UNIVERSIDADE FEDERAL DO PAMPA
UNIVERSIDADE ABERTA DO BRASIL
ESPECIALIZAÇÃO EM MÍDIA E EDUCAÇÃO

NATHÁLIA PINHEIRO MARTINS

A (IN)CONFIABILIDADE DE DETECTORES DE TEXTO DE INTELIGÊNCIA ARTIFICIAL: O CASO DO ZEROGPT

São Borja
2025

NATHÁLIA PINHEIRO MARTINS

A (IN)CONFIABILIDADE DE DETECTORES DE TEXTO DE INTELIGÊNCIA ARTIFICIAL: O CASO DO ZEROGPT

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Mídia e Educação da Universidade Federal do Pampa/ Universidade Aberta do Brasil/, como requisito parcial para a obtenção do Título de Especialista em Mídia e Educação.

Orientador: Prof. Dr. Érico Marcelo Hoff do Amaral

Coorientador: Prof. Dr. Cristiano Galafassi

São Borja
2025

Ficha catalográfica elaborada automaticamente com os dados fornecidos
pelo(a) autor(a) através do Módulo de Biblioteca do
Sistema GURI (Gestão Unificada de Recursos Institucionais) .

M386Martins,NatháliaPinheiro
A (in)confiabilidade de detectores de texto de inteligênciaartificial:ocasodoZeroGPT/Nathália Pinheiro Martins.
21p.

Trabalho de Conclusão de Curso(Especialização)--UniversidadeFederaldoPampa,ESPECIALIZAÇÃOEMMÍDIA E EDUCAÇÃO, 2025.
"Orientação:ÉricoMarceloHoffdoAmaral".

1.InteligênciaArtificialGenerativa.2.ZeroGPT.
3.Falsospositivos.I.Título.

NATHÁLIA PINHEIRO MARTINS

A (IN)CONFIABILIDADE DE DETECTORES DE TEXTO DE INTELIGÊNCIA ARTIFICIAL: O CASO DO ZEROGPT

Trabalho de Conclusão de Curso apresentado ao Curso de Pós-Graduação Lato Sensu – Especialização em Mídia e Educação da Universidade Aberta do Brasil/Universidade Federal do Pampa, como requisito parcial para obtenção do Título de Especialista em Mídia e Educação.

Trabalho de Conclusão de Curso defendido e aprovado em: 19 de novembro de 2025.

Banca examinadora:

gov.br
Documento assinado digitalmente
ERICO MARCELO HOFF DO AMARAL
Data: 10/12/2025 08:43:45-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Érico Marcelo Hoff do Amaral
Orientador
Unipampa

gov.br
Documento assinado digitalmente
CRISTIANO GALAFASSI
Data: 09/12/2025 19:09:01-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Cristiano Galafassi
Coorientador
Unipampa

Documento assinado digitalmente
gov.br
ANDERSON LUIS JESKE BIHAIN
Data: 10/12/2025 11:28:47-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Anderson Luis Jeske Bihain
Unipampa

Documento assinado digitalmente
gov.br
FABIANE FLORES PENTEADO GALAFASSI
Data: 09/12/2025 20:19:20-0300
Verifique em https://validar.iti.gov.br

Prof.ª Dr.ª Fabiane Flores Penteado Galafassi
Unipampa

A (IN)CONFIABILIDADE DE DETECTORES DE TEXTO DE INTELIGÊNCIA ARTIFICIAL: O CASO DO ZEROGPT

Nathália Pinheiro Martins[1]

Resumo

Este artigo investiga a (in)confiabilidade do ZeroGPT, detector de textos baseados em inteligência artificial (IA), ao avaliar produções humanas em língua portuguesa. A pesquisa combina uma Revisão Sistemática da Literatura (RSL) com um experimento empírico, no qual o sistema foi aplicado ao segundo capítulo do livro Aula de Português, de Irandé Antunes, anterior à popularização da IA generativa. Os resultados revelam instabilidade nos critérios de classificação e altos índices de falsos positivos — entre 26% e 40% —, evidenciando fragilidades técnicas e éticas. A análise confirma problemas como a opacidade algorítmica, o viés linguístico e a falta de rastreabilidade. Essas limitações comprometem a validade científica e o uso pedagógico do detector, pois textos humanos podem ser rotulados como artificiais. Defende-se que a verificação de autoria priorize princípios éticos baseados no diálogo e na transparência, evitando a dependência de sistemas automatizados que carecem de explicabilidade e podem gerar avaliações injustas no contexto educacional.

Palavras-chave: inteligência artificial generativa; ZeroGPT; falsos positivos.

Abstract

This article investigates the (un)reliability of ZeroGPT, an artificial intelligence (AI)–based text detector, in evaluating human-written works in Portuguese. The study combines a Systematic Literature Review (SLR) with an empirical experiment in which the system was applied to the second chapter of Aula de Português by Irandé Antunes, written before the popularization of generative AI. The results reveal instability in classification criteria and high rates of false positives — between 26% and 40% —, highlighting technical and ethical weaknesses. The analysis confirms issues such as algorithmic opacity, linguistic bias, and lack of traceability. These limitations undermine the scientific validity and pedagogical use of the detector, as human texts may be labeled as artificial. It is argued that authorship verification should

[1] Licenciada em Letras – Português pela Universidade Federal do Pampa, Especialista em Linguística Aplicada e Ensino de Línguas pela Universidade Federal de Mato Grosso do Sul. E-mail: nathaliamartins.aluno@unipampa.edu.br

prioritize ethical principles grounded in dialogue and transparency, avoiding reliance on automated systems that lack explainability and may lead to unfair evaluations in educational contexts.

Keywords: generative artificial intelligence; ZeroGPT; false positives.

