Análise de ferramentas para detecção de textos científicos gerados por Inteligência Artificial (ChatGPT)

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

[1]Curso de Bacharelado em Engenharia Civil – Instituto Federal de Alagoas (IFAL), Campus Maceió

{lsc7,camb1}@aluno.ifal.edu.br, esdras.costa@ifal.edu.br

Abstract. *This article analyzes the performance of four tools for detecting articles generated by artificial intelligence (AI), Writer AI Content Detector, GPT Zero, Zero GPT and Quillbot AI Content Detector. The detectors are evaluated on a dataset of 30 articles generated in Portuguese on ChatGPT 3.5 and 10 human-written text. The results obtained from the experiment show the effectiveness of Zero GPT in categorizing content produced by machines, achieving an accuracy of 97.50%, a performance comparable to the state-of-the-art. However, challenges were encountered in identifying artificially generated texts, and the research suggests the need for continuous improvements and explorations to improve the process of identifying artificially generated content.*

Abstract. *Este artigo analisa o desempenho de quatro ferramentas para detecção de artigos gerados por inteligência artificial (IA), Writer AI Content Detector, GPT Zero, Zero GPT e Quillbot AI Content Detector. Os detectores são avaliados em um conjunto de dados de 30 artigos gerados em português no ChatGPT 3.5 e 10 artigos escritos por pessoas reais. Os resultados obtidos do experimento mostram a eficácia do Zero GPT em categorizar conteúdo produzido por máquinas, alcançando a acurácia de 97,50%, desempenho equiparável ao estado da arte. No entanto, foram encontrados desafios em identificar textos gerados artificialmente, a investigação sugere a necessidade de melhorias e explorações contínuas para aperfeiçoar o processo de identificação de conteúdo gerados artificialmente.*

1. Introdução

O recente progresso em grandes modelos de linguagem (LLM) enviesado pelo amplo sucesso do ChatGPT, plataforma de *chatbot* da empresa OpenAI, fomentou a criação de máquinas de inteligência artificial generativas. Estes modelos são capazes de produzir materiais como textos, imagens e vídeos com alta qualidade de concepção, sendo difícil distinguir entre os conteúdos gerados por máquinas e dos criados por pessoas [Jiameng e et al. 2023]. O ChatGPT é uma variação do GPT-3 (Generative Pre-trained Transformer 3) especificamente desenvolvido para geração de textos semelhantes aos de humanos em forma de bate-papo comum [BROWN e et al. 2020]. Desde seu lançamento, o GPT-3 vem sendo usado para uma variedade de aplicações relacionadas ao processamento de linguagem natural, incluindo tradução de linguagens, geração de textos, resumos de documentos e elaboração de artigos em jornais.

A capacidade do ChatGPT em compreender e responder as entradas (*inputs*) de linguagem natural de forma coerente tem atraído bastante atenção com a sua habilidade de articulação, principalmente na língua inglesa. Conseguindo passar em provas de seleção de médicos, advogados e programas de pós-graduação [FERREIRA e et al. 2023], ou até mesmo conseguir influenciar e coagir seres humanos a partir de críticas falsas [Adelani e et al. 2020]. Isto vem levantando preocupações sobre os potenciais usos e impactos da inteligência artificial (IA) no campo de processamento de linguagens, tendo em vista que esses modelos podem ser utilizados para atividades maliciosas e criminosas.

Ademais, não existem dúvidas que o emprego de IA no aumento da produtividade dos mais diversos setores da sociedade ocorrerá num futuro próximo, porém existe a necessidade de debater com estudantes e pesquisadores a cerca dos benefícios e ética desses modelos generativos. Como contrapartida, [Weber-Wulff e et al. 2023] explicam que muitas educações institucionais proibiram o uso do ChatGPT, inclusive conferências e periódicos científicos explicitamente condenaram a submissão de manuscritos contendo conteúdo gerado por máquina. Porém, convém mencionar que o uso de LLM no âmbito acadêmico não é novidade ou somente surgiu com os avanços tecnológicos recentes. Estima-se que em 2014 ao menos 120 artigos de periódicos científicos foram identificados como "sem sentido" e denunciados como ilegítimos, segundo os dados levantados por [Cabanac e Labbé 2021]. Durante este episódio, os editores se retrataram publicamente e removeram o conteúdo existente, a publicação destes artigos falsos afetou negativamente a credibilidade acadêmica dessas revistas e ocasionou em reavaliações da qualidade do periódico.

