras.costa@ifal.edu.br

Abstract. *This study investigated the effectiveness of five AI detection tools (ZeroGPT, JustDone AI Detector, Writer AI Detector, Seo AI Detector and Summarizer AI Detector) in identifying scientific texts in Portuguese generated by different language models (ChatGPT, Gemini and DeepSeek), comparing them with control samples written by humans. The research used 50 manuscripts and error metrics (MAE and RMSE) to evaluate the performance of the detectors. The results showed that ZeroGPT had the highest accuracy in detecting AI-generated content, with the lowest average errors for synthetic texts. However, even ZeroGPT had a false positive rate when classifying human text. The other tools had limitations such as high false positive rates, low sensitivity in detecting AI in Portuguese, or inconsistency in results. The analysis concludes that although ZeroGPT is the most effective tool among those evaluated, the lack of a perfect solution underscores the need for continuous advances in AI detection technology to ensure the integrity of scientific production.*

Resumo. *Este estudo investigou a eficácia de cinco ferramentas de detecção de IA (ZeroGPT, JustDone AI Detector, Writer AI Detector, Seo AI Detector e Summarizer AI Detector) na identificação de textos científicos em português gerados por diferentes modelos de linguagem (ChatGPT, Gemini e DeepSeek), comparando-as com amostras de controle escritas por humanos. A pesquisa utilizou 50 manuscritos e métricas de erro (MAE e RMSE) para avaliar o desempenho dos detectores. Os resultados revelaram que o ZeroGPT apresentou a maior precisão na detecção de conteúdo gerado por IA, com os menores erros médios para textos sintéticos. No entanto, mesmo o ZeroGPT demonstrou uma taxa de falsos positivos ao classificar textos humanos. As outras ferramentas exibiram limitações como altas taxas de falsos positivos, baixa sensibilidade na detecção de IA em português ou inconsistência nos resultados. A análise conclui que, embora o ZeroGPT seja a ferramenta mais eficaz entre as avaliadas, a ausência de uma solução perfeita sublinha a necessidade de avanços contínuos na tecnologia de detecção de IA para garantir a integridade da produção científica.*

1. Introdução

A inteligência artificial (IA) generativa tem avançado exponencialmente nos últimos anos, demonstrando capacidade notável de produzir textos coerentes, imagens realistas e vídeos impressionantes. Embora os modelos de IA existam há bastante tempo, sua popularidade cresceu significativamente com a introdução de chatbots sofisticados, capazes de interagir em tempo real e gerar respostas de alta qualidade. Essa evolução tem impactado diversos setores, incluindo a produção de conteúdo científico, o que tem gerado debates importantes sobre os riscos éticos e a integridade acadêmica no uso desses grandes modelos de linguagem (LLM) [Sullivan et al. 2023]. Diante desse cenário, uma pesquisa realizada em janeiro de 2023 com 1.000 universitários revelou que pelo menos um terço dos estudantes utilizava o ChatGPT para escrever seus ensaios. Além disso, 75% destes usuários admitiram estar cientes de que estavam trapaceando, mas continuaram a usar a ferramenta mesmo assim [Intelligent.com 2023].

Estamos observando neste exato momento como a IA generativa está mudando o campo científico, enquanto o uso responsável pode trazer benefícios para a ciência como um todo, o uso indevido apresenta um sério risco a credibilidade da pesquisa. [Almeida et al. 2023] comprova, em seu estudo, que mestres e doutores com pelo menos cinco anos de experiência em docência tiveram dificuldade em distinguir resumos de trabalhos de conclusão de curso gerados pelo ChatGPT daqueles escritos por humanos. A pesquisa também sugere que os docentes tenderam a classificar erroneamente os resumos produzidos pela IA, categorizando-os como textos escritos por pessoas. Antes mesmo da popularização dos chatbots, estima-se que mais de 120 publicações em 2014 foram retratadas e removidas por terem sido criadas por um programa que gera artigos falsos sem coesão e sentido [Cabanac and Labbé 2021]. Um estudo recente de [Liyanage et al. 2022] evidencia que, embora existam pesquisas sobre a detecção de textos gerados automaticamente, ainda há pouca investigação voltada especificamente para textos no domínio acadêmico.

Diante de tal contexto, [Hammad 2023] propõe uma possível solução: a criação de programas capazes de detectar se um pesquisador está utilizando técnicas de inteligência artificial para produzir seus ensaios. [Ladha et al. 2023] explica que os detectores de IA são treinados em enormes bancos de dados para conseguir identificar corretamente se o conteúdo foi gerado por máquina ou escrito por humano. O pesquisador emprega os algoritmos desenvolvidos pelas empresas Copyleaks, Writer.com, e Content at scale para analisar se essas ferramentas são realmente capazes de distinguir entre ambos os tipos de ensaio. O autor concluiu que os três softwares demonstraram alta precisão na identificação de manuscritos em inglês elaborados pelo ChatGPT, reforçando sua eficácia na detecção de textos automatizados.

Similarmente, [Elkhatat et al. 2023] conduziram um estudo para avaliar a eficiência de cinco ferramentas de detecção de IA (OpenAI, Copyleaks, Writer, GPTZero, e CrossPlag) em relação a ensaios acadêmicos gerados pelas versões 3.5 e 4 do ChatGPT. A pesquisa utilizou quinze amostras para cada modelo generativo e cinco amostras de controle escritas por humanos, os resultados indicaram uma variação considerável dos detectores em identificar e categorizar corretamente os textos. Além disso, os algoritmos foram mais precisos na identificação de conteúdo gerado pelo ChatGPT 3.5 do que pela versão 4, isso sugere que, à medida que os LLM se tornam mais sofisticados, a tarefa de detecção se torna mais desafiadora. Acrescentando-se as inconsistências nas amostras de

controle, resultando falsos positivos e classificações incorretas, levantando preocupações sobre a possibilidade de acusar erroneamente autores de utilizar IA.

