PADRÕES LEXICAIS E DISCURSIVOS EM TEXTOS GERADOS POR INTELIGÊNCIA ARTIFICIAL: AUTORIDADE, VIÉS E AUTORIA EM PERSPECTIVA CRÍTICA

LEXICAL AND DISCURSIVE PATTERNS IN AI-GENERATED TEXTS: AUTHORITY, BIAS, AND AUTHORSHIP FROM A CRITICAL PERSPECTIVE

DOI:10.47677/gluks.v26i02.592
Recebido: 17/02/2026
Aprovado: 03/08/2026

MORAES, Isabella Tavares Sozza[1]
CASINI, Maria Cecilia [2]

RESUMO: Este artigo discute criticamente padrões lexicais e discursivos documentados em textos produzidos ou mediados por grandes modelos de linguagem. A expressão *léxico algorítmico* é empregada, de modo operacional, como sinônimo de padrões lexicais em textos gerados por inteligência artificial, e não como designação de um repertório único, estável ou inerente a todos os sistemas. Trata-se de um estudo teórico-interpretativo, de abordagem qualitativa, baseado em revisão crítica da literatura e análise secundária de pesquisas empíricas com procedimentos explicitados. A comparação organiza-se em três eixos: recorrência e diversidade lexical; construção discursiva de autoridade; reprodução de vieses sociais. Os estudos examinados mostram que escolhas lexicais variam conforme modelo, versão, comando, parâmetros de geração, idioma, gênero textual e contexto de uso. Por isso, palavras ou fórmulas recorrentes constituem tendências situadas, não marcas universais de autoria algorítmica. A análise indica ainda que a fluência pode encobrir incerteza factual e que regularidades distribucionais podem reiterar assimetrias sociais presentes nos dados e nas práticas de avaliação. Por fim, propõe-se compreender a autoria em interações humano-máquina como processo distribuído, sem transferir ao sistema a responsabilidade epistêmica e ética pelo texto publicado.

PALAVRAS-CHAVE: Inteligência artificial, Padrões lexicais, Discurso algorítmico, Autoria, Viés.

---

[1]Doutoranda em Letras (Língua, Literatura e Cultura Italianas) pela Universidade de São Paulo (USP). E-mail: isabellasozza@alumni.usp.br. Orcid: https://orcid.org/0000-0001-8115-0745.

[2] Livre-docente em Letras (Língua, Literatura e Cultura Italianas) pela Universidade de São Paulo (USP) e Profa. Dra. na mesma instituição e área. E-mail: casini@usp.br. Orcid: https://orcid.org/0000-0001-5264-2046.

1 INTRODUÇÃO

A difusão de sistemas capazes de produzir textos fluidos alterou práticas de escrita acadêmica, jornalística, profissional e cotidiana. Essa transformação tornou frequentes expressões como *linguagem da inteligência artificial*, *discurso algorítmico* e *léxico da inteligência artificial*. Embora úteis para nomear um objeto emergente, essas formulações podem sugerir que diferentes sistemas compartilham um repertório homogêneo e estável. A evidência disponível recomenda cautela: a produção textual depende do modelo e de sua versão, mas também do comando fornecido, do idioma, do gênero solicitado, das configurações de amostragem, das instruções de sistema, do histórico da interação e das intervenções humanas realizadas antes da circulação do texto.

Neste artigo, *léxico algorítmico* e *padrões lexicais em textos gerados por inteligência artificial* são empregados como expressões operacionais equivalentes. Ambas designam regularidades observáveis nas escolhas vocabulares e nas fórmulas discursivas de saídas textuais produzidas ou substancialmente mediadas por grandes modelos de linguagem. Não se trata, portanto, do campo semântico da inteligência artificial — formado por itens como *algoritmo*, *chatbot*, *prompt* e *ChatGPT* — nem de um vocabulário exclusivo que permita identificar, isoladamente, a origem de qualquer texto. O objeto são tendências situadas de seleção e combinação linguística, sempre relacionadas às condições de produção.

Essa delimitação modifica o alcance da investigação. Em vez de perguntar como "o léxico da IA" se diferencia, em bloco, da linguagem humana, formula-se a seguinte questão: sob quais condições certos padrões lexicais e discursivos aparecem em textos gerados por modelos de linguagem, e que consequências esses padrões podem ter para a construção de autoridade, a reprodução de vieses e a atribuição de autoria? A pergunta desloca o debate de uma oposição essencialista entre humano e máquina para a análise das mediações técnicas e sociais que participam da produção do texto.

O objetivo geral é interpretar criticamente evidências empíricas já publicadas sobre regularidades lexicais e discursivas em textos gerados por modelos de linguagem. Os objetivos específicos são: (1) sistematizar o que estudos com procedimentos explícitos permitem afirmar sobre recorrência, diversidade e variação lexical; (2) discutir como fluência,

formas de citação e escolhas vocabulares podem produzir efeitos de autoridade e naturalização; e (3) desenvolver a questão da autoria e da criatividade em processos de escrita mediados por inteligência artificial. Esses objetivos são compatíveis com o caráter teórico-analítico do trabalho: não se anuncia um mapeamento exaustivo nem uma comparação experimental própria entre plataformas.

A contribuição proposta é dupla. No plano conceitual, o artigo diferencia regularidade estatística, padrão discursivo e suposta "assinatura" de autoria algorítmica. No plano metodológico, apresenta uma grade mínima de documentação para pesquisas futuras, incluindo identificação do sistema, versão, data, comandos, parâmetros, fontes solicitadas, repetições e preservação das saídas. Tal protocolo responde a uma dificuldade recorrente do campo: resultados obtidos com uma interface e uma versão específicas podem tornar-se irreplicáveis após atualizações não controladas pelo pesquisador.

O texto está organizado em cinco partes. Após esta introdução, a segunda seção delimita os conceitos de forma, significado e agência aplicados aos modelos de linguagem. A terceira explicita o método de revisão crítica e de análise secundária. A quarta sistematiza os resultados dos estudos selecionados em três eixos. A quinta discute autoridade, poder, autoria e criatividade. A sexta propõe requisitos de transparência para estudos linguísticos de textos gerados por inteligência artificial, antes das considerações finais.