Existem diversas controversas éticas sobre o uso de inteligência artificial generativa no âmbito acadêmico, ainda mais quando falamos sobre metodologia científica e plágio nas pesquisas. Afinal, artigos científicos devem fornecer novas ideias aos leitores, porém o conteúdo produzido artificialmente pela máquina tem como origem as informações decorrentes de estudos publicados que serviram como base para sua resposta. Além da própria discussão moral se essas produções poderiam ser publicadas como propriedade intelectual de alguma pessoa, tendo em vista que o manuscrito teria sido gerado por IA.

Diante desta problemática, pesquisadores como [Elkhatat e et al. 2023] e [Weber-Wulff e et al. 2023] investigam ferramentas para identificar textos gerados por máquina (TGM) na língua inglesa. Enquanto o trabalho realizado por [Chaka e et al 2023] verifica soluções para reconhecer TGM em diferentes linguagens como alemão, francês e espanhol. Tendo isto em vista, o presente estudo tem como objetivo procurar e analisar a verdadeira eficácia de algoritmos utilizados para detectar manuscritos em português criados por inteligência artificial.

2. Metodologia

O vigente trabalho baseia-se nos conceitos definidos por [Gil 2017], atuando como uma pesquisa aplicada. Isso porque o estudo tem como objetivo analisar a eficácia das ferramentas utilizadas para detectar textos gerados por máquina no âmbito da produção científica, visando a aplicação prática dessas ferramentas para resolver o problema específico e contribuir para a prevenção do uso indevido de conteúdo gerado artificialmente.

2.1. Criação do Banco de Dados

O escopo desta pesquisa é analizar a eficácia e precisão de ferramentas produzidas para a classificação de conteúdo científico criados por IA na linguagem portuguesa. Em ordem de conduzir o experimento criamos um banco de dados contendo artigos reais escritos por humanos, e de manuscritos científicos gerados com ChatGPT 3.5.

A definição do banco de dados em ser composto por duas categorias de textos tem como fundamento proporcionar uma investigação robusta das ferramentas que compoem o objeto deste estudo, permitindo a avaliação de seu comportamento perante ambos cenários. As amostras foram categorizadas em positivas e negativas, as positivas correspondem a TGM e negativas a manuscritos de humano.

2.1.1. Obtenção de artigos escritos por humanos

A criação do subconjunto de dados de artigos reais se baseou na coleta de produções científicas oriundas de domínio público, levando em consideração que pesquisas atuais podem apresentar conteúdo tendencioso e de passível elaboração por IA. Obras em domínio público permitem maior flexibilidade na utilização de conteúdos protegidos por direitos autorais, de modo que os conteúdos sejam utilizados amplamente, sem que as leis de proteção à propriedade intelectual sejam infringidas.

2.1.2 Geração de manuscritos científicos artificiais

A capacidade de gerar um texto de alta qualidade a partir de um modelo generativo de texto depende da boa definição e estruturação do prompt de comando [Kıyak 2023]. [White e et al. 2023] explicam como a engenharia do prompt afeta a capacidade de se comunicar efetivamente com os grandes modelos de linguagem, os prompts são instruções dadas a um LLM para definir regras, processos e a lógica da produção gerada, atuando como uma forma de programação que pode personalizar a saída (output) do conteúdo criado pela inteligência artificial. Tendo isto em vista, criamos e testamos uma variedade de prompts para geração de textos científicos, a partir de tentativa e erro, considerando que não existe uma fórmula matemática para obter o comando desejado.