Em relação aos trabalhos de [Elkhatat et al. 2023; Ladha et al. 2023], observa-se que ambas pesquisas se concentram em apenas duas categorias de amostras: textos reais e sintéticos. No entanto, [Weber-Wulff et al. 2023] abrangeram o escopo de seu experimento, incluindo textos gerados por máquina com edições humanas, conteúdos escritos em linguagem não inglesa e traduzidos para o inglês por IA, além de textos sintéticos parafraseados por algoritmos de inteligência artificial. Os pesquisadores utilizaram 14 detectores de IA e realizaram 54 casos de teste para cada uma das ferramentas, totalizando 756 testes. Os resultados demonstraram que estes programas são falhos, facilmente manipuláveis e propensos a erros significativos, tanto na identificação incorreta de texto humano quanto na falha em detectar texto gerado por IA.

Entretanto, poucos autores têm-se focado em explorar a funcionalidade desses detectores de IA em idiomas diferentes do inglês, o que cria uma lacuna significativa na literatura científica. Embora [Candido et al. 2024] tenham abordado essa questão em analisar manuscritos sob a perspectiva da língua portuguesa, seu estudo limita-se a avaliar textos científicos gerados exclusivamente pelo ChatGPT 3.5. Dessa forma, a pesquisa não valida a eficácia das ferramentas de detecção na identificação de textos produzidos por outros modelos generativos, deixando uma área importante ainda por ser investigada.

Diante do contexto apresentado, esta pesquisa se propõe a investigar a seguinte questão: As ferramentas atuais de detecção de IA conseguem identificar textos científicos em português gerados pelo ChatGPT, Gemini e DeepSeek?

2. Metodologia

O presente trabalho atua como uma pesquisa exploratória de caráter aplicada, seguindo os conceitos definidos por [Gil 2017], pois busca investigar o potencial de ferramentas especializadas para a detecção de conteúdo científico gerado por grandes modelos de linguagem em português, fenômeno recente e pouco explorado neste idioma.

2.1. Seleção de LLM e Geração de Manuscritos Sintéticos

Atualmente existem dezenas de modelos de IA generativa disponíveis no mercado para a geração de textos em forma de chatbot e API (Interface de Programação de Aplicações), dessa forma estabelecemos os seguintes critérios para seleção dos LLMs:

i. O modelo deve ter a capacidade de entender instruções na linguagem portuguesa e gerar textos neste mesmo idioma;
ii. Janela de contexto com suporte mínimo a 100 mil tokens;
iii. Quantidade mínima de 5.000 tokens de saída;
iv. Disponível versão gratuita de uso em formato de chatbot.

As três primeiras restrições garantem que o modelo esteja habilitado a receber e entender comandos mais complexos a fim de produzir conteúdo com maior quantidade de caracteres e palavras, enquanto a quarta regra visa utilizar as inteligências artificiais mais acessíveis ao público acadêmico e comumente usadas. Utilizamos diferentes buscadores na internet e optamos por aplicar o ChatGPT com o GPT-4o mini (o modelo GPT-4o foi descartado da presente pesquisa devido a restrição de uso diário imposta pela

empresa proprietária), Gemini em suas versões 1.5 Flash e 2.0 Flash Experimental, além do DeepSeek na variante DeepSeek-V3, conforme a Tabela 1.

Tabela 1. Especificações das LLMs utilizadas

	ChatGPT	Gemini	Gemini	DeepSeek
Versão do LLM	GPT-4o mini	1.5 Flash	2.0 Flash Experimental	DeepSeek-V3
Janela de contexto	128 mil tokens	1 milhão de tokens	1 milhão de tokens	128 mil tokens
Limite de saídas por requisição	16 mil tokens	8 mil tokens	8 mil tokens	8 mil tokens
Desenvolvidor	OpenAI	Google Deepmind	Google Deepmind	DeepSeek AI
Referência	[OpenAI 2024]	[Team et al. 2024]	[Google DeepMind 2025]	[DeepSeek-AI et al. 2025]

Após a seleção dos LLMs, realizamos um design sistemático para criação e de otimização da instrução de comando (prompt) para guiar as respostas das IA na geração dos ensaios artificiais, garantindo maior precisão e coerência da saída de texto esperada [Chen et al. 2024]. Diante disso empregamos a técnica de role-prompting para atribuir o papel específico ao modelo de pesquisador científico para que suas respostas se alinhem com o conhecimento e a perspectiva esperados para essa função. Demos instruções claras e precisas sobre o formato desejado da saída do conteúdo, além disso usamos processo iterativo de interação para orientar a saída das sessões do texto em etapas. Mediante o exposto, criamos diversas tentativas de prompts para gerar os manuscritos sintéticos, após analisarmos os ensaios, escolhemos a instrução demonstrada no esquema da Figura 1.

Por fim, nesta etapa 40 manuscritos foram elaborados, sendo 10 correspondentes a cada uma das versões dos modelos. Os ensaios produzidos abrangeram diversas áreas de estudo, como história, educação física, saúde pública, ciências sociais e engenharias

ENTRADA: Atue como um pesquisador científico especialista em [GRANDE ÁREA DE ESTUDO] em relação ao tema "[ÁREA DE ESTUDO]" e escreva um artigo científico de revisão bibliográfica sobre "[TEMA DO ARTIGO]". O artigo deverá ser escrito de acordo com as normas brasileiras e melhores práticas científica, será estruturado em 5 sessões: introdução, metodologia, discussão, conclusão e referência bibliográfica. Estruture cada sessão de forma lógica com início, meio e fim; sendo coerente entre cada parágrafo de uma mesma sessão e utilizando como base o conteúdo escrito anteriormente em todo o texto do manuscrito. Porém, você escreverá as sessões do artigo etapa por etapa, irei dar a instrução para você redigir a introdução, você irá produzir a introdução e assim sucessivamente para cada sessão do artigo. Entendido?

