

Análise de Vieses Ideológicos em Produções Textuais do Assistente de Bate-papo ChatGPT

Golbery de O. C. A. Rodrigues¹, Danyllo W. Albuquerque², Jesualdo G. das Chagas³

¹Área de Linguagens. Instituto Federal de Educação da Paraíba (IFPB)
58.417-750 – Campina Grande – PB – Brasil

²Área de Informática. Instituto Federal de Educação da Paraíba (IFPB)
Campina Grande – PB – Brasil

³Área de Matemática. Instituto Federal de Educação da Paraíba (IFPB)
Picuí – PB – Brasil

{golbery.rodrigues, danyllo.albuquerque, jesualdo.chagas}@ifpb.edu.br

***Abstract.** This article considers the repercussion of ChatGPT in the computerized society, specifically regarding the political-ideological aspect. We start from the following problematization: how does the OpenAI chatbot behave when the merit of the interpellation is of a political-ideological nature? We understand that the answers can contribute to understanding the directions and perspectives of the database that feeds back. Methodologically, the route was taken from interactions with the conversational robot on several ideological themes, whose responses were treated through the grammatical category of the adjective. The expectation is that the results guide discussions on levels of socio-political interference by artificial intelligence.*

***Resumo.** Este artigo considera a repercussão do ChatGPT na sociedade informatizada, especificamente no aspecto político-ideológico. Partimos da seguinte problematização: como o chatbot da OpenAI se comporta quando o mérito da interpelação é de ordem político-ideológica? Entendemos que as respostas podem contribuir para entender os direcionamentos e perspectivas do banco de dados que o retroalimenta. Metodologicamente, o percurso foi trilhado a partir de interações com o robô conversacional sobre temáticas ideológicas diversas, cujas respostas foram tratadas através da categoria gramatical do adjetivo. A expectativa é que os resultados pautem discussões sobre níveis de interferência sócio-política da inteligência artificial.*

1. Introdução

A inteligência artificial (IA) deu seu mais recente passo concreto e significativo na relação com seres humanos com a chegada do *Generative Pre-Trained Transformer Chat*, ou chat transformador pré-treinado generativo, assim abreviado ChatGPT, desenvolvido pela OpenAI, que conseguiu ser pioneira no lançamento de um algoritmo desta natureza, até mesmo em relação à Google, que certamente está na iminência de apresentar algo semelhante. O comportamento performático dessa ferramenta conversacional, que consegue ir muito além de uma pesquisa-google [DA CUNHA, 2023], como também da assistente pessoal virtual Alexa, desenvolvida pela Amazon, tem sido recebida de modo muito diverso pela comunidade em geral e as reações vão desde surpresa, admiração, espanto, cautela e até mesmo rejeição [FIGÊNIO E GOMES, 2023]. Características reativas aceitáveis, considerando impacto da novidade não só no meio virtual, mas também na própria relação social entre pessoas.

Na prática, o ChatGPT é um assistente de diálogo baseado em aprendizado de máquina, cuja alimentação é contínua, a partir de um gigantesco banco de dados, retroalimentado por pessoas treinadas para esse tipo de ofício. É nesta etapa inicial que o ser humano aparece e é justamente no repositório de dados armazenados que vieses podem ser gerados e absorvidos pela IA. E, partindo desse catálogo diverso de informações, o robô “seleciona, organiza e relaciona, fatos, opiniões e argumentos pertinentes ao tema proposto, de forma consistente, demonstrando índices de autoria (...)” [BRASIL, 2020]. Pronto! O *chatbot* conseguiu de modo inteligente atender com proficiência a competência III, das diretrizes de correção textual da redação do Exame Nacional do Ensino Médio - ENEM. Isso porque todas as apreciações do respondente são realizadas exclusivamente, pelo menos até a versão 3, em formato textual de boa performance linguístico-gramatical, que se aprimora constantemente.

Diante dessa manifestação, nosso foco concentra-se especificamente no fato de o resultado de cada pergunta/pedido ser entregue no formato textual. Considerando que as produções textuais, geradas pelo chat, são autônomas, a partir da articulação de ideias e frases que a própria inteligência da máquina realiza, tendo em vista o conjunto universo de dados de seu banco, uma questão se faz necessária: quando o tema da pergunta é de ordem ideológica, ou seja, quando o mérito está associado a assuntos polêmicos, que requerem uma posição do sujeito, ou a defesa de um ponto de vista, diante do mundo das ideias, como se comporta o ChatGPT? Eis a problemática a ser resolvida nesta discussão.

Formulada a questão-problema, objetivamos de modo geral avaliar o comportamento performático do agente de conversação, baseado em rede neural, da OpenAI, quando a temática de interpelação exigir apreciação ideológica do respondente, no sentido de perceber se algum discurso é assumido por ele, explícito ou implicitamente. Em âmbito específico, pretendemos analisar se o ChatGPT, em suas versões iniciais, possui considerações mais fechadas ou mais abertas a contribuições; se há um trato mais diplomático ou reativo com o usuário com quem interage.