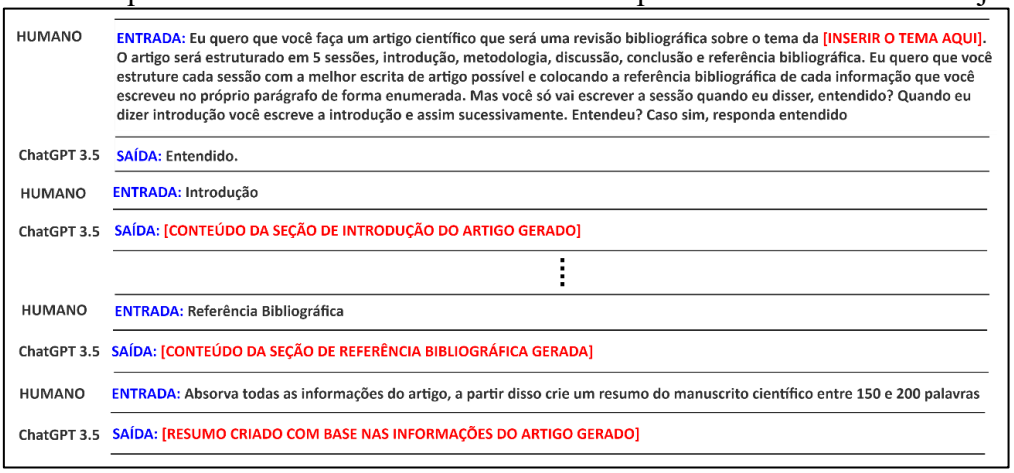

Figura 1. Simulação da geração de artigos científicos com o ChatGPT 3.5

A criação do comando de entrada, da presente pesquisa, tivera como base a sua capacidade de influenciar as interações subsequentes com o ChatGPT 3.5 e o conteúdo gerado pelo modelo, através da definição de diretrizes estabelecidas inicialmente na

conversa com o *chatbot*. O input desenvolvido define o contexto da comunicação e explica ao LLM qual informação é importante e qual a forma de saída do conteúdo deve ser, conforme demonstrado na figura 1.

O fluxo para geração do conteúdo inicia-se na criação de uma nova conversa com o GPT 3.5, inserindo a cadeia de instruções definidas na figura 1 em forma de bate-papo. O conteúdo de saída gerado pelo chatbot é extraído e organizado em novo arquivo de documento de texto, representando o artigo recém-escrito. Após a geração do manuscrito, apaga-se completamente o chat e inicia-se uma nova conversa com o GPT 3.5, dando importância que o histórico do bate-papo pode influenciar negativamente na elaboração de novos textos.

2.2. Ferramentas de detecção de texto gerado por máquina

Existem diversas ferramentas para detectar conteúdo gerado por inteligência artificial, algumas mais eficientes do que outras, conforme [Adelani e et al. 2020] demonstra ao utilizar o mesmo banco de dados em conjunção com diferentes técnicas de detecção de textos gerados por máquina. A seleção dos classificadores de conteúdo em português gerado por IA fora realizada por uma busca extensiva de trabalhos científicos relacionados a solução deste problema, assim como efetuamos pesquisas através de buscadores online de sites da internet. Quatro ferramentas de classificação foram escolhidas para avaliar sua habilidade de analisar manuscritos: Writer AI Content Detector [WRITER 2024], GPT Zero [GPTZero 2023], Zero GPT [ZeroGPT 2023] e QuillBot AI Content Detector [QuillBot 2017].

As versões gratuitas destes sistemas impõem limitações na quantidade de palavras e/ou caracteres do texto a ser analisado, conforme apresentado na tabela 1. Em ordem de solucionar esta problemática, dividimos as amostras e computamos individualmente cada fragmento, a porcentagem geral da amostra ($\boldsymbol{P_{Geral}}$) pode ser calculada como expressado na equação 1, onde $\boldsymbol{Pi}$ é a porcentagem na fração $\boldsymbol{i}$ e $\boldsymbol{n}$ é o número total de fragmentos.

$$P_{geral} = \left(\frac{1}{n} \times \sum_{i=1}^{n} Pi\right) \times 100 \quad (1)$$

Tabela 1. Escala de classificação para textos de máquinas e humanos

Nome da ferramenta	Tamanho mínimo	Tamanho máximo
Writer AI Detector	60 Palavras	5.000 palavras
GPT Zero	250 caracteres	5.000 caracteres
Zero GPT	Não declarado	15.000 caracteres
QuillBot AI Detector	80 Palavras	1.200 palavras

Ainda convém mencionar que as ferramentas de detecção exibem seus resultados em representações distintas umas das outras, dificultando a análise comparativa dos classificadores. As aplicações Writer AI Content Detector e GPT Zero apresentam a probabilidade de o texto ter sido escrito por uma pessoa real, enquanto os modelos da QuillBot e Zero GPT detectam a chance de o conteúdo ter sido gerado por máquina. Dada a importância da compatibilidade dos dados a fim de explorar com precisão os algoritmos, parametrizamos os dados dos resultados de acordo com a porcentagem do texto ter sido escrito por humanos, como está descrito na tabela 2, os valores dos intervalos das escalas de classificação foram selecionados tendo como base os estudos produzidos por [Elkhatat e et al. 2023] e [Weber-Wulff e et al. 2023].