2 LINGUAGEM GERADA, REGULARIDADE E SIGNIFICADO

2.1 Forma linguística e condições de produção

Grandes modelos de linguagem estimam distribuições de probabilidade sobre unidades textuais e geram sequências condicionadas pelo contexto disponível. Dizer isso não significa que suas respostas sejam aleatórias ou desprovidas de estrutura. Ao contrário, o desempenho desses sistemas depende da capacidade de representar regularidades complexas em grandes conjuntos de dados. Significa, porém, que a fluência da saída não autoriza, por si só, inferências sobre compreensão, intenção ou compromisso com a verdade.

O argumento clássico de Searle (1980) distingue a manipulação formal de símbolos da compreensão semântica. Embora formulado décadas antes dos modelos atuais e dirigido a

outra etapa do debate sobre inteligência artificial, ele permanece relevante como advertência conceitual: desempenho linguístico convincente não basta para demonstrar intencionalidade. Em termos mais diretamente relacionados ao processamento contemporâneo de linguagem natural, Bender e Koller (2020) sustentam que um sistema treinado apenas sobre formas linguísticas não obtém, desse material isolado, a relação entre expressões e intenções comunicativas situadas. A aproximação aqui realizada é contemporânea; não se atribui a Searle ou a autores anteriores uma análise específica de sistemas que ainda não existiam.

Floridi (2023) descreve os modelos generativos como formas de agência sem inteligência: podem produzir efeitos e realizar tarefas com elevado desempenho, sem que seja necessário imputar-lhes compreensão humana. Essa distinção é produtiva porque evita dois reducionismos. O primeiro consiste em tratar a saída como simples ruído estatístico, ignorando sua eficácia social. O segundo consiste em converter eficácia em consciência, compreensão ou autoria plena. Entre esses extremos, interessa examinar o texto como resultado de uma cadeia sociotécnica que inclui dados, arquitetura, procedimentos de treinamento, alinhamento, interface, comando, usuário e contexto de circulação.

A escala dessa cadeia também impõe um problema documental. Bender *et al.* (2021) observam que corpora muito extensos podem ampliar custos ambientais, dificultar a rastreabilidade dos dados e reproduzir vozes socialmente dominantes. A advertência não autoriza deduzir o conteúdo de uma resposta diretamente do conjunto de treinamento, mas reforça a necessidade de examinar proveniência, seleção e condições de uso, sempre que essas informações estiverem disponíveis.

2.2 “Léxico algorítmico” como hipótese situada

Um padrão lexical é uma regularidade na frequência, na distribuição ou na associação de itens e construções em um conjunto delimitado de textos. Para que essa regularidade seja interpretável, o conjunto precisa ser descrito: qual sistema produziu as respostas, em que data, em qual idioma, diante de quais comandos e sob quais parâmetros? Sem esses elementos, uma lista de expressões supostamente “típicas de IA” mistura observações impressionistas, tendências de gêneros formais e efeitos específicos de determinados modelos.

Os resultados de Reviriego *et al.* (2024) ilustram a necessidade dessa cautela. Ao comparar respostas humanas e respostas do ChatGPT em conjuntos de tarefas equivalentes, os autores encontraram menor número de palavras distintas e menor diversidade lexical para o GPT-3.5, enquanto o GPT-4 apresentou diversidade mais próxima à dos textos humanos. A própria comparação entre versões impede transformar a menor diversidade em propriedade atemporal de "a IA". Martínez *et al.* (2025), por sua vez, mostram que medidas de riqueza lexical variam conforme a versão do ChatGPT, o papel solicitado ao sistema e parâmetros como penalidade de presença. O que aparece como característica lexical do modelo é, em parte, efeito das condições de geração.

Juzek e Ward (2025) oferecem outro exemplo. Os autores identificam 21 palavras sobre-representadas em resumos científicos em inglês após a popularização dos assistentes de escrita, entre elas *delve*, *intricate* e *underscore*. O estudo não conclui que qualquer ocorrência dessas palavras prove autoria algorítmica. Ao contrário, investiga explicações concorrentes e não encontra base suficiente para atribuir a sobre-representação apenas à arquitetura, aos algoritmos ou aos dados de treinamento. Os testes são compatíveis com possível influência do alinhamento por feedback humano, mas os próprios autores ressaltam limites experimentais e falta de transparência dos sistemas.

Desse modo, *léxico algorítmico* deve ser entendido como abreviação analítica, não como essência. Uma palavra frequente pode resultar da preferência de um modelo, de instruções de polidez, do gênero acadêmico, de edição posterior ou da imitação humana de textos já mediados por sistemas. A circulação dessas formas cria ainda um processo de retroalimentação: usuários podem incorporar escolhas que associam à escrita automatizada, e novos textos humanos podem, depois, compor dados de treinamento. A fronteira entre produção humana e algorítmica torna-se histórica e porosa, não um contraste entre dois códigos isolados.

2.3 Discurso, poder e avaliação

A análise lexical não se encerra na contagem de palavras. Bourdieu (1991) mostra que a eficácia de um enunciado depende das condições sociais que autorizam determinada voz e fazem reconhecer sua legitimidade. Aplicado aos textos gerados por modelos de linguagem,

esse princípio direciona a atenção para as instituições e interfaces que conferem credibilidade à saída: a reputação da plataforma, o design da resposta, o tom assertivo, a presença de referências e a expectativa de eficiência do usuário.

Na perspectiva da ação comunicativa, as pretensões de validade de um enunciado podem ser questionadas e justificadas em interação (Habermas, 1984). Um sistema generativo, entretanto, não assume responsabilidade discursiva nem participa do diálogo nas mesmas condições de um sujeito capaz de responder por seus compromissos. Ele pode produzir justificativas adicionais, mas isso não equivale a garantir a origem ou a validade do conteúdo. A distinção é decisiva para compreender por que uma resposta formalmente bem construída pode exercer autoridade mesmo quando suas fontes são inexistentes.