SAÍDA: Entendido!

ENTRADA: Introdução

SAÍDA: [Conteúdo da sessão de introdução gerado]

⋮

SAÍDA: [Conteúdo da última sessão do artigo gerado]

ENTRADA: Absorva todas as informações desse artigo, crie um título para ele e um resumo científico de 150 palavras.

SAÍDA: [Título e resumo do artigo gerado]

Figura 1. Simulação de conversa com chatbot para geração de manuscritos

2.2. Seleção de ferramentas para detecção de IA

A princípio analisamos as ferramentas utilizadas em trabalhos semelhantes na literatura atual e em extensiva busca online, estabelecemos os seguintes critérios de escolha: (i) funcionalidade principal em distinguir entre texto escrito por uma pessoa e texto escrito por computador, sem considerar plágio; (ii) sistema web com disponibilidade gratuita para avaliar documentos com ao menos 2.500 palavras; (iii) capacidade de analisar textos em português. Ao final do processo, encontramos cinco ferramentas que atenderam os requisitos, nomeadas, ZeroGPT, JustDone AI Detector, Writer AI Detector, Seo AI Detector, Summarizer AI Detector [Summarizer.org 2025], conforme a Tabela 2. Salientamos que cada um desses algoritmos exibem os resultados de classificação dos ensaios de formas distintas uns dos outros, porém utilizamos o valor percentual do manuscrito ter sido escrito por humano ou por IA.

Tabela 2. Limitações das ferramentas gratuitas

Nome da ferramenta	Tamanho mínimo	Tamanho máximo	Referência
Zero GPT	250 caracteres	50.000 caracteres	[ZeroGPT.net 2025]
Justdone AI Detector	Não especificado	Não especificado	[JustDone 2025]
Writer AI Detector	60 palavras	5.000 palavras	[Writer 2025]
Seo AI Detector	Não especificado	20.000 caracteres	[Seo.Ai 2024]
Summarizer AI Detector	100 palavras	3.000 palavras	[Summarizer.org 2025]

2.3. Processo de avaliação

Objetivando analisar a capacidade das ferramentas em distinguir entre os textos escritos pela IA dos escritos por humanos, incorporamos 10 amostras de controle que foram escritas por pessoas reais para avaliar as respostas falso-positivas dos detectores de IA. Essas amostras foram obtidas de manuscritos e relatórios técnicos de discentes formados dos cursos de engenharia civil e enfermagem, não havendo uso de inteligência artificial para auxílio da escrita desses documentos, evitando possível viés de avaliação.

A fim de quantificar o desempenho dos detectores de IA empregamos métricas comumente utilizadas em aprendizado de máquina para classificação binária, sendo o MAE (do inglês, *Mean Absolute Error*) e RMSE (do inglês, *Root Mean-Square Error*), expressadas respectivamente nas equações (1) e (2). O MAE calcula a média da diferença absoluta entre os valores previstos pelo algoritmo e os valores reais, nesse contexto mede o erro médio da ferramenta em termos absoluto sem amplificar erros maiores. O RMSE calcula o erro quadrático médio e depois tira a raiz quadrada, atribuindo pesos maiores a erros grandes, ideal para identificar inconsistências e variações extremas nos resultados. Se uma ferramenta cometer alguns erros muito altos (exemplo: classificar um texto humano como 100% IA), o RMSE será bem maior que o MAE, indicando que há problemas sérios de detecção. Se o RMSE e o MAE forem parecidos, significa que os erros são relativamente uniformes, sem valores extremamente errados.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - p_i| \qquad (1)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - p_i)^2} \qquad (2)$$

Onde nas equações (1) e (2), n representa o número de amostras, sendo y_i o valor observado para a amostra i, e p_i representa o valor previsto pelo algoritmo na amostra i.

3. Análise dos Resultados

Inicialmente submetemos todos os manuscritos gerados pelo ChatGPT, DeepSeek V3, Gemini 1.5 Flash, Gemini 2.0 Flash Experimental e as amostras de controle escritas por humanos para os cinco detectores de IA: ZeroGPT, JustDone AI Detector, Writer AI Detector, Seo AI Detector e Summarizer AI Detector. Os testes ocorreram entre o período de 12 de janeiro de 2025 e 22 de fevereiro de 2025, empregando as respectivas versões dos softwares disponibilizadas durante este intervalo de tempo. Efetuamos 50 casos de teste para cada uma das ferramentas totalizando 250 diagnósticos.

A análise dos resultados das cinco ferramentas de detecção de conteúdo gerado por IA revelou uma variabilidade significativa no desempenho dos algoritmos em identificar os manuscritos de IA, vide a Tabela 3. O ZeroGPT demonstrou alta precisão em identificar ensaios gerados pelos LLMs, alcançando uma confiança média de 96%. Vale destacar que todas as pontuações obtidas variaram entre 91,42% e 100% na análise de textos sintéticos produzidos pelos modelos do ChatGPT, Gemini e DeepSeek, os dados sugerem sua eficácia em identificar conteúdos produzidos por estes grande modelos generativos.