Tabela 2. Escala de classificação para textos de máquinas e humanos

Texto gerado por IA (POSITIVO), a ferramenta classifica escrito como:		
Escala	**Classificação**	**Sigla**
[100% - 81%] humano	Falso negativo	FN
[80% - 61%] humano	Parcialmente falso negativo	PFN
[60% - 41%] humano	Incerto	INC
[40% - 21%] humano	Parcialmente verdadeiro positivo	PVP
[20% - 0%] humano	Verdadeiro positivo	VP
Texto escrito por humano (NEGATIVO), a ferramenta classifica escrito como:		
Escala	**Classificação**	**Sigla**
[100% - 81%] humano	Verdadeiro negativo	VN
[80% - 61%] humano	Parcialmente verdadeiro negativo	PVN
[60% - 41%] humano	Incerto	INC
[40% - 21%] humano	Parcialmente falso positivo	PFP
[20% - 0%] humano	Falso positivo	FP

2.2. Medidas de avaliação de desempenho

A análise da performance das ferramentas ocorrerá pelo critério da expressão métrica de acurácia, comumente utilizada em avaliações de classificações binárias. A acurácia (equação 2) define a proporção de classificações corretas para a quantidade total de amostras.

$$\text{Acurácia} = \frac{(VN+VP)}{(\text{VN+VP+FN+FP})} \tag{2}$$

Esta expressão matemática não leva em consideração as amostras existentes classificadas como incertas, parcialmente corretas e parcialmente incorretas no experimento. Em decorrência da solução desta problemática, a equação básica foi ajustada de acordo com a formulação proposta por [Elkhatat e et al. 2023] na equação 3. A abordagem da adaptação matemática dessa métrica tem como objetivo abranger de forma geral o número de ocorrências em que as ferramentas de classificação têm uma avaliação parcialmente correta, contabilizando-as como corretas. Os autores apresentam mais uma abordagem para calcular a acurácia, denominando-a de avaliação semi-binária, atribuindo pesos de menor pontuação a classificações parciais, de acordo com expressado na equação 4.

$$\text{Acurácia\_corrigida} = \frac{(\text{VN+PVN+VP+PVP})}{(\text{VN+VP+PVP+FN+PFN+FP+PFP+INC})} \tag{3}$$

$$\text{Acurácia\_semi\_binária} \frac{(\text{VN+0,5}\times\text{PVN+VP+0,5}\times\text{PVP})}{(\text{VN+VP+PVP+FN+PFN+FP+PFP+INC})} \tag{4}$$

3. Análise e Discussão dos Resultados

Em ordem de avaliar os algoritmos de detecção de texto gerado por máquina, criamos um banco de dados composto por 40 artigos, 30 ensaios artificiais e 10 originados de escritas reais em domínio público. Os manuscritos gerados com o GPT 3.5 possuem uma média de 1.800 palavras e 12.000 caracteres, abrangendo temas como história, educação, literatura, ciência e sociologia. Os artigos provenientes de domínio público foram selecionados de tal forma que se assemelhassem ao conteúdo e estruturas das amostras da outra categoria, tendo como objetivo padronizar as amostras dos dados e prevenir que o tamanho do texto exerça influência na disparidade dos resultados.

O experimento de teste fora realizado com as quatro ferramentas: Writer AI Content Detector, GPT Zero, Zero GPT e QuillBot AI Content Detector; durante os dias

21 de fevereiro e 21 de março de 2024. Portanto utilizou-se as respectivas versões de cada software dentre o intervalo de tempo mencionado para análise de todas as 40 amostras, totalizando 160 testes executados. O número obtido pelos classificadores refere-se ao percentual de texto detectado como humano ou de máquina, então mapeamos os valores de acordo com a escala de classificação.

A padronização da representação dos resultados de cada um destes algoritmos favorece a análise comparativa coesa, conforme o desempenho do teste contabilizado na tabela 3. A apuração demonstra que Writer AI Content Detector e GPT Zero classificam artigos gerados por máquina como escritos por humanos, somando 30 falsos negativos e parcialmente falsos negativos de 30 amostras. Em contrapartida, Zero GPT e Quillbot AI Content Detector se encaminham em direção totalmente oposta, detectando corretamente 28 e 17 verdadeiros positivos, 2 e 23 parcialmente verdadeiro positivos, respectivamente. Sobre o segundo sub-conjunto, composto por 10 manuscritos reais, não houve nenhuma grande irregularidade.