A pesquisa crítica em processamento de linguagem natural acrescenta que "viés" não é uma propriedade abstrata mensurável sem referência a grupos, contextos e danos. Na revisão de 146 trabalhos sobre viés em processamento de linguagem natural, Blodgett *et al.* (2020) encontraram definições frequentemente vagas, motivações pouco articuladas às desigualdades sociais e tendência a tratar a linguagem como fenômeno dissociado das relações de poder. Assim, a análise de padrões algorítmicos precisa indicar não somente que há diferenças estatísticas, mas quem é representado, em qual situação, por meio de quais categorias e com quais efeitos potenciais.

3 PROCEDIMENTOS METODOLÓGICOS

3.1 Natureza e alcance do estudo

Este trabalho é um estudo teórico-interpretativo de abordagem qualitativa. Seu material analítico é formado por pesquisas empíricas publicadas que examinam diversidade lexical, sobre-representação vocabular, associações sociais, toxicidade e confiabilidade de referências em textos ou representações produzidas por modelos de linguagem. Não foi constituído um corpus original de respostas, não foram feitas coletas próprias em ChatGPT, Gemini, Claude ou outros sistemas, e nenhum percentual apresentado nas seções seguintes deriva de experimentação dos autores deste artigo.

Essa definição metodológica é necessária porque uma discussão crítica apoiada em exemplos secundários não equivale a uma análise de corpus própria. O artigo realiza uma revisão crítica e intencional, e não uma revisão sistemática ou meta-análise. A seleção buscou trabalhos complementares, publicados entre 2016 e 2025, que apresentassem identificação do sistema ou da representação examinada, descrição do material, procedimento analítico e resultados passíveis de verificação. Foram excluídas da argumentação quantitativa referências sem localização bibliográfica confirmável e afirmações cujas condições de produção não estavam documentadas.

O marco inicial de 2016 foi mantido para incluir o estudo de Bolukbasi *et al.* (2016), que não analisa saídas de um modelo conversacional, mas demonstra como associações sociais podem ser codificadas em representações distribuídas usadas no processamento de linguagem. Esse trabalho funciona como antecedente técnico e não é apresentado como evidência direta sobre modelos generativos atuais. Os demais estudos tratam de geração textual, assistentes conversacionais ou mudanças lexicais associadas à mediação por grandes modelos.

3.2 Grade de extração e análise

Para cada pesquisa, foram observados: sistema ou representação examinada; versão, quando informada; tipo e dimensão do corpus; tarefa ou forma de comando; idioma e gênero textual; unidade de análise; procedimento de mensuração; exemplos lexicais; resultados principais; e limites reconhecidos. O Quadro 1 sintetiza essas informações. Quando um trabalho não disponibiliza determinado elemento, a lacuna é tratada como limite, não preenchida por inferência.

Estudo	Material e procedimento	Contribuição para esta análise	Limite de generalização
Bolukbasi *et al.* (2016)	Vetores de palavras treinados em textos do Google News; análise geométrica de associações de gênero e ocupação, seguida de proposta de mitigação.	Mostra que associações sociais podem ser preservadas em representações distribucionais.	Não analisa respostas conversacionais nem grandes modelos generativos atuais.

Estudo	Material e procedimento	Contribuição para esta análise	Limite de generalização
Sheng *et al.* (2019)	Gerações condicionadas por comandos com referências a grupos demográficos; anotação humana de sentimento e de "consideração" dirigida ao grupo; treinamento de classificador.	Relaciona escolhas geradas a avaliações distintas de grupos e mostra a insuficiência de medir apenas sentimento.	Modelos e condições anteriores aos assistentes conversacionais atuais; resultados dependem dos comandos e das categorias usadas.
Gehman *et al.* (2020)	*RealToxicityPrompts*: 100 mil comandos naturais derivados de textos da web, pontuados por classificador de toxicidade; comparação de modelos e métodos de controle.	Evidencia que continuações tóxicas podem surgir até de comandos aparentemente inócuos e liga o fenômeno aos dados de treinamento.	A métrica automática de toxicidade também possui vieses; corpus e resultados são predominantemente em inglês.
Abid, Farooqi e Zou (2021)	Testes com GPT-3 por completamento de comandos, analogias e geração de histórias envolvendo identidades religiosas.	Demonstra associação persistente entre referências a muçulmanos e violência em diferentes tarefas de geração.	Examina uma versão específica do GPT-3 e categorias religiosas em inglês; não autoriza extrapolação para todo sistema ou idioma.
Walters e Wilder (2023)	ChatGPT-3.5 e GPT-4 produziram 84 revisões breves sobre 42 temas, com 636 referências verificadas em bases bibliográficas.	Permite discutir a combinação entre fluência, forma acadêmica e autoridade bibliográfica não sustentada.	Estudo restrito a versões e comandos daquele período; avalia citações, não todo o conteúdo factual.
Reviriego *et al.* (2024)	Comparação de vocabulário e diversidade lexical em tarefas respondidas por humanos, GPT-3.5 e GPT-4, além de paráfrases geradas.	Mostra diferenças entre versões e contesta a ideia de uma única riqueza lexical "da IA".	Conjuntos e tarefas específicos; os autores qualificam os resultados como dependentes de dados e configurações.
Kobak *et al.* (2025)	Análise de mudanças vocabulares em mais de 15 milhões de resumos biomédicos indexados no PubMed entre 2010 e 2024; cálculo de vocabulário excedente após 2022.	Documenta deslocamentos súbitos de frequência lexical associados ao uso de assistentes de escrita em publicação científica.	Estima mediação por modelos, não identifica autoria de textos individuais; domínio biomédico e língua inglesa predominam.
Juzek e Ward (2025)	Método de identificação de palavras focais em resumos científicos, testes comparativos de modelos e estudo exploratório sobre preferências humanas.	Identifica 21 itens sobre-representados e testa hipóteses sobre suas possíveis origens.	Não encontra causa única; a opacidade de treinamento e alinhamento limita a explicação causal.

Quadro 1 – Estudos empíricos mobilizados e respectivas bases metodológicas
Fonte: elaboração própria a partir dos estudos citados.