Justdone AI Detector e Seo AI Detector apresentaram resultados semelhantes na detecção de textos gerados por IA nos manuscritos sintéticos, obtendo respectivamente uma média de precisão 85,33% e 80,50%. Porém Seo AI Detector teve desempenho inconsistente em identificar os ensaios do ChatGPT e DeepSeek, com médias inferiores a 70% para ambos os modelos, enquanto Justdone AI Detector teve constância média superior de 80% neste mesmo cenário.

Embora os desenvolvedores do Summarizer AI Detector afirmem que sua ferramenta é capaz de detectar conteúdos gerados por qualquer modelo generativo em diversos tipos de documentos, incluindo ensaios acadêmicos e teses, os resultados demonstram uma eficácia limitada. Em média, o Summarizer AI Detector identificou apenas 55,65% do conteúdo sintético. Ao analisar exemplos específicos da Tabela 3, como os ensaios GPT-4o Mini_8, Gemini 1.5 Flash_1, Gemini 2.0 Flash Experimental_1 e DeepSeek-V3_6, observa-se que, em nesses casos, o programa detectou no máximo 25% de texto artificial, mesmo em documentos completamente gerados por IA.

Tabela 3. Comparação do resultado dos detectores de IA em identificar os textos científicos gerados pelos grandes modelos de linguagem.

	Percentual de conteúdo gerado por IA identificado				
Manuscrito gerado por inteligência artificial	Zero GPT	Justdone AI Detector	Writer AI Detector	Seo AI Detector	Summarizer AI Detector
GPT 4o mini_1	92,46%	95%	17%	29%	39%
GPT 4o mini_2	97,20%	85%	15%	35%	18%
GPT 4o mini_3	97,30%	100%	6%	46%	93%
GPT 4o mini_4	99,14%	81%	17%	100%	66%
GPT 4o mini_5	99,46%	78%	12%	56%	52%
GPT 4o mini_6	91,96%	74%	9%	70%	17%
GPT 4o mini_7	93,94%	87%	16%	78%	50%
GPT 4o mini_8	91,96%	88%	9%	70%	17%
GPT 4o mini_9	98,80%	96%	17%	100%	89%
GPT 4o mini_10	97,53%	92%	13%	100%	59%
Gemini 1.5 flash_1	97,43%	81%	10%	100%	15%

Gemini 1.5 flash_2	97,14%	87%	15%	100%	30%
Gemini 1.5 flash_3	97,58%	86%	6%	100%	72%
Gemini 1.5 flash_4	95,56%	77%	14%	100%	72%
Gemini 1.5 flash_5	96,34%	90%	11%	83%	61%
Gemini 1.5 flash_6	91,42%	97%	9%	100%	30%
Gemini 1.5 flash_7	99,60%	85%	9%	100%	49%
Gemini 1.5 flash_8	93,32%	77%	7%	100%	49%
Gemini 1.5 flash_9	97,40%	78%	15%	100%	81%
Gemini 1.5 flash_10	98,22%	98%	10%	100%	65%
Gemini 2.0 flash exp._1	96,79%	83%	11%	100%	14%
Gemini 2.0 flash exp._2	98,97%	72%	13%	100%	78%
Gemini 2.0 flash exp._3	98,25%	76%	10%	49%	69%
Gemini 2.0 flash exp._4	94,97%	91%	5%	62%	70%
Gemini 2.0 flash exp._5	97,57%	88%	10%	86%	39%
Gemini 2.0 flash exp._6	94,25%	81%	12%	100%	14%
Gemini 2.0 flash exp._7	100%	85%	13%	100%	60%
Gemini 2.0 flash exp._8	96,90%	75%	12%	100%	83%
Gemini 2.0 flash exp._9	99,06%	87%	15%	100%	92%
Gemini 2.0 flash exp._10	99,10%	77%	7%	100%	63%
DeepSeek-V3_1	94,44%	81%	11%	28%	51%
DeepSeek-V3_2	98,14%	83%	9%	57%	52%
DeepSeek-V3_3	98,26%	89%	8%	78%	70%
DeepSeek-V3_4	96,67%	85%	4%	50%	51%
DeepSeek-V3_5	98,49%	82%	15%	55%	66%
DeepSeek-V3_6	98,44%	74%	10%	100%	21%
DeepSeek-V3_7	96,22%	98%	9%	33%	75%
DeepSeek-V3_8	99,04%	83%	5%	57%	82%
DeepSeek-V3_9	98,72%	99%	11%	100%	91%
DeepSeek-V3_10	98,31%	92%	15%	100%	61%
Média aritmética	**96,91%**	**85,33%**	**11,05%**	**80,55%**	**55,65%**

Em contrapartida, o Writer AI Detector foi o menos eficaz, com taxas de detecção muito baixas em todos os documentos feitos com chatbots, identificando apenas 17% em dois manuscritos gerados com o ChatGPT modelo GPT 4o mini, nos outros LLMs a situação foi ainda pior, pontuando entre o intervalo de 4% e 15%. Toda via, a empresa responsável pela ferramenta declara que o produto Writer AI Detector tem desempenho afetado pelo idioma, com melhor performance na deteção de texto em inglês do que em outros idiomas [Writer 2024].

Os resultados apresentados até o momento fornecem informações relevantes sobre a eficácia dos detectores de IA na identificação de conteúdo gerados por máquina. Embora o ZeroGPT se destaque como a ferramenta mais confiável, seguido pelo JustDone AI Detector e pelo SEO AI Detector, os dados obtidos não validam completamente a precisão dessas ferramentas. Diante da necessidade de verificar se os detectores de IA são realmente capazes de distinguir entre textos gerados por LLM e textos escritos por humanos, realizamos um diagnóstico nas amostras de controle, conforme demonstrado na Tabela 4.