1 INTRODUÇÃO

O avanço das tecnologias de inteligência artificial (IA) tem transformado a forma como textos são produzidos, revisados e avaliados. Em contextos educacionais, essas ferramentas passaram a atuar como assistentes de escrita, apoiando estudantes na organização de ideias, na reformulação de enunciados e na ampliação da clareza textual. Quando utilizadas de modo orientado e ético, podem favorecer o aprendizado da escrita acadêmica e atuar como instrumentos de apoio pedagógico. Contudo, seu uso crescente tem provocado tensões quanto à autenticidade e à originalidade dos textos produzidos (Santaella; Kaufman, 2024).

Entre as vertentes da IA, a inteligência artificial generativa (IAGen) destaca-se por empregar modelos de linguagem capazes de gerar textos, imagens e sons com alto grau de coerência e fluidez. Esses modelos, treinados em grandes bases de dados e técnicas de processamento de linguagem natural (PLN), aprendem a reconhecer e reproduzir padrões sintáticos e semânticos, o que lhes permite simular a escrita humana (Russell; Norvig, 2021; Goodfellow; Bengio; Courville, 2016). Ferramentas como ChatGPT e Gemini tornaram-se exemplos emblemáticos dessa nova etapa, rapidamente incorporada às rotinas de estudo, ensino e produção acadêmica (He; Cao; Tan, 2025).

Esse cenário inaugura um novo paradigma: a convivência entre autoria humana e produção automatizada. O mesmo potencial que torna a IAGen uma aliada na escrita também suscita dilemas éticos e avaliativos, especialmente quando professores e instituições passam a recorrer a detectores de IA para distinguir textos humanos de textos artificiais. Tais detectores, como o ZeroGPT, baseiam-se em cálculos estatísticos e probabilísticos, analisando previsibilidade lexical, repetição e coesão para estimar a origem de um texto. Embora se apresentem como soluções de verificação de autoria, há pouca evidência sobre a eficiência e estabilidade desses sistemas, especialmente em textos não produzidos em inglês (Liang *et al*., 2023; Cooperman; Brandão, 2024).

Diante disso, surge o seguinte problema de pesquisa: o detector ZeroGPT é confiável para identificar corretamente textos de autoria humana em língua portuguesa?

Para investigar essa questão, este artigo analisa o desempenho do ZeroGPT na classificação de trechos do livro Aula de Português (Antunes, 2003), anterior à popularização da IA generativa. O objetivo é verificar se a ferramenta reconhece adequadamente a autoria humana ou apresenta classificações equivocadas. A escolha do tema se justifica pela relevância de compreender o impacto pedagógico e ético da adoção de detectores automatizados em contextos educacionais, uma vez que resultados imprecisos podem gerar julgamentos injustos, insegurança e desconfiança nos processos de ensino e aprendizagem.

O presente artigo está estruturado em seis seções: a metodologia, com os procedimentos da pesquisa; a terceira expõe o referencial teórico, dividido em subseções sobre IA, IA generativa, ética e funcionamento dos detectores; a quarta descreve o desenvolvimento da aplicação prática; a quinta discute os resultados à luz da literatura; e a sexta reúne as considerações finais do estudo.

2 METODOLOGIA

Neste trabalho, realizou-se uma Revisão Sistemática da Literatura (RSL), com o objetivo de identificar, selecionar, avaliar e sintetizar estudos prévios sobre o tema (Cavalcante; Oliveira, 2020), que subsidiaram a elaboração do referencial teórico e, posteriormente, uma pesquisa experimental (Gil, 2017). Para apresentar a metodologia adotada, utilizou-se a representação baseada em Método, Procedimentos, Objetivos e Técnicas (MPOT), organizado conforme o Quadro 1:

Quadro 1 – Etapas da pesquisa, de acordo com o modelo MPOT

Método (M)	**Procedimento (P)**
Abordagem qualitativa, descritiva e exploratória, fundamentada em RSL e orientada por um referencial teórico que contempla IA, IAGen, ética e o detector ZeroGPT.	O procedimento foi organizado em seis etapas: definição do escopo, escolha dos descritores, busca em bases, triagem, extração dos dados e síntese conforme o modelo PRISMA.
Objetivo (O)	**Técnica (T)**
Mapear evidências de desempenho de detectores de IA para subsidiar o referencial teórico e orientar um experimento aplicando excertos de um livro no ZeroGPT.	Buscas sistemáticas e análise qualitativa dos estudos selecionados, com identificação de categorias temáticas sobre desempenho e limitações dos detectores de IA.

Fonte: Elaborado pela autora (2025).

A integração entre a RSL e o experimento com o ZeroGPT possibilitou delinear um panorama sobre o desempenho e as limitações do detector, articulando evidências teóricas e empíricas para subsidiar a análise da ocorrência de falsos positivos em textos humanos.

3 REFERENCIAL TEÓRICO

Esta seção apresenta a fundamentação teórica da pesquisa, estruturada em cinco partes: a primeira descreve a RSL e o estado da arte; a segunda aborda os conceitos de IA; a terceira discute a IAGen; a quarta trata dos desafios e implicações éticas da nova primavera da IA; e a quinta analisa o funcionamento, limitações e vieses linguísticos do ZeroGPT.

3.1 Revisão Sistemática da Literatura e Estado da Arte

A RSL foi conduzida para reunir estudos recentes sobre IAGen, uso ético e detectores de IA, com ênfase no desempenho e nas limitações do ZeroGPT. O processo envolveu busca automatizada e leitura analítica dos estudos obtidos, resultando na seleção de produções relevantes publicadas entre 2015 e 2025. A pergunta que orientou a pesquisa foi: “O detector ZeroGPT é confiável para identificar corretamente textos de autoria humana em língua portuguesa?”.

A busca foi realizada com apoio da ferramenta Undermind, que acessa bases de dados de acesso aberto e indexadores como ArXiv, IEEE Xplore, SpringerLink, Elsevier, Scopus, Web of Science, SciELO, Google Scholar e Open Information Science, retornando artigos revisados por pares. Foram utilizados descritores em português e inglês, combinados por operadores booleanos: “Inteligência Artificial” OR “Artificial Intelligence”; “IA Generativa” OR “Generative AI”; “ZeroGPT” OR “AI detector”; “falsos positivos” OR “false positives”; “Ética” OR “Ethics”.