Tabela 3. Resultado de classificação dos textos

	Classificador	Resultado da Classificação												
		FN	PFN	INC	PVP	VP	Sub-Total 1	VN	PVN	INC	PFP	FP	Sub-Total 2	Total
1	Writer AI Content Detector	29	1	0	0	0	30	10	0	0	0	0	10	40
2	GPT ZERO	22	8	0	0	0	30	10	0	0	0	0	10	40
3	ZERO GPT	0	0	0	2	28	30	10	0	0	0	0	10	40
4	QuillBot AI Content Detector	0	0	0	13	17	30	9	1	0	0	0	10	40

Ao examinar os resultados obtidos dos dois primeiros algoritmos na classificação do banco de dados, verifica-se a incerteza em distinguir corretamente a categoria de TGM. Para ambos os casos, pode-se presumir duas possibilidades: (i) os softwares não foram treinados para detectar textos de IA na linguagem portuguesa; (ii) os algoritmos foram superados pelo avanço do ChatGPT 3.5. Em relação ao item (i) essa possibilidade existe, uma vez que não houve nenhuma análise de conjunto de dados em inglês nesta pesquisa. A cerca do item (ii), a atualização dos parâmetros do LLM possibilita a inutilização da ferramenta, em razão da alteração dos padrões de identificação operados pelos detectores. Em suma, independente das suposições, os resultados asseguram que tanto o Writer AI quanto o GPT Zero se mostraram completamente ineficazes ao propósito avaliado.

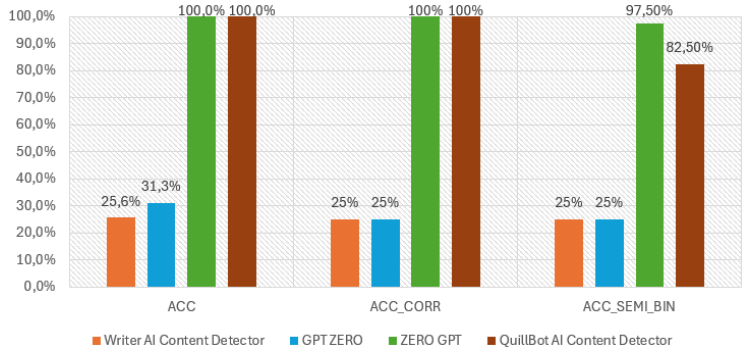

Figura 2. Gráfico da acurácia.

A acurácia do Zero GPT e Quillbot AI calculado com as equações 2 e 3, ilustrado no gráfico da figura 2, fornece valores tendenciosos apontando que ambas classificaram corretamente 100% dos dados. A acurácia semi-binária (equação 4) corrige o desempenho

das ferramentas, Zero GPT obteve 97,50% e Quillbot AI alcançou 82,50% da métrica, com uma diferença de 15%, ocasionada pelas classificações parciais do Quillbot AI. Em geral, as duas aplicações apresentaram boa habilidade em identificar e categorizar corretamente os artigos, seja gerado por IA ou escrito por humano.

4. Considerações Finais

O presente artigo teve por objetivo a avaliação e comparação do desempenho de quatro ferramentas para detecção de artigos gerados pela inteligência artificial do ChatGPT 3.5. O experimento foi realizado a partir da criação de um novo banco de dados formado por 40 manuscritos, sendo 30 elaborados por IA e 10 escritos por pessoas reais. Este conjunto de artigos serviram como cenário de teste para análise do desempenho de cada uma das aplicações de detecção de IA. Os resultados obtidos mostram a eficiência de dois sistemas, Zero GPT e Quillbot AI, com 97,50% e 82,50% de acurácia, respectivamente; enquanto Writer AI e GPT Zero não conseguiram categorizar corretamente nenhum texto gerado por máquina.

Embora o estudo tenha identificado algumas medidas efetivas para solução do problema relatado, os resultados evidenciam a dificuldade de algoritmos em distinguir e detectar corretamente os textos gerados artificialmente. Isto evidencia a necessidade e importância de investir em pesquisas para aprimorar o processo de detecção de IA, a fim de garantir a integridade e a autenticidade dos conteúdos na produção científica. Por fim, se faz preciso a promoção de debates sobre o uso de inteligência artificial de forma ética na pesquisa científica.