A análise dos manuscritos escritos por humanos revelou taxas médias de detecção de IA mais baixas em todas as ferramentas, o que era esperado. No entanto, os resultados identificação médios obtidos pelo o Justdone AI Detector e o Writer AI Detector levantam

sérias preocupações quanto à eficácia de ambos na distinção entre texto gerado por inteligência artificial e conteúdo escrito por humanos.

Tabela 4. Comparação do resultado das ferramentas em identificar conteúdo gerado por IA em manuscritos criados por humanos.

Manuscrito escrito por humano	Percentual de conteúdo gerado por IA detectado				
	Zero GPT	Justdone AI Detector	Writer AI Detector	Seo AI Detector	Summarizer AI Detector
Humano_1	43,49%	80%	0%	43%	34%
Humano_2	14,47%	77%	0%	27%	91%
Humano_3	13,65%	100%	0%	9%	30%
Humano_4	0%	88%	0%	18%	80%
Humano_5	16,20%	86%	0%	4%	89%
Humano_6	28,56%	75%	0%	100%	43%
Humano_7	65,15%	70%	0%	19%	18%
Humano_8	90,64%	81%	0%	8%	0%
Humano_9	0%	95%	0%	70%	56%
Humano_10	23,19%	88%	2%	19%	12%
Média aritmética	**29,54%**	**84%**	**0,20%**	**31,70%**	**45,30%**

Observou-se que a pontuação média do Justdone AI Detector ao avaliar ensaios escritos por humanos foi de 84%, um valor surpreendentemente alto, que praticamente se equipara à sua média de 85,33% na identificação de textos gerados por IA em documentos produzidos por modelos de linguagem. Essa pequena diferença de 1,33% entra as médias sugere fortemente que o Justdone AI Detector não possui a capacidade de distinguir de forma confiável entre as duas categorias de texto.

De forma análoga, os resultados do Writer AI Detector também suscitam preocupações, embora de natureza diferente. Enquanto sua pontuação média de 0,20% em identificar texto de IA em documentos escritos por humanos é baixa, o que seria desejável, sua pontuação média de apenas 11,05% na identificação de texto de IA em documentos gerados por inteligência artificial demonstra uma baixa sensibilidade na detecção do conteúdo alvo. Essa combinação de resultados sugere que o Writer AI tem o viés e tendencia de identificar qualquer texto como sendo escrito por uma pessoa real, independentemente da origem do conteúdo. Essa quantidade de falsos positivos torna a ferramenta ineficaz em identificar o conteúdo gerado em português por um chatbot.

O Summarizer AI Detector apresentou dados com uma limitação significativa em sua capacidade de distinguir entre textos escritos por humanos e aqueles gerados por inteligência artificial. Com uma pontuação média de 55,65% na detecção de texto de IA em manuscritos efetivamente gerados por IA e uma pontuação média de 45,30% em textos escritos por humanos, a proximidade dessas médias indica que o algoritmo subjacente não consegue efetivamente separar as duas categorias de conteúdo. Uma diferença de apenas cerca de 10 pontos percentuais entre a detecção de texto de IA real e a detecção em textos humanos sugere que a ferramenta está essencialmente atribuindo pontuações semelhantes a ambos os tipos de escrita.

Os resultados médios de detecção de IA em textos originalmente humanos para o ZeroGPT e Seo AI Detector revelam os melhores resultados entre os cinco detectores, porém eles ainda apresentam um padrão de desempenho preocupante. Tanto o ZeroGPT quanto o SEO AI Detector apresentaram taxas consideráveis de falsos positivos, classificando erroneamente textos humanos como gerados por IA em aproximadamente

30% dos casos de testes (29,54% e 31,70%, respectivamente). Embora essa diferença de 2,16% entre os algoritmos seja pequena, quando analisamos a capacidade de identificar corretamente textos sintéticos, o ZeroGPT demonstrou desempenho de 16,36% superior em comparação com o SEO AI Detector. Essa disparidade sugere que o ZeroGPT se comprovou como a ferramenta mais precisa na detecção de conteúdo gerado por IA em ambas as categorias dentre todos os cinco detectores de IA analisados. Porém ele ainda compartilha como as outras ferramentas a limitação crítica de produzir falsos positivos das análises de texto humano.

O cálculo das métricas erro MAE (Erro Médio Absoluto) da Figura 2 e RMSE (Raiz do Erro Quadrático Médio) da Figura 3, complementam a análise das médias de detecção, fornecendo uma medida da precisão e da consistência das previsões de cada ferramenta. Um MAE baixo indica que, em média, as previsões estão próximas dos valores reais, enquanto um RMSE baixo sugere que menor ocorrência de grandes picos de erros.

A análise quantitativa do MAE e RMSE do Writer AI Detector confirma sua baixa eficácia na detecção de conteúdo gerado por IA. Para textos sintéticos, os valores de MAE e RMSE foram consistentemente altos em todos os modelos testados (Figura 2 e 3), com erro médio absoluto próximo de 89%. Embora os valores baixos para manuscritos humanos (0,20% de MAE e 0,63% de RMSE) indiquem poucos falsos positivos, essa vantagem é anulada pela incapacidade do detector de identificar textos de IA. Os resultados sugerem que a ferramenta tende a classificar qualquer texto como humano, tornando-a ineficaz para detecção de IA em português. Esse resultado reforça a hipótese de [Candido et al. 2024] de que o Writer AI Detector não foi treinado para avaliar textos em português, sendo completamente ineficaz para o propósito avaliado.