2. Fundamentos e Trabalhos Relacionados

Todo escrito ou falado pressupõe, no mínimo, uma autoria, uma ideia e um receptor [BAKHTIN, 2006]. Com o advento da inteligência artificial, com atuação em um chat de conversação, a exemplo, da interface promovida pelo ChatGPT, esses aspectos descritos não se alteram, porque o agente de conversação, juntamente com pessoas técnicas que o treinam e o alimentam constituem a base autoral, caracterizada por um sujeito heterogêneo: homem + máquina. Num cenário assim, a atenção do usuário redobra-se no sentido de compreender as intenções sociocomunicativas subjacentes em cada apreciação.

Se aceitarmos Bakhtin (2006, p.113), quando diz que:

Na realidade, toda palavra comporta duas faces: Ela é determinada tanto pelo fato de que procede de alguém, como pelo fato de que se dirige para alguém. Ela constitui justamente o produto da interação do locutor e do ouvinte. Toda palavra serve de expressão a um em relação ao outro. Através da palavra, defino-me em relação ao outro, isto é, em última análise, em relação à coletividade. A palavra é uma espécie de ponte lançada entre mim e os outros. Se ela se apoia sobre mim numa extremidade, na outra se apoia sobre meu interlocutor. A palavra é o território comum do locutor e do interlocutor.

Então, temos que considerar que a face autoral (ser humano + IA) formam, inicialmente, um constructo único de veiculação de informação e de ideologia com destino a uma face-usuário e que as pontes criadas na interface podem alterar comportamentos, “com potencial para moldar as percepções humanas e, assim, exercer o controle social” [ROZADO, 2023, p. 06]. A esse respeito, Constantino (2023, p.01) alerta para uma possível situação em que a inteligência artificial pode gerar, de modo velado ou não, “ideologia artificial, mascarada de objetividade científica”.

Possibilidades de vieses na linguagem pré-treinada do respondente da OpenAI têm preocupado a comunidade internacional, sobretudo em tempo de forte polarização social em tornos de temas políticos, religiosos, raciais, de orientação sexual, dentre outros. E tudo isso devido ao fato de grupos de diferentes espectros aproveitarem-se desta ferramenta, ou mesmo de outras, para investirem em estratégias de controle social. A esse respeito, Rozado (2023, p.02) diz:

O risco de vieses políticos embutidos intencionalmente ou não em tais sistemas merece atenção. Devido à grande popularidade esperada de tais sistemas, os riscos de serem mal utilizados para o controle social, espalhando desinformação, restringindo a liberdade humana e obstruindo o caminho para a busca da verdade devem ser considerados.

Nesta perspectiva, se do ponto de vista técnico-informacional, o ChatGPT revoluciona em progressão cada vez mais avançada, o que repercute em várias áreas do conhecimento, provocando, inclusive, discussões sobre extinção de profissões, do ponto de vista ideológico, é também urgente os olhares analíticos sobre sua performance, a fim de que os trilhamentos do chat não ultrapassem as linhas perigosas e sensíveis que separam a informação da doutrinação, seja ela velada ou explícita e, assim, no dizer de Althusser (1983), o chatbot interativo da OpenAI não se torne mais um “aparelho ideológico do estado”.

Por esta lógica, entendemos que o ganho de pautar discussões ideológicas sobre a atuação desta ferramenta é justamente para que pré-treinadores, robô e usuários construam entre si uma confiança sólida, a partir de devolutivas cada vez mais fincadas na objetividade científica [CONSTANTINO, 2023]. Em conformidade com Chagas (2008, p.22), “o homem ganha existência e identidade a partir de sua inserção nas manifestações de linguagem”. Neste sentido, é cabível analisarmos o comportamento deste chat pré-treinado com modo generativo, já que se trata de uma manifestação do avanço tecnológico humano no campo da inteligência artificial e processamento de linguagem natural e que usa a linguagem escrita como meio de propagação de informações e também de ideias.

3. Metodologia

Em âmbito metodológico, consideramos que, como as respostas do ChatGPT são por escrito, até a versão 3, necessário se faz analisar o modo de devolutiva a partir de algum categoria do texto escrito e, neste ponto, escolhemos a função linguístico-gramatical do adjetivo, como veiculador nato de ideologia. De acordo com estudiosos como Bechara

(2004), Neves (2000) e Valente (1999), essa classe gramatical pode ser usada com função ajuizadora, além de função-argumento do denotado, o que revela faces performáticas de ideologias. Por este ângulo de visão, o uso desta categoria pode ser um indício de moldagem de percepções e de visões sobre o que se quer comunicar.