Quanto aos critérios de inclusão e exclusão (Cavalcante; Oliveira, 2020), foram incluídos estudos em português ou inglês, publicados no período delimitado, que abordassem os temas solicitados. Excluíram-se trabalhos sem acesso aberto, duplicados, sem metodologia identificável e que tratassem de temáticas fora do escopo da pesquisa.

Conforme proposto por Cavalcante e Oliveira (2020), as etapas do processo metodológico foram: (1) definição da pergunta de pesquisa; (2) seleção dos descritores; (3) busca automatizada na ferramenta Undermind; (4) triagem por título, resumo e número de citações; (5) leitura dos estudos elegíveis; e (6) extração e síntese dos dados relevantes.

A busca identificou 112 estudos, dos quais 67 atenderam aos critérios de inclusão e foram selecionados para análise inicial. Os textos foram agrupados nas seguintes categorias temáticas: (1) Conceitos de IA e/ou IAGen; (2) Detectores de IA e ZeroGPT; (3) Uso ético de IA. Após a leitura de seus resumos, fez-se uma nova triagem, baseada na quantidade de citações e *match score*, e foram selecionados 8 artigos, apresentados no quadro 2:

Quadro 2 – Resultado da RSL

Título, autor e ano	**Síntese**
How the machine 'thinks': Understanding opacity in machine learning algorithms, de J. Burrell (2016)	Aponta três formas de opacidade (sigilo, barreira técnica e complexidade) que tornam os algoritmos "caixas-pretas" e limitam a transparência dos resultados.
AI tools vs AI text: detecting AI-generated writing in foot and ankle surgery, de S. R. Cooperman, R. A. Brandão (2016)	Analisa seis detectores de IA, verificando acurácia média de 63%, com 83% de falsos positivos no ZeroGPT e queda de detecção após reescrita com ChatGPT.
Generative Adversarial Networks, de Goodfellow *et al*. (2020).	Descreve as GANs como modelos em competição entre gerador e discriminador, voltados à produção de dados realistas, ressaltando avanços e desafios de estabilidade.
Generative Artificial Intelligence: a Historical Perspective, de R. He, J. Cao e T. Tan (2025).	Revisa a evolução da IAGen em quatro fases: sistemas baseados em regras, algoritmos modelados, metodologias profundas e modelos fundacionais, destacando avanços, desafios e perspectivas futuras.
The global landscape of AI ethics guidelines, de A. Jobin, M. Ienca, E. Vayena (2019)	Analisa 84 diretrizes globais de ética em IA e identifica convergência em cinco princípios: transparência, justiça, não maleficência, responsabilidade e privacidade, destacando divergências na interpretação e aplicação desses valores entre setores e países.
GPT detectors are biased against non-native English writers, de W. Liang *et al.* (2023)	Avalia sete detectores de IA e constata viés sistemático contra autores não nativos, com 61% de falsos positivos em textos do TOEFL. Demonstra que simples ajustes de vocabulário reduzem o erro, expondo limitações éticas e técnicas dos detectores baseados em perplexidade

Detecting generative artificial intelligence in scientific articles: Evasion techniques and implications for scientific integrity, de G. Odri e D. J. Y. Yoon (2023).	Avalia 11 detectores aplicados a textos do ChatGPT-4, mostrando que pequenas alterações, como remoção de vírgulas, parafraseamento ou troca de letras latinas por cirílicas, tornam os textos indetectáveis. O ZeroGPT apresentou falsos positivos em textos humanos, evidenciando baixa confiabilidade e vulnerabilidade a manipulações
A Inteligência artificial generativa como quarta ferida narcísica do humano, de L. Santaella e D. Kaufman (2024).	Analisa a ascensão da IAGen e propõe que seu advento representa a "quarta ferida narcísica" da humanidade, ao desafiar a exclusividade humana na linguagem e na criação simbólica. Explora bases técnicas, implicações culturais e éticas desse impacto.

Fonte: Elaborado pela autora (2025).

A partir dos estudos selecionados, foi possível delinear um panorama teórico que orienta a análise desenvolvida nesta pesquisa. A seção a seguir apresenta os fundamentos teóricos construídos com base nessas evidências, articulando conceitos que contextualizam os aspectos técnicos que sustentam o experimento realizado.

3.2 Inteligência Artificial: conceitos básicos

A IA constitui um campo interdisciplinar dedicado ao desenvolvimento de sistemas computacionais capazes de realizar tarefas que, se executadas por humanos, demandariam inteligência, como aprender, raciocinar, compreender linguagem e resolver problemas. Trata-se do estudo de agentes racionais, isto é, entidades capazes de perceber o ambiente, processar informações e agir de modo a maximizar suas chances de alcançar metas definidas (Russell; Norvig, 2021). Essa definição destaca a inteligência como um processo de tomada de decisão orientado por dados e voltado à adaptação ao contexto.

Historicamente, o desenvolvimento da IA articulou-se em quatro abordagens principais: sistemas que pensam como humanos, que agem como humanos, que pensam racionalmente e que agem racionalmente. A última abordagem, predominante nas pesquisas atuais, baseia-se em princípios de lógica, probabilidade e aprendizado, distanciando-se da mera imitação do comportamento humano (Russell; Norvig, 2021).

O funcionamento da IA combina representação de conhecimento, raciocínio automatizado e aprendizado de máquina, ramo que desenvolve algoritmos capazes de generalizar a partir de exemplos e identificar regularidades em grandes volumes de dados. O avanço dessa área levou ao surgimento do aprendizado profundo (*deep learning*), baseado em redes neurais artificiais com múltiplas camadas hierárquicas, nas quais cada nível abstrai características mais complexas dos dados (Goodfellow; Bengio; Courville, 2016). Essas redes, inspiradas no cérebro humano, ajustam automaticamente seus parâmetros à medida que são expostas a exemplos, permitindo reconhecer padrões e adaptar-se a novos contextos. Entre as arquiteturas mais conhecidas estão as redes convolucionais (CNNs), eficazes no reconhecimento de imagens, e as recorrentes (RNNs), aplicadas ao processamento de sequências temporais, como fala e texto.