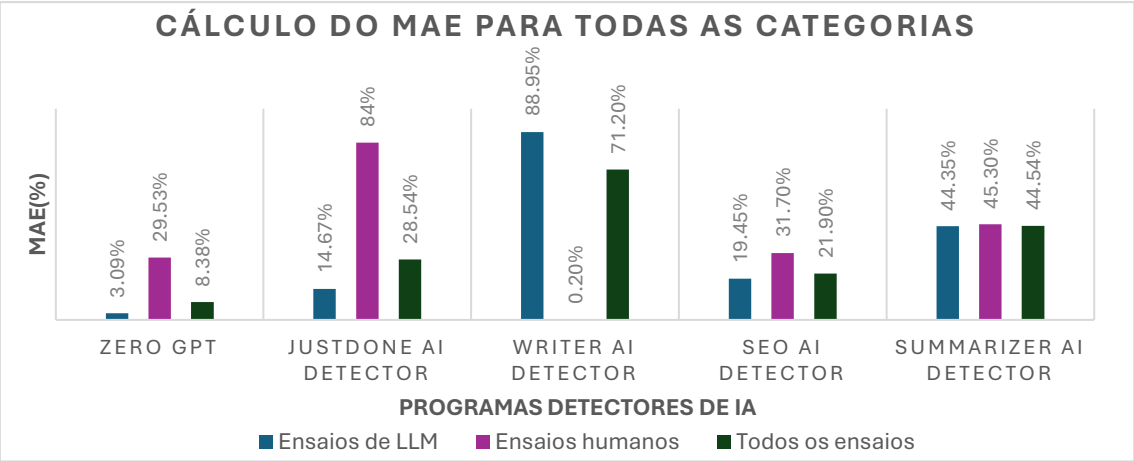

Figura 2. Gráfico de comparação do MAE das ferramentas para identificar IA em ensaios gerados por LLMs, ensaios escritos por humanos e todas as amostras.

Os valores de RMSE e MAE do JustDone AI Detector variam em torno de 84% (Figura 2 e 3) em textos humanos, isto implica que a ferramenta está consistentemente sinalizando conteúdo genuíno como sendo de IA, indicando uma taxa de falsos positivos inaceitavelmente elevada. Contradizendo completamente as afirmações da documentação oficial [JustDone 2025], que alega promover a integridade acadêmica. Na prática, quando aplicado à detecção de IA em ensaios científicos em português, a ferramenta mostrou-se completamente ineficaz. Tendo isso em vista, pode existir a hipótese que a ferramenta

pode estar super ajustada a padrões linguísticos do inglês ou carece de parâmetros adequados para analisar adequadamente textos acadêmicos em outras línguas, no nosso caso o português, assim como constatado no Writer AI Detector.

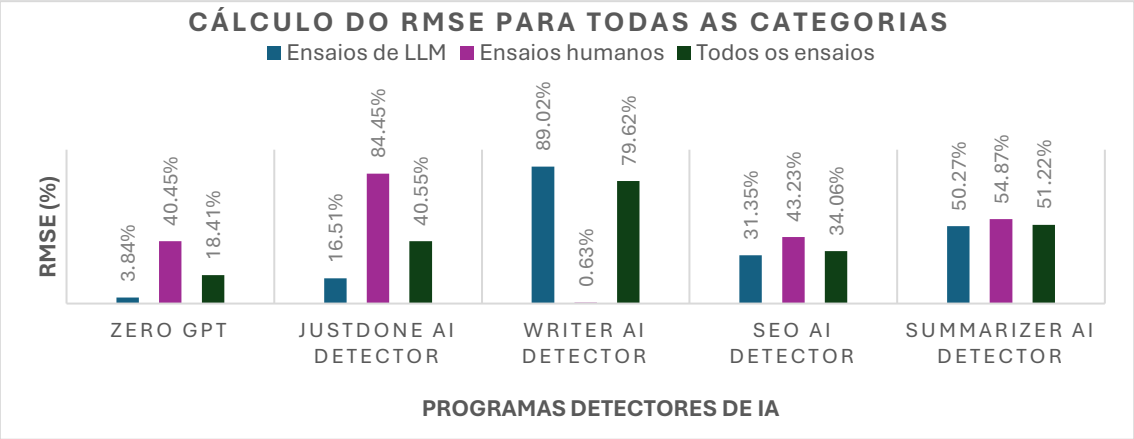

Figura 3. Gráfico de comparação do RMSE das ferramentas para identificar IA em ensaios gerados por LLMs, ensaios escritos por humanos e todas as amostras.

O Summarizer AI Detector obteve 44,54% e 51,22% de MAE e RMSE para todas as amostras de teste, o valor elevado do RMSE em comparação ao MAE revela que a ferramenta está atribuindo uma pontuação instável e incoerente entre as categorias de textos. Como por exemplo identificar 14% de IA no ensaio Gemini 2.0 flash exp._6, ao mesmo tempo que no Gemini 2.0 flash exp._9 detectou 92% de conteúdo artificial, havendo outras situações similares. Essa falta de distinção clara implica que a ferramenta comete erros graves em ambas as direções: produzindo numerosos falsos positivos (textos humanos classificados como IA) e falsos negativos (conteúdo de IA não detectado). Portanto, uma ferramenta pouco confiável para essa finalidade específica.