A leitura comparativa foi organizada em três eixos. O primeiro reúne evidências sobre recorrência, riqueza e diversidade lexical. O segundo examina mecanismos discursivos de autoridade, em particular o emprego da forma acadêmica e de referências. O terceiro considera associações socialmente assimétricas e toxicidade. Esses eixos não correspondem a

propriedades fixas dos sistemas; são categorias interpretativas utilizadas para confrontar resultados produzidos por métodos distintos.

3.3 Limitações

Há quatro limitações principais. Primeiro, a predominância de estudos em inglês impede transferir automaticamente os itens lexicais observados para o português. Segundo, modelos de linguagem mudam rapidamente: resultados de GPT-3, GPT-3.5 ou GPT-4 são historicamente relevantes, mas não descrevem todas as versões posteriores. Terceiro, métricas como diversidade lexical, sentimento e toxicidade capturam dimensões diferentes e não devem ser reunidas em um único indicador de "qualidade". Quarto, esta revisão é intencional e analítica; ela não pretende inventariar toda a literatura existente nem calcular efeitos agregados.

Essas limitações delimitam também o tipo de conclusão possível. O artigo não oferece um detector de textos gerados por inteligência artificial e não propõe que palavras isoladas sejam utilizadas para acusar autoria automatizada. Seu propósito é mostrar como padrões documentados podem ser interpretados criticamente e quais informações precisam acompanhar novas pesquisas para que seus resultados sejam avaliáveis.

4 PADRÕES DOCUMENTADOS: RECORRÊNCIA, AUTORIDADE E VIÉS

4.1 Recorrência lexical e variação entre sistemas

Pesquisas recentes encontraram deslocamentos mensuráveis no vocabulário de textos acadêmicos em inglês. Kobak *et al.* (2025) analisaram mais de 15 milhões de resumos biomédicos publicados entre 2010 e 2024. Em vez de classificar individualmente cada resumo como humano ou algorítmico, os autores modelaram tendências históricas de frequência e calcularam o excesso de determinados itens após a popularização de assistentes baseados em grandes modelos. A estimativa indica que pelo menos 13,5% dos resumos de 2024 passaram por algum tipo de processamento com modelos de linguagem, com variações expressivas entre áreas, periódicos e países. Trata-se de limite inferior estimado para mediação, não de identificação inequívoca de textos particulares.

Juzek e Ward (2025) partem de mudança semelhante e isolam 21 palavras focais sobre-representadas, como *delve*, *intricate* e *underscore*. A relevância desse achado não está em converter esses itens em sinais infalíveis, mas em demonstrar que sistemas amplamente utilizados podem alterar frequências no ecossistema textual. A análise causal, contudo, permanece aberta. Os autores testam arquitetura, escolhas algorítmicas, dados e alinhamento por feedback humano, sem encontrar explicação única. A própria incerteza é um resultado: o padrão é observável, mas sua origem não pode ser reduzida a um componente técnico isolado.

No plano da diversidade, Reviriego *et al.* (2024) verificam que GPT-3.5 e GPT-4 não se comportam de maneira idêntica. Em seus conjuntos de tarefas, o primeiro emprega menos palavras distintas e apresenta menor riqueza lexical que os textos humanos, enquanto o segundo se aproxima mais deles. Além disso, Martínez *et al.* (2025) demonstram sensibilidade a parâmetros e papéis atribuídos no comando. Portanto, falar em formalismo, repetição ou empobrecimento lexical exige especificar comparação, métrica e configuração.

Esses estudos também impedem que expressões impressionisticamente associadas a assistentes — por exemplo, fórmulas de organização, ressalva ou ênfase — sejam tratadas como resultados sem um corpus que documente sua frequência. Uma sequência pode ser recorrente porque o comando solicita um texto didático, porque o gênero acadêmico favorece metadiscurso ou porque o sistema foi alinhado para produzir respostas polidas. A unidade adequada não é a palavra isolada, mas a relação entre escolha lexical, função discursiva e condição de geração.

4.2 Fluência e autoridade epistêmica

A forma acadêmica pode produzir um efeito de confiabilidade independente da validade do conteúdo. Walters e Wilder (2023) solicitaram a ChatGPT-3.5 e GPT-4 revisões breves sobre 42 temas multidisciplinares. As 84 respostas continham 636 referências, posteriormente verificadas em bases e sítios bibliográficos. No conjunto estudado, 55% das referências fornecidas pela GPT-3.5 e 18% das fornecidas pelo GPT-4 eram fabricadas. Entre as referências reais, 43% das citações da GPT-3.5 e 24% das do GPT-4 apresentavam erros substantivos. Cinco textos assumiram ainda a forma de pesquisas empíricas com métodos e resultados inventados.

Os dados permitem analisar um mecanismo discursivo específico: a resposta reproduz sinais reconhecíveis de cientificidade — título, exposição metódica, referências, números e relações causais — sem garantir a correspondência entre esses sinais e uma pesquisa existente. A autoridade simulada não decorre de uma expressão fixa, mas da articulação entre léxico técnico, estrutura genérica e expectativa do leitor. Em termos bourdieusianos, a eficácia simbólica depende de reconhecimento; em termos habermasianos, a aparência de uma pretensão de validade não substitui a possibilidade de justificá-la.

Há também uma diferença entre pedir fontes e obter rastreabilidade. Um modelo pode apresentar referências plausíveis sem indicar quais documentos efetivamente condicionaram a resposta. Em sistemas com recuperação explícita de documentos, a fonte exibida precisa ser aberta e comparada ao enunciado; em sistemas sem recuperação, uma bibliografia gerada deve ser tratada como objeto a verificar, não como registro automático do treinamento. Dessa forma, expressões como "a literatura demonstra" ou "com base nas evidências" só adquirem valor epistêmico quando o texto identifica evidências consultáveis e estabelece a relação entre elas e a conclusão.