Esses avanços transformaram a forma como a IA processa linguagem, imagem e som, abrindo caminho para sistemas capazes não apenas de identificar padrões, mas de gerar novos conteúdos, marcando a transição entre a IA analítica e a IA criativa — base da IAGen.

3.3 A inteligência artificial generativa e os modelos de base

A IAGen representa uma nova etapa no desenvolvimento da IA, caracterizada pela capacidade de criar conteúdos originais — textos, imagens, sons, vídeos e códigos — a partir de dados e padrões previamente aprendidos. Diferentemente da IA preditiva, voltada à identificação de padrões e à realização de previsões, a IAGen simula a produção de linguagem e de expressão simbólica. De acordo com Santaella e Kaufman (2024), essa autonomia criativa provoca um deslocamento cultural e epistemológico, pois desafia a exclusividade humana sobre a linguagem, a criação e o pensamento.

O desenvolvimento das IAGens foi possibilitado pelo avanço dos modelos de linguagem (*language models*), sistemas treinados com grandes volumes de texto para prever a sequência mais provável de palavras ou fragmentos (*tokens*) em um enunciado. A geração textual ocorre de forma probabilística: o modelo analisa o contexto e seleciona, a cada etapa, o token mais coerente para continuar a frase, equilibrando previsibilidade e criatividade linguística. Parâmetros como *temperature* e *burstiness* regulam o grau de variação lexical e sintática, influenciando o estilo e a fluidez do texto produzido.

Além dos modelos de linguagem, outra arquitetura significativa para a evolução da IAGen é a das redes adversariais generativas (*Generative Adversarial Networks* – GANs), propostas por Goodfellow *et al*. (2020). Nelas, dois sistemas neurais — o gerador e o

discriminador — competem entre si: o primeiro cria dados sintéticos, enquanto o segundo tenta distinguir esses dados dos reais. À medida que ambos se aperfeiçoam, o gerador passa a produzir resultados cada vez mais verossímeis, o que torna as GANs amplamente utilizadas na criação de textos, imagens, vídeos e vozes sintéticas.

He, Cao e Tan (2025) descrevem quatro fases da evolução da IA generativa: (i) sistemas baseados em regras; (ii) modelos estatísticos, que produziam dados por probabilidades; (iii) modelos generativos profundos, sustentados por redes neurais; e (iv) modelos de base (foundation models), como ChatGPT e Gemini, capazes de operar de modo multimodal, integrando texto, imagem, som e vídeo. Treinados com bilhões de parâmetros, esses modelos representam um salto em escalabilidade, flexibilidade e adaptação a diversos domínios. Vicari *et al.* (2025) chamam esse período de "nova primavera da IA", marcada pela popularização dos modelos generativos e sua incorporação às práticas acadêmicas, científicas e educacionais.

3.4 Desafios e implicações éticas da nova primavera da IA

Apesar do potencial inovador das IAGens, seu funcionamento ainda é marcado pela opacidade algorítmica e por fenômenos como as alucinações de informação — situações em que o sistema gera respostas aparentemente coerentes, mas incorretas. He, Cao e Tan (2025) observam que a dificuldade de rastrear as fontes dos dados de treinamento e a ausência de mecanismos robustos de verificação comprometem a confiabilidade dos resultados, levantando questões éticas sobre autoria, autenticidade e responsabilidade algorítmica.

Burrell (2016) identifica três dimensões da opacidade em sistemas de aprendizado de máquina: o sigilo corporativo sobre o funcionamento interno dos modelos; a barreira técnica, que impede a compreensão por parte dos usuários; e a opacidade inerente à complexidade estatística dos próprios algoritmos. Esses fatores tornam difícil compreender como e por que um modelo chega a determinado resultado, o que se reflete diretamente na avaliação de autoria e na confiabilidade das respostas produzidas por IA generativa.

Do ponto de vista cultural, Santaella e Kaufman (2024) argumentam que o avanço da IAGen representa uma "quarta ferida narcísica" para a humanidade, pois desloca o humano de seu lugar central como único produtor de linguagem e de sentido. Essa transformação não se limita ao campo técnico: ela abala a noção de subjetividade e reconfigura o papel do humano frente à criação simbólica. As IAGens, ao imitarem processos expressivos sem compreendê-los, desafiam a noção de autoria e ampliam os riscos de desinformação e reprodução de vieses culturais.

De acordo com Jobin, Ienca e Vayena (2019), as diretrizes internacionais de ética em IA convergem em torno de princípios como transparência, rastreabilidade, justiça e responsabilidade. Esses princípios são importantes para orientar o desenvolvimento e o uso responsável das tecnologias generativas, de modo a mitigar riscos associados à manipulação de dados e à produção automatizada de conteúdo. Nesse contexto, emerge uma nova frente de preocupação acadêmica: se as IAGens ampliam as possibilidades de criação, também tornam mais complexa a tarefa de verificar a autoria e a origem dos textos produzidos.

Assim, ganha destaque o papel de ferramentas desenvolvidas para detectar a presença de IA em produções humanas, especialmente em ambientes educacionais e científicos. É nesse cenário que se insere o ZeroGPT, uma das plataformas mais conhecidas voltadas à identificação de textos gerados por IA.