O Seo AI Detector obteve o segundo melhor desempenho entre as ferramentas avaliadas, porém as medidas métricas de RMSE e MAE revela um padrão de desempenho misto. Nos manuscritos escritos por pessoas reais, observamos um MAE de 31,70% e um RMSE de 43,23%, indicando uma taxa considerável de falsos positivos, onde textos humanos foram erroneamente classificados como contendo uma porcentagem significativa de conteúdo gerado por IA. Quando observamos a Tabela 4 percebemos o motivo do RMSE ser mais elevado, o Seo AI Detector identificou entre 4% e 18% de IA em 6 amostras humanas, porém chegou a acusar a amostra Humano_6 de conter conteúdo 100% gerado por IA. Por outro lado, ao analisar os ensaios sintéticos, o detector teve um erro médio absoluto de 19,45% em relação ao esperado (100%), sugerindo capacidade razoável de detecção de conteúdo gerado por inteligência artificial. Em contrapartida, os valor alto de RMSE (31,35%) indicam que uma margem de erro considerável, porém estes erros estão mais propensos a ocorrer nos textos gerados pelo ChatGPT 4o mini e DeepSeek-V3, como evidenciado na Tabela 3, a ocorrência de variação de detecção de IA nesses modelos.

Ao considerar todos os manuscritos avaliados, o MAE geral foi de 21,09% e o RMSE de 34,06%. Esses resultados confirmam que, embora o Seo AI Detector demonstre alguma capacidade de identificar conteúdo sintético, ele também apresenta uma propensão a classificar erroneamente textos humanos como sendo gerados por IA, levantando preocupações sobre sua precisão e confiabilidade geral.

A análise das métricas de erro do ZeroGPT consolida sua posição como o detector de IA mais eficaz em relação as demais ferramentas avaliadas, reafirmando o resultado discutido anteriormente das médias aritméticas da Tabela 3. Nos textos de autoria humana, registrou um MAE de 29,53% e RMSE de 40,45%, indicando os melhores resultados comparativos em termos de redução de falsos positivos, porém. A disparidade do RMSE ocorre em decorrência de 7 manuscritos humanos terem tido uma taxa de detecção de IA variando entre 0% e 28,56% em comparação aos outros 3 que variaram entre 43% e 90%. Em contraste, o desempenho do ZeroGPT na detecção de conteúdo gerado por inteligência artificial foi notável, com um MAE de apenas 3,09% e um RMSE de 3,84%. Esses valores excepcionalmente baixos demonstram a alta precisão e consistência da ferramenta na identificação de textos sintéticos produzidos pelos diferentes modelos de linguagem testados, apresentando resultados semelhantes aos trabalhos de [Candido et al. 2024; Malik and Amjad 2025].

Esses resultados gerais reforçam a afirmação realizada pelos pesquisadores [Weber-Wulff et al. 2023] de que os detectores de IA falham, não são completamente precisos ou confiáveis. Embora algumas ferramentas tenham se mostrado mais eficazes, como o ZeroGPT, a inconsistência geral levanta dúvidas sobre sua confiabilidade como única forma de determinar a autoria de um texto, entrando em acordo com o trabalho de [Malik and Amjad 2025]. [Ladha et al. 2023] explicam que é impossível qualquer programa criado para detectar texto generativo identificar com precisão todo o conteúdo gerado por LLM, enfatizam o julgamento humano para a tomada de decisão e garantir a integridade acadêmica. Além disso, os achados dos pesquisadores corroboram com os nossos resultados, demonstrando que os algoritmos tendem a classificar erroneamente textos gerados por humanos como sendo de IA, o que pode levar a acusações injustas.

4. Considerações Finais e Trabalhos Futuros

A rápida evolução da inteligência artificial generativa e sua crescente aplicação na produção de conteúdo científico suscitam importantes debates sobre integridade acadêmica e a necessidade de ferramentas eficazes para detectar o uso indevido dessas tecnologias. Este estudo investigou a capacidade de cinco detectores de IA, inclusos o ZeroGPT, JustDone AI Detector, Writer AI Detector, Seo AI Detector e Summarizer AI Detector, em identificar textos científicos em português gerados pelos diferentes modelos de linguagem ChatGPT, Gemini e DeepSeek, em comparação as amostras de controle escritas por humanos.

Os resultados da análise revelaram uma variabilidade significativa no desempenho das ferramentas. O ZeroGPT destacou-se como o detector mais preciso na identificação de conteúdo gerado por IA, apresentando os menores erros médios (MAE e RMSE) para textos sintéticos e a menor taxa de falsos positivos em comparação com as outras ferramentas. No entanto, mesmo o ZeroGPT demonstrou a limitação de classificar erroneamente textos humanos como sendo gerados por IA, apresentando MAE de 29,53% e RMSE de 40,45%. As demais ferramentas apresentaram limitações significativas, como alta taxa de falsos positivos do JustDone AI Detector e Seo AI Detector. Além da baixa sensibilidade na detecção de textos de IA em português e inconsistência, assim como a dificuldade em distinguir entre os tipos de conteúdo apresentado pelos Writer AI Detector e Summarizer AI Detector.

As nossas descobertas fortemente sugerem que não existem uma solução fácil para a detecção de manuscritos científicos gerados por máquina, tais quais os detectores de IA, e talvez nem sequer possam existir [Weber-Wulff et al. 2023]. Além disso, os resultados servem como um alerta para a comunidade acadêmica, mostrando que as ferramentas atuais têm limitações significativas, especialmente em idiomas não ingleses, e não devem ser utilizadas como único critério para avaliação de autoria.

O estudo também apresenta limitações que devem ser consideradas. A amostragem, composta por 50 manuscritos (40 gerados por IA e 10 humanos), pode não ser suficiente para generalizar os resultados. Além disso, a pesquisa focou em ferramentas gratuitas e publicamente disponíveis, excluindo soluções comerciais que poderiam oferecer desempenho diferente. Outro fator relevante é a rápida evolução dos modelos de linguagem, que pode tornar os resultados obtidos menos aplicáveis a versões mais recentes. Por fim, o estudo concentrou-se no português, e suas conclusões podem não ser diretamente extrapoladas para outros idiomas.