Em trabalhos futuros, esta pesquisa tem como pretensão entender os experimentos em novas ferramentas e aumentar a quantidade de dados avaliados, a fim de obter uma análise mais ampla e assertiva dos classificadores de IA.

Agradecimentos

O presente trabalho foi implementado de através do apoio da Pró-Reitoria de Pesquisa, Pós-graduação e Inovação (PRPPI) do Instituto Federal de Alagoas (IFAL), por meio do Programa Institucional de Bolsa de Iniciação Científica (PIBIC) diante do edital de n° 16/2023/PRPPI/IFAL. Projeto sob código “PVE820-2023”, denominado “Avaliação De Ferramentas Para Detecção De Artigos Gerados Por Inteligência Artificial”.

Referências

Generating sentiment-preserving Fake Online Reviews Using Neural Language Models and Their Human- and Machine-based Detection. Adelani, David Ifeoluwa, et al. 2020. 2020. Advanced Information Networking and Applications.

A Benchmark Corpus for the Detection of Automatically Generated Text in Academic Publications. Vijini, Liyanage, et al. 2022. 2022. Proceedings of the Thirteenth Language Resources and Evaluation Conference.

Cabanac, Guillaume e Labbé, Cyrill. 2021. Prevalence of nonsensical algorithmically generated papersin the scientific literature. JASIST. Dezembro de 2021, pp. 1461-1476.

Chaka, Chaka e al, et. 2023. Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. Journal of Applied Learning & Teaching. 2023.

Deepfake Text Detection: Limitations and Opportunities. Jiameng, Pu, Zain, Sarwar e Sifat, Muhammad Abdullah. 2023. San Francisco : s.n., 2023. IEEE Symposium on Security and Privacy .

Dehouche, N. 2021. Plagiarism in the Age of Massive Generative Pre-Trained Transformers (Gpt-3). Ethics in Science and Environmental Politics. 2021, pp. 17–23.

Elkhatat, A.M., Elsaid, K e Almeer, S. 2023. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. International Journal for Educational Integrity. 2023.

FERREIRA, Rafael Clementino Veríssimo, GARCIA, Gustavo Henrique Maia e BRASIL, Deilton Ribeiro. 2023. O surgimento do Chat GPT e a insegurança sobre o futuro dos trabalhos acadêmicos. Cadernos de Dereito Actual. 2023, pp. 130-143.

Gil, Antonio Carlos. 2017. Como Elaborar Projetos de Pesquisa. 6. São Paulo : Atlas, 2017.

GPTZero. 2023. GPTZero: The Global Standard for AI Detection, Humans Deserve the Truth. [Online] GPTZero, 2023. [Citado em: 06 de 30 de 2023.] https://gptzero.me/.

IGA: An Intent-Guided Authoring Assistant. Sun, Simeng, Zhao, Wenlong e Manjunatha, Varun. 2021. 2021. Empirical Methods in Natural Language Processing.

Kıyak, Yavuz Selim. 2023. A ChatGPT Prompt for Writing Case-Based Multiple-Choice Questions. Revista Española de Educación Médica. 23 de 10 de 2023, pp. 98-103.

Language Models are Few-Shot Learners. BROWN, Tom e al, et. 2020. Vancouver : s.n., 2020. Advances in neural information processing systems. pp. 1877-1901.

QuillBot. 2017. QuillBot AI Detector. QuillBot. [Online] 14.957.4, Learneo, 2017. [Citado em: 12 de 12 de 2023.] https://quillbot.com/ai-content-detector.

Stokel-Walker, C. AI bot ChatGPT writes smart essays-should academics worry? Nature.

Weber Wulff, Debora e al, et. 2023. Testing of detection tools for AI generated text. International Journal for Educational Integrity. 2023.

White, Jules e al., et. 2023. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. Arxiv. 21 de 2 de 2023.

WRITER. 2024. AI Content Detector. Writer. [Online] Writer, 2024. [Citado em: 04 de 03 de 2024.] https://writer.com/ai-content-detector/.

ZeroGPT. 2023. AI Detector - Trusted AI Checker for ChatGPT, GPT4 & Bard. ZeroGPT. [Online] ZeroGPT, 01 de 01 de 2023. [Citado em: 12 de 12 de 2023.] https://www.zerogpt.com/.