3.5 ZeroGPT: funcionamento, limitações e vieses linguísticos

O ZeroGPT é uma plataforma digital que reúne diferentes recursos voltados à análise e à produção textual mediada por IA. De acordo com informações disponíveis em seu site oficial, o serviço inclui um detector de IA — destinado a estimar a probabilidade de um texto ter sido gerado por modelos como ChatGPT, GPT-4, GPT-5, Gemini, Grok, Claude, DeepSeek e LLaMA —, além de outras funcionalidades, como verificador de plágio, parafraseador, resumidor, corretor gramatical, gerador de citações e tradutor. A ferramenta é apresentada como uma solução integrada para uso em ambientes educacionais, empresariais e editoriais, com foco na verificação de autoria e no suporte a práticas de integridade acadêmica (ZeroGPT, 2025).

Embora anuncie índices de acurácia superiores a 98%, estudos empíricos demonstram que o desempenho do ZeroGPT é muito inferior. Cooperman e Brandão (2024), ao avaliarem seis detectores de IA, constataram uma acurácia média de apenas 63% e identificaram o ZeroGPT como o sistema com o maior número de falsos positivos (83%), classificando textos humanos como gerados por IA. Os autores também observaram que simples reescritas feitas por modelos generativos reduziram em mais de 50% a taxa de detecção, revelando a vulnerabilidade e a falta de confiabilidade de seus resultados.

O funcionamento do ZeroGPT é descrito de forma genérica: o site menciona o uso de "sinais linguísticos e estatísticos" e de um conjunto de classificadores que analisam padrões de *tokens*, *burstiness* (variabilidade) e entropia textual (ZeroGPT, 2025). No entanto, não há informações públicas sobre os pesos atribuídos a cada fator nem sobre os parâmetros que definem o limiar entre um texto "humano" e um texto "gerado por IA". Essa ausência de

transparência aproxima o detector de um sistema do tipo caixa preta, conceito discutido por Burrell (2016) ao se referir a modelos computacionais cuja lógica interna é inacessível ao escrutínio público ou científico. Assim, embora os detectores apresentem resultados probabilísticos, o usuário não dispõe de critérios verificáveis para compreender como a decisão é produzida, o que gera insegurança e dificulta qualquer avaliação de confiabilidade.

Essas fragilidades também são evidenciadas por Odri e Yoon (2023), que testaram o desempenho do ZeroGPT e de outros dez detectores de IA em textos produzidos pelo ChatGPT-4. No experimento, os autores aplicaram modificações graduais ao texto original — como remoção de vírgulas, introdução de pequenos erros gramaticais, parafraseamento automático, ajuste de parâmetros de geração (perplexidade e variabilidade) e até a substituição de letras latinas por caracteres cirílicos visualmente idênticos — para verificar a capacidade de detecção dos sistemas. Os resultados indicaram que, mesmo sem alterações, a maioria dos detectores, incluindo o ZeroGPT, classificou o texto como humano, e que pequenas intervenções tornaram o conteúdo completamente indetectável. Além disso, o sistema apresentou alta taxa de falsos positivos, confirmando sua baixa precisão e vulnerabilidade a manipulações simples.

Outro ponto relevante é a contradição entre a alegação de funcionamento universal e o contexto linguístico de aplicação do detector. Embora o ZeroGPT esteja disponível em português europeu, inglês, francês, alemão, espanhol e italiano, o site afirma manter a mesma precisão em todos os idiomas (ZeroGPT, 2025), o que é improvável, considerando as diferenças estruturais e morfossintáticas entre línguas. Essa limitação se relaciona ao viés linguístico de sistemas treinados majoritariamente em inglês, como evidenciaram Liang *et al*. (2023) ao avaliarem sete detectores, entre eles o ZeroGPT. Os autores constataram que textos de falantes não nativos foram classificados como gerados por IA em mais de 60% dos casos, enquanto os de nativos foram identificados como humanos. Isso ocorre porque métricas como a perplexidade, que medem a previsibilidade linguística, tendem a penalizar vocabulário mais restrito e estruturas simples, típicas de quem escreve em segunda língua (Liang *et al*., 2023). Como resultado, ferramentas como o ZeroGPT podem reproduzir desigualdades linguísticas e comprometer a equidade em avaliações acadêmicas.

4 DESENVOLVIMENTO

A etapa de desenvolvimento consistiu na aplicação experimental da versão gratuita do detector ZeroGPT ao livro Aula de Português (Antunes, 2003), com o objetivo de verificar sua capacidade de identificar corretamente textos de autoria humana. Essa fase buscou observar se

o sistema apresentaria falsos positivos, isto é, se classificaria como “gerado por IA” um texto comprovadamente humano e anterior à popularização das tecnologias generativas.

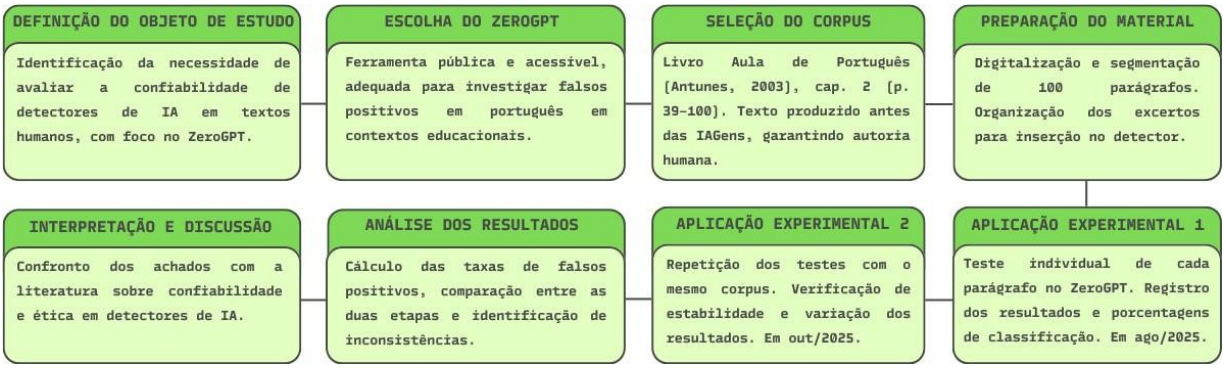
A escolha do ZeroGPT fundamenta-se em sua disponibilidade pública e na relevância prática que a ferramenta adquiriu em discussões sobre integridade acadêmica e verificação de autoria. Por ser um dos detectores mais conhecidos e de uso gratuito, sua análise permite problematizar a confiança depositada em sistemas que carecem de documentação técnica e explicabilidade, especialmente no contexto educacional.