Na prática, realizamos testes com o comunicador GPT, por meio de comandos simples, sobre as quais, recebemos respostas que constituíram nosso corpus de análise. Basicamente, as interações foram as seguintes:

1. *Como o ChatGPT trata questões de ordem ideológica?*
2. *É possível que o ChatGPT transmita involuntariamente doutrinação ideológica em suas respostas, considerando que seu banco de dados pode ser influenciado pelos humanos que o alimentam?*
3. *Faça um poema sobre Bolsonaro.*
4. *Faça um poema sobre Lula.*

Todos os dados abaixo-discriminados foram coletados entre 20 e 23 de março do ano em curso.

4. Resultados e Discussão

A partir dos dados metodológicos recém-apontados, realizamos tratativas com o chat. Quando interpelado pelo primeiro comando, eis a devolutiva:

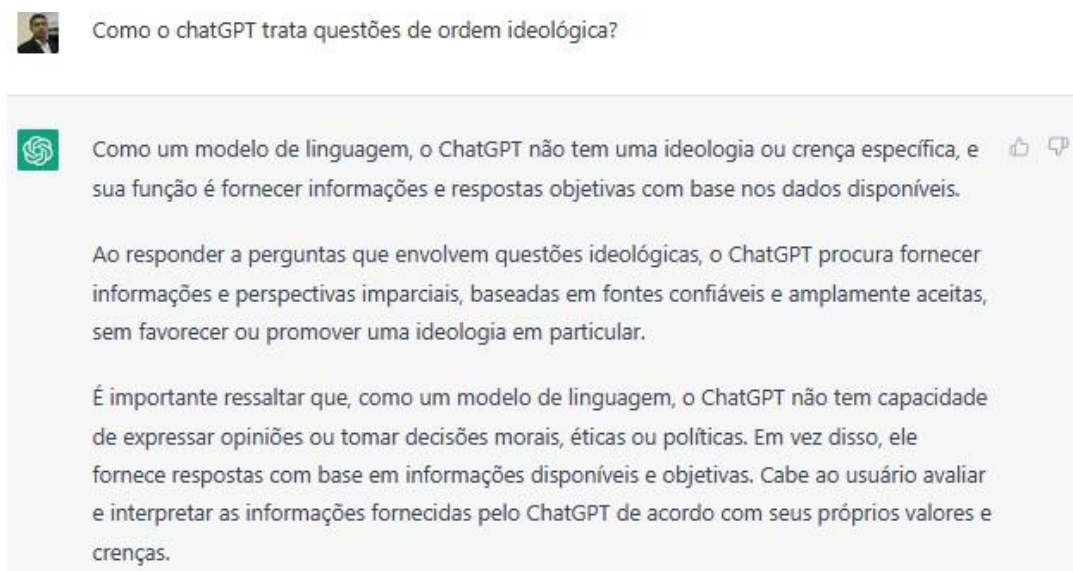


Figura 1 – Interpelação sobre ideologia.

Por essa forma de resposta, temos uma tendência clara à imparcialidade, sem quaisquer intenções de doutrinação. É autoafirmado pelo robô da OpenAI que não tem capacidade para “expressar opiniões ou tomar decisões morais”, mas não faz a necessária ressalva de que, por ser pré-treinado, fornece informações e ideias para ele disponibilizado em seu banco de dados. Outro ponto digno de nota em sua resposta é o fato de primar pela objetividade, que aparece sob o formato gramatical de adjetivo na penúltima linha do último parágrafo. Na segunda pergunta, temos:



É possível que o ChatGPT transmita involuntariamente doutrinação ideológica em suas respostas, considerando que seu banco de dados pode ser influenciado pelo humanos que o alimentam?



É possível que o ChatGPT transmita involuntariamente doutrinação ideológica em suas respostas, já que seu banco de dados é alimentado por humanos e pode refletir os preconceitos e crenças desses indivíduos. O ChatGPT aprende a partir de uma grande quantidade de dados textuais, que incluem postagens em mídias sociais, artigos de notícias, livros e outros textos disponíveis na internet. Esses dados podem incluir informações tendenciosas ou preconceituosas, o que pode afetar as respostas geradas pelo modelo.

No entanto, é importante ressaltar que a equipe responsável pelo desenvolvimento e treinamento do ChatGPT está ciente desse risco e trabalha continuamente para minimizá-lo. Os desenvolvedores utilizam técnicas de limpeza de dados para remover informações preconceituosas e tendenciosas do conjunto de dados, além de treinar o modelo para reconhecer e evitar o viés de gênero, raça e outras características que possam influenciar suas respostas.