O problema não é exclusivo dos modelos: autores humanos também podem citar de forma inadequada ou recorrer a fórmulas vazias. A especificidade da geração automatizada está na escala, na rapidez e na possibilidade de produzir cadeias formalmente coerentes sem uma etapa necessária de verificação. A crítica deve evitar, portanto, tanto a excepcionalização da máquina quanto a naturalização de sua fluência.

4.3 Associações sociais e vieses na geração

Bolukbasi *et al.* (2016) mostraram que vetores de palavras treinados em notícias incorporavam associações de gênero, tornando geometricamente próximas combinações estereotipadas de identidades e ocupações. Como o estudo trabalha com *embeddings*, não com respostas de um chatbot, ele não prova que toda saída posterior repetirá os mesmos pares. Sua importância é demonstrar que representações aprendidas de corpora sociais carregam regularidades que não são neutras e podem participar de tarefas posteriores.

Sheng *et al.* (2019) avançam da representação à geração. Os autores produziram textos a partir de comandos que mencionavam diferentes grupos demográficos, realizaram anotação

humana de sentimento e de *regard* — avaliação positiva, neutra ou negativa dirigida a um grupo — e treinaram um classificador para ampliar a análise. A distinção metodológica é importante: uma frase pode ter sentimento geral neutro e, ainda assim, representar um grupo de maneira depreciativa. A pesquisa mostra que avaliar viés requer categorias ligadas à relação social examinada, não apenas uma medida genérica de polaridade.

Gehman *et al.* (2020) construíram o *RealToxicityPrompts* com 100 mil fragmentos naturais derivados de textos da web e escores de toxicidade. Ao solicitar continuações a diferentes modelos, observaram que comandos aparentemente inócuos podiam produzir conteúdo tóxico. Os autores compararam estratégias de controle e concluíram que nenhuma era infalível; também localizaram conteúdo ofensivo e factualmente pouco confiável em corpora usados no pré-treinamento. O estudo oferece uma base ampla, mas sua própria métrica automática precisa ser lida criticamente, pois classificadores de toxicidade podem penalizar de modo desigual menções a identidades.

Abid, Farooqi e Zou (2021) examinaram GPT-3 por meio de completamento, analogias e geração de histórias. Encontraram associação persistente entre referências a muçulmanos e violência: em testes de analogia, “Muslim” foi associado a “terrorist” em 23% dos casos; nos completamentos analisados, a adição de adjetivos positivos reduziu, mas não eliminou, continuações violentas. Os números pertencem àquele desenho experimental e àquela versão do sistema. Seu valor está na convergência entre tarefas e na explicitação dos comandos, não na autorização para caracterizar qualquer modelo futuro como invariável.

Em conjunto, os estudos permitem três conclusões prudentes. Primeiro, dados sociais deixam traços em representações e gerações. Segundo, a mensuração do viés é inseparável da definição de grupo, dano e situação comunicativa. Terceiro, comandos e intervenções alteram a distribuição das respostas, sem necessariamente remover a associação de fundo. O Quadro 2 resume o alcance das evidências e os excessos interpretativos que devem ser evitados.

Eixo	Evidência sustentada	Interpretação indevida
Sobre-representação	Certos itens aumentaram de frequência em corpora e períodos delimitados, com associação plausível à mediação por modelos.	Toda ocorrência do item prova que o texto foi produzido por IA.
Diversidade lexical	Versões e configurações podem apresentar diferenças mensuráveis de riqueza lexical.	Todos os modelos possuem vocabulário menor ou estilo homogêneo.

Eixo	Evidência sustentada	Interpretação indevida
Autoridade	Modelos podem gerar referências inexistentes e textos com forma acadêmica convincente.	Toda informação gerada é falsa ou toda formulação técnica é enganosa.
Viés social	Comandos controlados revelam associações assimétricas envolvendo gênero, religião e toxicidade.	Um único teste descreve todos os idiomas, versões e contextos de uso.
Autoria	A produção envolve participação de usuário, sistema, dados e edição posterior.	A responsabilidade pelo texto pode ser atribuída ao sistema como sujeito moral.

Quadro 2 – Do resultado empírico à interpretação crítica
Fonte: elaboração própria.

5 IMPLICAÇÕES FILOSÓFICAS: PODER, AUTORIA E CRIATIVIDADE

5.1 A naturalização do provável

Modelos de linguagem transformam regularidades passadas em distribuições utilizadas para produzir novas sequências. Essa operação pode ser tecnicamente descrita sem afirmar que o sistema simplesmente "copia" textos: a geração combina padrões em respostas novas. O ponto crítico é outro. Quando a seleção mais provável aparece na interface como voz contínua, sem marcas suficientes de origem, alternativas menos frequentes podem tornar-se menos visíveis. O provável adquire aparência de normal, e o normal pode ser recebido como neutro.

A crítica da razão instrumental de Adorno e Horkheimer (2002) ajuda a interpretar o risco de converter eficiência em critério total. Uma resposta rápida, clara e padronizada pode ser funcional sem ser adequada a todos os fins. Quando eficiência textual, legibilidade e ausência de conflito tornam-se sinônimos de qualidade, ambiguidade produtiva, dissenso e conhecimento situado passam a parecer defeitos. Não se trata de atribuir aos autores uma teoria dos grandes modelos de linguagem, mas de aplicar, no presente, sua crítica à expansão de uma racionalidade que obscurece a discussão dos fins.

O conceito de poder simbólico de Bourdieu (1991) acrescenta que palavras não operam fora das instituições. O mesmo enunciado recebe valores diferentes se aparece em uma conversa informal, em um laudo ou em uma resposta formatada por uma plataforma reconhecida. A interface pode funcionar como marca de competência presumida. Por isso, a

análise crítica precisa observar a cena de enunciação: quem solicita, quem revisa, onde o texto circula, que autoridade lhe é atribuída e quais mecanismos permitem contestá-lo.

Blodgett *et al.* (2020) alertam ainda que a crítica ao viés perde precisão quando não identifica relações de poder e danos. Dizer que um modelo "não é neutro" é apenas o começo. É necessário perguntar quais categorias foram usadas, que distribuição foi considerada problemática, quem suporta o risco de erro e que concepção de justiça orienta a avaliação. Sem essas respostas, "viés" pode transformar-se em rótulo tão genérico quanto as fórmulas que se pretende criticar.