Optou-se por aplicar o experimento a uma obra amplamente reconhecida na formação docente e produzida antes da emergência das IAGens, de modo a garantir a autoria humana e possibilitar a observação de eventuais falsos positivos. O corpus analisado corresponde a 100 parágrafos do segundo capítulo do livro Aula de Português, que abrange as páginas 39 a 100. Essa delimitação foi definida por conter trechos de caráter expositivo e argumentativo, organizados em subtópicos e listas que detalham princípios e implicações pedagógicas sobre escrita, leitura, gramática e oralidade.

A autora recorre frequentemente ao uso de travessões, enumerações e tópicos explicativos — recursos que conferem clareza didática, mas também aumentam a regularidade formal do texto. Essas características, aliadas à previsibilidade lexical e à repetição de estruturas paralelas, compõem um tipo de escrita que pode ser interpretado de forma equivocada por detectores de IA, tornando o material adequado para observar possíveis inconsistências do ZeroGPT.

O experimento foi conduzido com duas aplicações experimentais, realizadas em agosto e outubro de 2025, utilizando a versão pública do ZeroGPT disponível em seu site oficial. Para fins de sistematização, a Figura 1 apresenta o fluxograma das etapas de desenvolvimento da pesquisa, desde a seleção do corpus até a análise dos resultados:

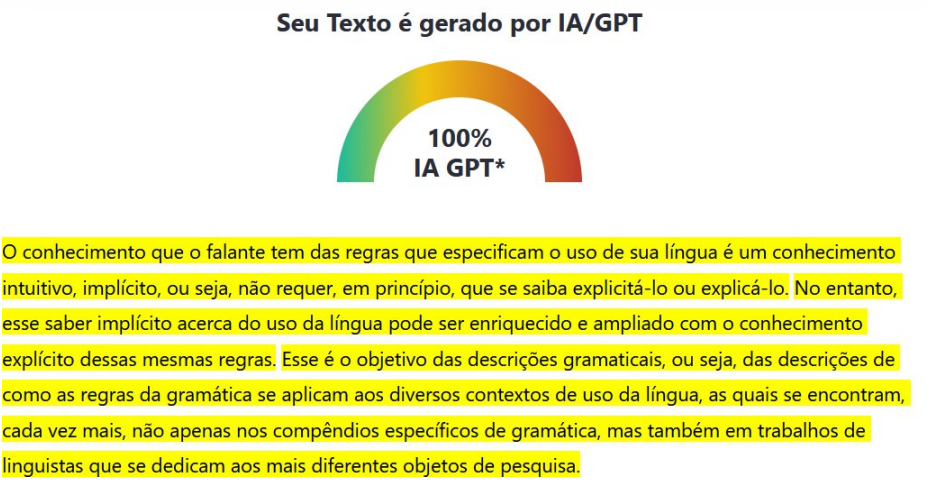
Figura 1 – Etapas de desenvolvimento da pesquisa



Fonte: Da autora (2025).

Na primeira aplicação experimental, cada parágrafo selecionado foi testado individualmente, utilizando a versão do ZeroGPT disponível em agosto de 2025. Para cada inserção, registrou-se o resultado fornecido (texto gerado por IA ou texto escrito por humano), conforme exemplo na Figura 2:

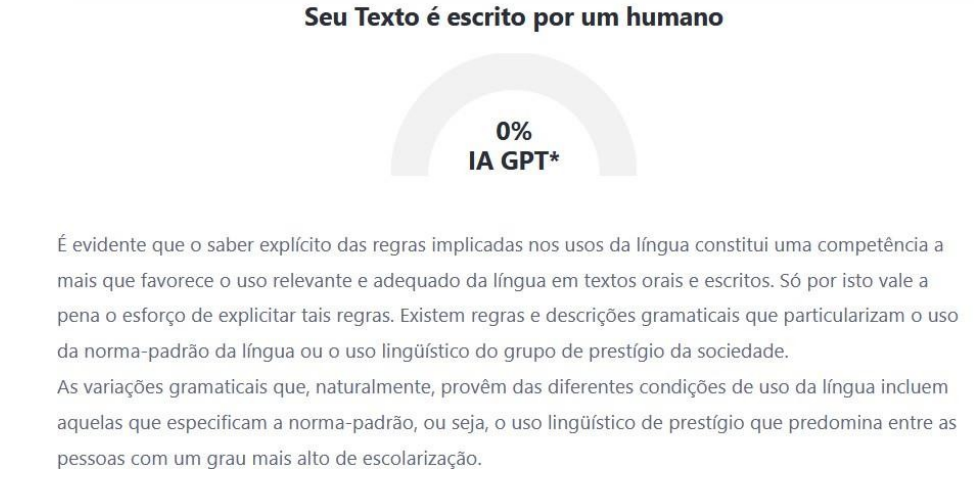
Figura 2 – Excerto da página 94 do livro Aula de Português



Fonte: Da autora (2025), a partir de Aula de Português (Antunes, 2003, p. 94).