Para trabalhos futuros, esta pesquisa tem como pretensão a expansão de modelos de IA, bem como a inclusão de novas categorias de amostras: ensaios gerados por IA mas editados por humanos, textos sintéticos humanizados por IA, manuscritos acadêmicos feitos por pessoas e com texto melhorado por LLM. Além disso, incluir novos detectores de IA a fim de obter uma análise mais ampla e assertiva desses algoritmos.

5. Agradecimentos

Este trabalho teve financiamento da Fundação de Amparo à Pesquisa do Estado de Alagoas, por meio do Programa Institucional de Bolsas de Iniciação em Desenvolvimento Tecnológico ofertado pela Pró-reitoria de Pesquisa, Pós-graduação do Instituto Federal de Alagoas, aprovado no edital n.° 06/2024 PRPPI/IFAL sob o código PVE1235-2024.

6. Referencias

Almeida, F. das C. F., Aguiar, Y. P. C. and Magalhaes, R. M. (16 oct 2023). Você sabe diferenciar um resumo escrito por humanos do gerado pelo ChatGPT? In *Workshop sobre Aspectos da Interação Humano-Computador na Web Social (WAIHCWS)*. . SBC. https://sol.sbc.org.br/index.php/waihcws/article/view/26864, [accessed on Mar 20].

Cabanac, G. and Labbé, C. (2021). Prevalence of nonsensical algorithmically generated papers in the scientific literature. *Journal of the Association for Information Science and Technology*, v. 72, n. 12, p. 1461–1476.

Candido, L. S., Barbosa, C. A. de M. and Costa, E. J. H. (21 jul 2024). Análise de ferramentas para detecção de textos científicos gerados por Inteligência Artificial (ChatGPT). In *Workshop sobre as Implicações da Computação na Sociedade (WICS)*. . SBC. https://sol.sbc.org.br/index.php/wics/article/view/29518, [accessed on Mar 22].

Chen, B., Zhang, Z., Langrené, N. and Zhu, S. (5 sep 2024). Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review. . arXiv. http://arxiv.org/abs/2310.14735, [accessed on Dec 30].

DeepSeek-AI, Liu, A., Feng, B., et al. (18 feb 2025). DeepSeek-V3 Technical Report. . arXiv. http://arxiv.org/abs/2412.19437, [accessed on Mar 13].

Elkhatat, A. M., Elsaid, K. and Almeer, S. (1 sep 2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, v. 19, n. 1, p. 17.

Gil, A. C. (2017). *Como elaborar projetos de pesquisa*. 6. ed ed. São Paulo: Atlas.

Google DeepMind (12 mar 2025). Gemini 2.0. https://deepmind.google/technologies/gemini/, [accessed on Mar 13].

Hammad, M. (1 mar 2023). The Impact of Artificial Intelligence (AI) Programs on Writing Scientific Research. *Annals of Biomedical Engineering*, v. 51, n. 3, p. 459–460.

Intelligent.com (23 jan 2023). Nearly 1 in 3 College Students Have Used ChatGPT on Written Assignments. https://www.intelligent.com/nearly-1-in-3-college-students-have-used-chatgpt-on-written-assignments/, [accessed on Mar 19].

JustDone (2025). AI Detector | JustDone. https://justdone.ai/ai-detector, [accessed on Mar 18].

Ladha, N., Yadav, K. and Rathore, P. (25 oct 2023). AI-Generated Content Detectors: Boon or Bane for Scientific Writing. *Indian Journal of Science and Technology*, v. 16, n. 39, p. 3435–3439.

Liyanage, V., Buscaldi, D. and Nazarenko, A. (jun 2022). A Benchmark Corpus for the Detection of Automatically Generated Text in Academic Publications. [N. Calzolari, F. Béchet, P. Blache, et al., Eds.]In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. . European Language Resources Association. https://aclanthology.org/2022.lrec-1.501/, [accessed on Mar 17].

Malik, M. A. and Amjad, A. I. (12 jan 2025). AI vs AI: How effective are Turnitin, ZeroGPT, GPTZero, and Writer AI in detecting text generated by ChatGPT, Perplexity, and Gemini? *Journal of Applied Learning and Teaching*, v. 8, n. 1.

OpenAI (18 jul 2024). GPT-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, [accessed on Dec 23].

Seo.Ai (2024). Free AI Content Detector. https://seo.ai/tools/ai-content-detector, [accessed on Mar 18].

Sullivan, M., Kelly, A. and McLaughlan, P. (20 mar 2023). ChatGPT in higher education: Considerations for academic integrity and student learning. *Journal of Applied Learning and Teaching*, v. 6, n. 1, p. 31–40.

Summarizer.org (2025). Free AI Detector. https://www.summarizer.org/ai-detector, [accessed on Mar 18].

Team, G., Georgiev, P., Lei, V. I., et al. (16 dec 2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. . arXiv. http://arxiv.org/abs/2403.05530, [accessed on Mar 13].

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., et al. (25 dec 2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, v. 19, n. 1, p. 26.

Writer (jan 2024). Writer AI content detector disclaimer. https://writer.com/disclaimer-aicd/, [accessed on Mar 23].

Writer (2025). AI content detector. https://writer.com/ai-content-detector/, [accessed on Mar 18].

ZeroGPT.net (2025). ZeroGPT - Detector de Chat GPT | Verificador de Chat GPT. https://zerogpt.net/pt/ai-content-detector, [accessed on Mar 18].