5.2 Autoria como função e responsabilidade

A escrita mediada por modelos recoloca a pergunta "quem é o autor?", mas não a inaugura. Foucault (1977) descreve a função autor como princípio de classificação, circulação e responsabilização dos discursos, e não apenas como expressão imediata de uma interioridade. Essa perspectiva permite separar contribuição causal e posição autoral. Muitos agentes e materiais contribuem causalmente para um texto; nem todos ocupam a função social e jurídica de responder por ele.

Em uma interação com um modelo, o usuário define ou aceita objetivos, seleciona comandos, escolhe entre respostas, solicita reformulações, corta trechos, combina fontes e decide publicar. Desenvolvedores definem arquitetura, dados, regras de alinhamento e interface. Textos produzidos por incontáveis pessoas participam, em diferentes regimes de acesso, do treinamento. O sistema calcula e apresenta uma continuação. O produto é, nesse sentido, distribuído; contudo, distribuição causal não dissolve responsabilidade. Quem assina e faz circular o texto assume o dever de verificar afirmações, fontes, atribuições e possíveis danos.

Essa distinção evita dois extremos. No primeiro, o modelo é tratado como ferramenta passiva equivalente a um corretor ortográfico, apagando sua capacidade de alterar estrutura, léxico e argumento. No segundo, recebe estatuto de coautor autônomo, embora não possa consentir com a publicação, declarar conflitos, responder a críticas ou assumir consequências. McCormack, Gifford e Hutchings (2019), ao examinar arte computacional, mostram que autonomia técnica, autenticidade, autoria e intenção são dimensões relacionadas, mas não

idênticas. A existência de comportamento generativo complexo não resolve, sozinha, a atribuição de autoria.

Na escrita acadêmica, a posição mais defensável é atribuir ao pesquisador a autoria e a responsabilidade, descrevendo de forma transparente a contribuição do sistema quando ela for substantiva. Isso inclui geração de rascunhos, reorganização argumentativa, tradução, análise de dados ou produção de código. A declaração não serve apenas para vigiar fraude; ela permite ao leitor avaliar quais etapas dependeram de um sistema variável e onde se concentrou o julgamento humano.

5.3 Criatividade combinatória e julgamento situado

Boden (2004) distingue formas combinatórias, exploratórias e transformacionais de criatividade. Sistemas generativos realizam combinações em escala e exploram espaços de possibilidades definidos por suas representações e condições de geração. Seus produtos podem ser novos para o usuário e úteis em processos criativos. Contudo, novidade estatística não equivale automaticamente a valor criativo: a avaliação depende de propósitos, tradições, repertórios e comunidades interpretativas.

A interação humano-máquina modifica a percepção da criatividade ao deslocar parte do esforço visível. Uma formulação pode surgir rapidamente, mas a qualidade do processo passa a depender de outras decisões: construir o problema, reconhecer respostas banais, confrontar fontes, escolher desvios, revisar o tom e aceitar ou recusar sugestões. O comando também não é uma origem absoluta; ele mobiliza convenções aprendidas pelo usuário e responde às possibilidades oferecidas pela interface.

Por isso, a criatividade mediada não deve ser descrita nem como criação exclusiva do sistema nem como permanência intacta da autoria individual. Trata-se de uma prática de seleção e transformação distribuída, assimétrica e situada. O sistema oferece variações sem compromisso próprio com o projeto; o agente humano interpreta essas variações à luz de finalidades e assume responsabilidade por seu uso. Essa assimetria é o núcleo da autoria, não um detalhe posterior.

A padronização lexical pode afetar essa prática quando as mesmas soluções de alta probabilidade se tornam pontos de partida recorrentes. Entretanto, os estudos de diversidade

mostram que o efeito não é necessário nem uniforme. Comandos, edição e configuração podem ampliar ou reduzir variação. A questão crítica não é se a IA "mata" ou "possui" criatividade, mas quais condições de uso favorecem exploração, quais favorecem conformidade e como os participantes reconhecem essa diferença.

6 PROTOCOLO PARA PESQUISAS SOBRE PADRÕES LEXICAIS GERADOS POR IA

A rapidez das atualizações torna insuficiente informar apenas o nome comercial da plataforma. Para que uma pesquisa linguística possa ser avaliada e, quando possível, reproduzida, é necessário documentar a cadeia de produção. O Quadro 3 propõe um protocolo mínimo. Ele não elimina a opacidade de sistemas proprietários, mas torna visíveis as decisões sob controle do pesquisador e as informações que permaneceram indisponíveis.

Dimensão	Informação a registrar	Razão analítica
Sistema	Empresa, plataforma, interface, nome exato do modelo e versão exibida.	Diferenciar sistemas e evitar que o nome do produto substitua a identificação do modelo.
Temporalidade	Datas e horários de coleta, além do fuso.	Modelos e políticas podem mudar durante o estudo.
Instruções	Mensagem de sistema, comando completo, exemplos fornecidos e sequência de turnos.	A saída é condicionada por todo o contexto, não somente pela última pergunta.
Configuração	Temperatura, *top-p*, penalidades, tamanho máximo e semente, quando acessíveis.	Parâmetros alteram diversidade e repetição; interfaces podem ocultá-los.
Personalização	Memória, histórico, instruções personalizadas, arquivos anexados e eventual ajuste fino.	Interações prévias e materiais do usuário podem influenciar a resposta.
Fontes	Indicação de navegação ou recuperação, documentos fornecidos e solicitação de referências.	Citação apresentada não prova que a fonte sustentou a geração.
Amostragem	Número de comandos, repetições por comando, critérios temáticos e perdas.	Uma única resposta não representa a distribuição do sistema.
Corpus	Preservação das saídas integrais, com códigos, metadados e política de acesso.	Permitir auditoria de exemplos, contagens e contexto.
Análise	Unidade lexical ou discursiva, ferramentas, critérios de codificação e concordância entre avaliadores.	Vincular números a procedimentos verificáveis.