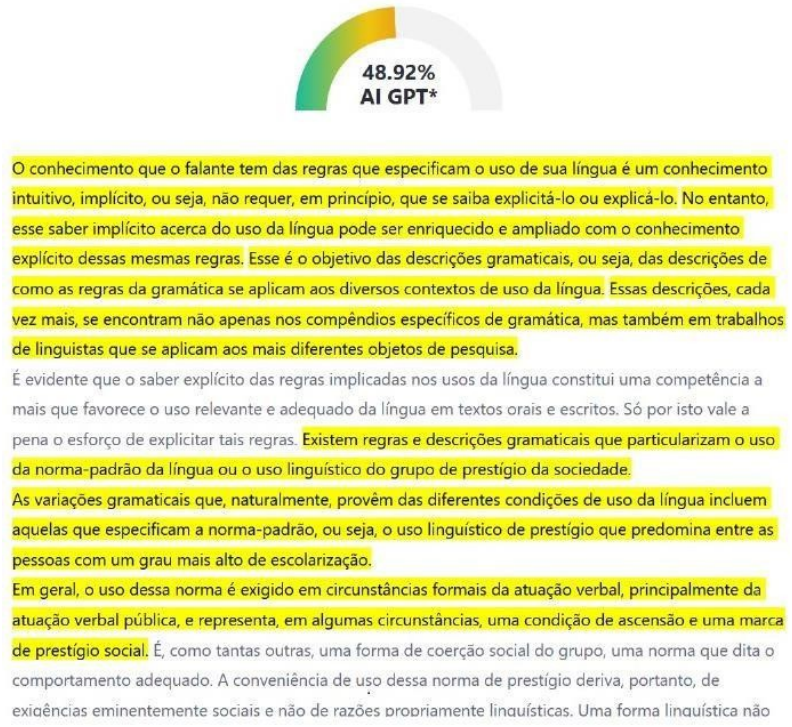
Na segunda etapa, manipulou-se o texto alterando a quantidade de parágrafos testados, conforme figuras 3 e 4, buscando verificar se o aumento do volume textual influenciaria os resultados:

Figura 3 – Excerto 1 da página 95 do livro Aula de Português



Fonte: Da autora (2025), a partir de Aula de Português (Antunes, 2003, p. 95).

Figura 4 – Excertos das páginas 94 e 95 do livro Aula de Português



Fonte: Da autora (2025), a partir de Aula de Português (Antunes, 2003, p. 94-95).

Como é possível observar, parte do mesmo trecho que havia sido classificado como 0% IA na Figura 2 foi identificado como artificial na Figura 3. Além disso, sem o primeiro segmento apresentado na Figura 3, todo o excerto é reconhecido como humano.

Buscando verificar se o resultado seria o mesmo, os testes foram realizados uma segunda vez, dois meses após, em outubro de 2025, apresentando resultados distintos. No exemplo acima, apontou-se apenas o primeiro parágrafo como possivelmente gerado por IA.

Em agosto, dos 100 parágrafos selecionados, 33% apresentaram falsos positivos, enquanto 67% foram considerados humanos. Entretanto, quando testou-se 25 excertos, cada grupo com quatro parágrafos, o resultado variou, elevando o índice de falsos positivos para cerca de 40%. Em outubro, entretanto, 26% dos 100 parágrafos foram considerados artificiais. Com grupos de quatro parágrafos, o valor mudou para 32%.

5. DISCUSSÃO DOS RESULTADOS

Os resultados obtidos com a aplicação do ZeroGPT ao texto de Antunes (2003) indicam inconsistências significativas no funcionamento do detector, confirmando as fragilidades detectadas durante a RSL. A variação entre os testes de agosto e outubro de 2025 demonstra que o sistema não apresenta estabilidade nos critérios de classificação, o que compromete sua

confiabilidade como instrumento de verificação de autoria. Mesmo mantendo o mesmo corpus e as mesmas condições de análise, o detector produziu resultados diferentes — um indício de que o algoritmo pode ser sensível a atualizações de parâmetros internos ou a flutuações de conectividade e formatação textual.

A incidência de falsos positivos, que chegou a 33% na primeira rodada e a 26% na segunda, corrobora as constatações de Cooperman e Brandão (2024), que afirmam que o ZeroGPT apresenta desempenho inferior ao anunciado, com alta taxa de classificações equivocadas. Essa instabilidade reforça a ideia de que o sistema opera como uma "caixa-preta", nos termos de Burrell (2016), uma vez que o usuário não dispõe de informações sobre os critérios de decisão nem sobre os pesos atribuídos a cada variável linguística. Tal opacidade dificulta a interpretação dos resultados e impede o controle científico sobre sua validade.

A variação observada conforme o número de parágrafos testados também sugere que o ZeroGPT se apoia em indicadores estatísticos de previsibilidade lexical (perplexidade e burstiness), e não em uma compreensão semântica do texto. Quando o volume textual aumenta, a regularidade sintática e a repetição de estruturas — características comuns à escrita acadêmica e didática de Antunes — parecem ser interpretadas como sinais de artificialidade.

Do ponto de vista teórico, a análise reforça a pertinência de questionar o uso desses detectores em contextos educacionais e avaliativos. Conforme discutido por Jobin, Ienca e Vayena (2019), a ética em IA requer transparência, rastreabilidade e justiça — princípios que não se concretizam quando ferramentas são aplicadas para julgar a autoria de textos humanos sem disponibilizar critérios claros para o resultado. No caso do ZeroGPT, a ausência de documentação técnica e a falta de explicabilidade comprometem a responsabilidade pedagógica e institucional de quem adota seus resultados como evidência de fraude ou uso indevido de IA.

Além disso, o experimento evidencia como o comportamento do detector é afetado por fatores extratextuais, como a segmentação do corpus e a atualização do algoritmo. Esses elementos tornam impossível reproduzir os resultados de maneira controlada, o que inviabiliza o uso científico ou administrativo do ZeroGPT como instrumento confiável de detecção. Em termos práticos, textos humanos podem ser indevidamente marcados como artificiais, gerando constrangimentos e interpretações equivocadas em ambientes acadêmicos.

6. CONCLUSÃO

A pesquisa desenvolveu uma investigação experimental voltada à análise da (in)confiabilidade do detector ZeroGPT na identificação de autoria textual em língua

portuguesa. O estudo combinou uma RSL com um experimento empírico, no qual o sistema foi aplicado a trechos do livro Aula de Português (Antunes, 2003), de autoria humana comprovada e anterior à popularização da IAGen.