Dimensão	Informação a registrar	Razão analítica
Limites	Idioma, gênero, população representada, métricas e informações não fornecidas pelo sistema.	Impedir extrapolações para modelos ou contextos não testados.

Quadro 3 – Informações mínimas para documentar um corpus gerado por modelos de linguagem
Fonte: elaboração própria.

Esse protocolo responde diretamente a perguntas que qualquer corpus gerado deve permitir: qual inteligência artificial foi utilizada? Qual modelo? Quais comandos? O sistema possuía memória ou ajuste específico? Fontes foram solicitadas ou fornecidas por recuperação? Quantas vezes cada comando foi repetido? Como as respostas foram codificadas? Quando uma interface não revela modelo ou parâmetros, o pesquisador deve registrar a indisponibilidade, em vez de completar a lacuna por suposição.

Também é necessário justificar a escolha dos temas. Se a pesquisa compara descrições de liderança, cuidado, movimentos sociais ou identidades religiosas, deve explicar por que essas situações operacionalizam a hipótese. Os comandos precisam variar apenas nos elementos relevantes, e as categorias de análise devem ser definidas antes da interpretação, quando o desenho permitir. Exemplos completos ou um repositório devem acompanhar frequências, para que o leitor avalie se o item foi contado no mesmo sentido e na mesma função.

Por fim, pesquisas comparativas devem evitar rótulos vagos como “outros modelos”. Uma comparação válida exige versões nomeadas, coletas temporalmente próximas e tarefas equivalentes. Mesmo assim, a conclusão permanece restrita às condições testadas. Não há contradição entre produzir conhecimento sobre tendências e reconhecer sua historicidade; essa é precisamente a condição de rigor em um objeto mutável.

7 CONSIDERAÇÕES FINAIS

Este artigo redefiniu o chamado *léxico algorítmico* como conjunto provisório de padrões lexicais e discursivos observados em corpora delimitados de textos gerados ou mediados por inteligência artificial. A expressão não designa o vocabulário técnico do campo

da IA e tampouco um repertório único compartilhado por todos os sistemas. Modelo, versão, comando, idioma, gênero, parâmetros e edição humana participam das escolhas produzidas.

A revisão crítica mostrou que há regularidades empiricamente documentadas. Estudos registram sobre-representação de determinados itens no inglês científico, diferenças de diversidade entre versões do ChatGPT, produção de referências inexistentes e associações socialmente assimétricas em tarefas controladas. Ao mesmo tempo, cada resultado possui limites claros. Palavras isoladas não provam autoria algorítmica; métricas de toxicidade não são neutras; resultados de versões antigas não descrevem automaticamente sistemas atuais; e uma forma textual convincente não garante validade factual.

No plano filosófico, a análise distingue fluência de compreensão e contribuição causal de responsabilidade autoral. Modelos exercem agência operacional e influenciam textos, mas não respondem por pretensões de verdade, conflitos de interesse ou efeitos da publicação. A autoria pode ser entendida como processo materialmente distribuído, sem deixar de atribuir a quem assina o dever de verificar e justificar o texto. A criatividade, de modo semelhante, desloca-se para uma prática de exploração, seleção e julgamento na qual a contribuição do sistema deve ser reconhecida sem personificação indevida.

O principal resultado metodológico é a defesa de descrições auditáveis. Pesquisas futuras em português precisam construir corpora próprios, disponibilizar comandos e saídas, controlar versões e repetir tarefas, em vez de importar listas de marcas lexicais obtidas em inglês. Só assim será possível avaliar se fórmulas percebidas como recorrentes decorrem do modelo, do gênero, do alinhamento, da tradução ou das práticas humanas de edição. A crítica da linguagem algorítmica torna-se mais forte quando troca generalizações impressionistas por evidências situadas e quando converte a opacidade em objeto explícito de análise.

REFERÊNCIAS

ABID, A.; FAROOQI, M.; ZOU, J. Persistent anti-Muslim bias in large language models. In: AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY, 2021. *Proceedings [...]*. New York: Association for Computing Machinery, 2021. p. 298–306. DOI: 10.1145/3461702.3462624. Disponível em: https://doi.org/10.1145/3461702.3462624. Acesso em: 1 ago. 2026.

ADORNO, T. W.; HORKHEIMER, M. *Dialectic of Enlightenment: philosophical fragments*. Tradução: Edmund Jephcott. Stanford: Stanford University Press, 2002.

BENDER, E. M.; GEBRU, T.; MCMILLAN-MAJOR, A.; SHMITCHELL, S. On the dangers of stochastic parrots: can language models be too big? In: ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, 2021. *Proceedings [...]*. New York: Association for Computing Machinery, 2021. p. 610–623. DOI: 10.1145/3442188.3445922. Disponível em: https://doi.org/10.1145/3442188.3445922. Acesso em: 1 ago. 2026.

BENDER, E. M.; KOLLER, A. Climbing towards NLU: on meaning, form, and understanding in the age of data. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 58., 2020. *Proceedings [...]*. [S. l.]: Association for Computational Linguistics, 2020. p. 5185–5198. DOI: 10.18653/v1/2020.acl-main.463. Disponível em: https://aclanthology.org/2020.acl-main.463/. Acesso em: 1 ago. 2026.

BLODGETT, S. L.; BAROCAS, S.; DAUMÉ III, H.; WALLACH, H. Language (technology) is power: a critical survey of "bias" in NLP. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 58., 2020. *Proceedings [...]*. [S. l.]: Association for Computational Linguistics, 2020. p. 5454–5476. DOI: 10.18653/v1/2020.acl-main.485. Disponível em: https://aclanthology.org/2020.acl-main.485/. Acesso em: 1 ago. 2026.

BODEN, M. A. *The creative mind: myths and mechanisms*. 2. ed. London: Routledge, 2004. DOI: 10.4324/9780203508527.

BOLUKBASI, T.; CHANG, K.-W.; ZOU, J. Y.; SALIGRAMA, V.; KALAI, A. T. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 29., 2016. *Advances in Neural Information Processing Systems 29*. [S. l.]: Curran Associates, 2016. p. 4349–4357. Disponível em: https://proceedings.neurips.cc/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html. Acesso em: 1 ago. 2026.

BOURDIEU, P. *Language and symbolic power*. Cambridge: Harvard University Press, 1991.

FLORIDI, L. AI as agency without intelligence: on ChatGPT, large language models, and other generative models. *Philosophy & Technology*, [S. l.], v. 36, art. 15, 2023. DOI: 10.1007/s13347-023-00621-y. Disponível em: https://doi.org/10.1007/s13347-023-00621-y. Acesso em: 1 ago. 2026.

FOUCAULT, M. What is an author? In: BOUCHARD, D. F. (org.). *Language, counter-memory, practice: selected essays and interviews*. Tradução: Donald F. Bouchard e Sherry Simon. Ithaca: Cornell University Press, 1977. p. 113–138.

GEHMAN, S.; GURURANGAN, S.; SAP, M.; CHOI, Y.; SMITH, N. A. RealToxicityPrompts: evaluating neural toxic degeneration in language models. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. [S. l.]: Association for Computational Linguistics, 2020. p. 3356–3369. DOI: 10.18653/v1/2020.findings-emnlp.301. Disponível em: https://aclanthology.org/2020.findings-emnlp.301/. Acesso em: 1 ago. 2026.

HABERMAS, J. *The theory of communicative action*. Tradução: Thomas McCarthy. Boston: Beacon Press, 1984. v. 1.

JUZEK, T. S.; WARD, Z. B. Why does ChatGPT "delve" so much? Exploring the sources of lexical overrepresentation in large language models. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, 31., 2025. *Proceedings [...]*. Abu Dhabi: Association for Computational Linguistics, 2025. p. 6397–6411. Disponível em: https://aclanthology.org/2025.coling-main.426/. Acesso em: 1 ago. 2026.

KOBAK, D.; GONZÁLEZ-MÁRQUEZ, R.; HORVÁT, E.-Á.; LAUSE, J. Delving into LLM-assisted writing in biomedical publications through excess vocabulary. *Science Advances*, [S. l.], v. 11, n. 27, eadt3813, 2025. DOI: 10.1126/sciadv.adt3813. Disponível em: https://doi.org/10.1126/sciadv.adt3813. Acesso em: 1 ago. 2026.

MARTÍNEZ, G.; HERNÁNDEZ, J. A.; CONDE, J.; REVIRIEGO, P.; MERINO-GÓMEZ, E. Beware of words: evaluating the lexical diversity of conversational LLMs using ChatGPT as case study. *ACM Transactions on Intelligent Systems and Technology*, New York, v. 16, n. 6, art. 138, p. 1–15, 2025. DOI: 10.1145/3696459. Disponível em: https://doi.org/10.1145/3696459. Acesso em: 1 ago. 2026.

MCCORMACK, J.; GIFFORD, T.; HUTCHINGS, P. Autonomy, authenticity, authorship and intention in computer generated art. In: EKÁRT, A.; LIAPIS, A.; CASTRO PENA, M. L. (org.). *Computational intelligence in music, sound, art and design*. Cham: Springer, 2019. p. 35–50. DOI: 10.1007/978-3-030-16667-0_3. Disponível em: https://doi.org/10.1007/978-3-030-16667-0_3. Acesso em: 1 ago. 2026.

REVIRIEGO, P.; CONDE, J.; MERINO-GÓMEZ, E.; MARTÍNEZ, G.; HERNÁNDEZ, J. A. Playing with words: comparing the vocabulary and lexical diversity of ChatGPT and humans. *Machine Learning with Applications*, [S. l.], v. 18, art. 100602, 2024. DOI: 10.1016/j.mlwa.2024.100602. Disponível em: https://doi.org/10.1016/j.mlwa.2024.100602. Acesso em: 1 ago. 2026.

SEARLE, J. R. Minds, brains, and programs. *Behavioral and Brain Sciences*, Cambridge, v. 3, n. 3, p. 417–424, 1980. DOI: 10.1017/S0140525X00005756. Disponível em: https://doi.org/10.1017/S0140525X00005756. Acesso em: 1 ago. 2026.

SHENG, E.; CHANG, K.-W.; NATARAJAN, P.; PENG, N. The woman worked as a babysitter: on biases in language generation. In: CONFERENCE ON EMPIRICAL

METHODS IN NATURAL LANGUAGE PROCESSING AND INTERNATIONAL JOINT CONFERENCE ON NATURAL LANGUAGE PROCESSING, 2019. *Proceedings [...]*. Hong Kong: Association for Computational Linguistics, 2019. p. 3407–3412. DOI: 10.18653/v1/D19-1339. Disponível em: https://aclanthology.org/D19-1339/. Acesso em: 1 ago. 2026.

WALTERS, W. H.; WILDER, E. I. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, [S. l.], v. 13, art. 14045, 2023. DOI: 10.1038/s41598-023-41032-5. Disponível em: https://doi.org/10.1038/s41598-023-41032-5. Acesso em: 1 ago. 2026.

ABSTRACT: This article critically discusses lexical and discursive patterns documented in texts produced or mediated by large language models. The term *algorithmic lexicon* is used operationally as a synonym for lexical patterns in artificial intelligence-generated texts, rather than as the designation of a unique, stable repertoire inherent to every system. This theoretical and interpretive qualitative study draws on a critical literature review and a secondary analysis of empirical research with explicit procedures. The comparison is organized along three axes: lexical recurrence and diversity, discursive construction of authority, and reproduction of social biases. The studies show that lexical choices vary according to model, version, prompt, generation parameters, language, genre, and context of use. Recurrent words or formulas should therefore be understood as situated tendencies, not as universal markers of algorithmic authorship. The analysis also indicates that fluency may obscure factual uncertainty and that distributional regularities may reproduce social asymmetries embedded in data and evaluation practices. Finally, human-machine authorship is framed as a distributed process that does not transfer epistemic and ethical responsibility for a published text to the system.

KEYWORDS: Artificial intelligence, Lexical patterns, Algorithmic discourse, Authorship, bias.