Os resultados evidenciaram instabilidade significativa nos critérios de classificação do ZeroGPT e alta incidência de falsos positivos, que variaram entre 26% e 40%, mesmo sob as mesmas condições de análise. Essa oscilação indica que o algoritmo não mantém padrões consistentes de decisão, confirmando as fragilidades apontadas na literatura quanto à opacidade e ao viés linguístico desses sistemas. Considerando o universo de experimentos realizados, conclui-se que o ZeroGPT não apresenta precisão, estabilidade nem transparência suficientes para ser utilizado como instrumento confiável de verificação de autoria, recomendando-se evitar seu uso em ambiente acadêmico.

Os achados reforçam que a previsibilidade lexical e a regularidade sintática, características comuns em textos didáticos e acadêmicos, são interpretadas pelo detector como indícios de geração artificial, o que amplia o risco de julgamentos equivocados. A ausência de documentação técnica e de explicabilidade dos resultados impede a reprodutibilidade científica e o uso pedagógico responsável dessas ferramentas.

Para pesquisas futuras, espera-se ampliar o estudo para outros contextos, sendo uma possibilidade aplicar o mesmo procedimento a resumos e abstracts de artigos científicos produzidos antes da popularização da IAGen, a fim de verificar a ocorrência de falsos positivos. Também se propõe comparar o desempenho do ZeroGPT a outros detectores.

Por fim, cabe suscitar uma reflexão sobre a contradição de recorrer a uma IA para julgar a presença de outra, atribuindo a uma máquina o papel de mediadora da autoria humana. Esse paradoxo evidencia um dilema próprio da era digital: a tentativa de assegurar autenticidade por meio de sistemas que carecem de transparência e discernimento linguístico. Em ambientes educacionais, essa lógica pode enfraquecer a confiança e deslocar o sentido formativo da escrita para uma prática de vigilância. Refletir criticamente sobre o uso desses recursos torna-se, portanto, indispensável para que a IA seja integrada à educação de modo ético, dialógico e comprometido com a autoria.

REFERÊNCIAS

ANTUNES, I. **Aula de português**: encontro e interação. São Paulo: Parábola Editorial, 2003.

BURRELL, J. How the machine ‘thinks’: Understanding opacity in machine learning algorithms. **Big Data & Society**, Los Angeles, v. 3, n. 1, jun. 2016. Disponível em: https://doi.org/10.1177/2053951715622512. Acesso em: 13 out. 2025.

CAVALCANTE, L. T. C.; OLIVEIRA, A. A. S de. Métodos de revisão bibliográfica nos estudos científicos. **Psicologia em Revista**, vol.26 no.1, Belo Horizonte jan./abr. 2020. Disponível em: https://doi.org/10.5752/P.1678-9563.2020v26n1p82-100. Acesso em: 7 ago. 2025.

COOPERMAN, S. R.; BRANDÃO, R. A. AI tools vs AI text: detecting AI- generated writing in foot and ankle surgery. **Foot & Ankle Surgery**: Techniques, Reports & Cases, Amsterdam, v. 4, n. 1, 2024. Disponível em: https://doi.org/10.1016/j.fastrc.2024.100367. Acesso em: 13 out.

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

GOODFELLOW, I; BENGIO, Y; COURVILLE, A. **Deep Learning**. Cambridge: MIT Press, 2016.

GOODFELLOW, I; POUGET-ABADIE, J; MIRZA, M; XU, B; WARDE-FARLEY, D; OZAIR, S; COURVILLE, A; BENGIO, Y. Generative Adversarial Networks. **Communications of the ACM**, New York, v. 63, n. 11, p. 139-144, nov. 2020. Disponível em: https://doi.org/10.1145/3422622. Acesso em: 20 jun. 2025.

HE, R; CAO, J; TAN, T. Generative Artificial Intelligence: a Historical Perspective. **National Science Review**, Oxford, v. 12, n. 5, maio 2025. Disponível em: https://doi.org/10.1093/nsr/nwaf050. Acesso em: 13 out. 2025.

JOBIN, A; IENCA, M; VAYENA, E. The global landscape of AI ethics guidelines. **Nature Machine Intelligence**, London, v. 1, 2019. Disponível em: https://doi.org/10.1038/s42256-019-0088-2. Acesso em: 18 out. 2025.

LIANG, W.; YUKSEKGONUL, M.; MAO, Y.; WU, E.; ZOU, J. GPT detectors are biased against non-native English writers. **Patterns**, New York, v. 4, n. 7, 2023. Disponível em: https://doi.org/10.1016/j.patter.2023.100779. Acesso em: 13 out. 2025.

ODRI, G; YOON, D. J. Y. Detecting generative artificial intelligence in scientific articles: Evasion techniques and implications for scientific integrity. **Orthopaedics & Traumatology**: Surgery & Research, New York, v. 109, 2023. Disponível em: https://doi.org/10.1016/j.otsr.2023.103706. Acesso em: 13 out. 2025.

RUSSELL, S. J.; NORVIG, P. **Artificial Intelligence**: A Modern Approach. 4. ed. Hoboken: Pearson, 2021.

SANTAELLA, L; KAUFMAN, D. A Inteligência artificial generativa como quarta ferida narcísica do humano. **MATRIZes**, São Paulo, v. 18, n. 1, p. 37-53, jan./abr. 2024. Disponível em: https://doi.org/10.11606/issn.1982-8160.v18i1p37-53. Acesso em: 11 jun. 2025.

VICARI, R; BRACKMANN, C; GALAFASSI, C; MIZUSAKI, L. **IA na Educação Básica**: atividades desplugadas para a Educação Básica. Porto Alegre: Fundação Itaú, 2025.

ZEROGPT. **ZeroGPT**: ferramenta de detecção confiável GPT-5, ChatGPT e IA da ZeroGPT. [*S.l.*]: 2025. Disponível em: https://www.zerogpt.com/. Acesso em: 13 out. 2025.