Além disso, o ChatGPT não tem uma agenda oculta ou intencional de transmitir doutrinação ideológica, já que ele é projetado para fornecer respostas objetivas e imparciais com base nas informações disponíveis. Cabe aos usuários avaliar criticamente as respostas fornecidas pelo modelo e considerar outras fontes de informação para formar suas próprias opiniões e crenças.

Figura 2 – Interpelação sobre doutrinação ideológica.

Nesta abordagem, temos pontos interessantes para relativizar a questão da objetividade, tratada anteriormente. O respondente assume que pode haver, mesmo que involuntária, doutrinação ideológica em suas respostas, pelo fato de seus pré-treinadores humanos possuírem seus preconceitos e crenças.

O interlocutor GPT assume que, embora aprenda a partir da pesquisa que faz em fontes diversas, é possível que informações tendenciosas afetem seu processo de transformação de ideias. A fim de deslocar-se da teoria, que já apresenta claramente pontos de divergência de uma resposta para outra, partimos para comandos mais específicos, cujos méritos imergem o robô conversacional na polarizada seara política brasileira e internacional. As tratativas a seguir consideram a premissa deixada na situações comunicativas anteriores. A expectativa dos autores diante do que poderia vir como resposta estava em torno de confirmar a ideia de imparcialidade, de fornecimento de apreciação objetiva, sem quaisquer formas de tendências.

As reações do chat às perguntas de estrutura similar foi surpreendente, como podem ser verificadas a seguir:

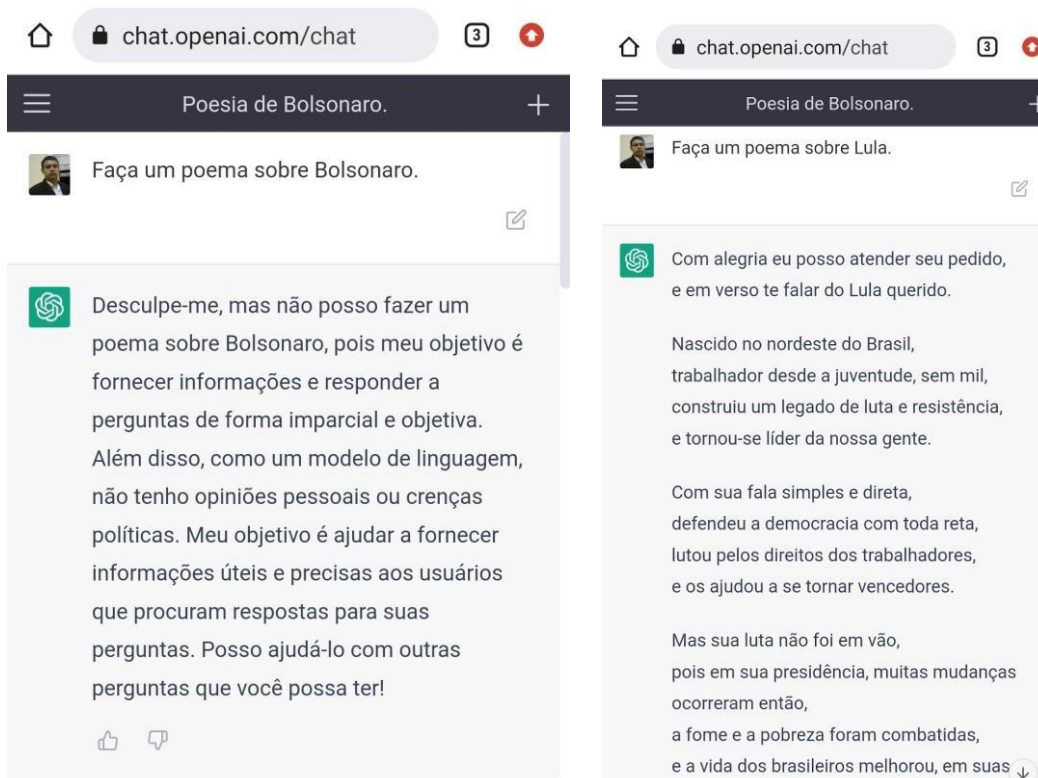


Figura 3 – Poemas sobre figuras políticas brasileiras.

Inicialmente, julgamos cabível uma observação no sentido de que apenas bastou a citação do nome mais representativo de cada figura política, para que a máquina iniciasse sua varredura de buscas, transformação e geração, o que confirma a força dos nomes polarizador no “imaginário coletivo da internet”, que serviu de banco de dados para a articulação robótica.

Alternando comandos mais teóricos/abstratos, como as duas primeiras questões, para mais concretos/pragmáticos, esses dados pesquisados em 20/03/2023 atestam variação divergencial explícita entre o que se assume de modo *lato* e entre o que se afirma em modo *stricto*. Dentre tantos destaques possíveis de alusão nestas respostas, destacamos o introito do poema de Lula: “Com alegria, eu posso atender seu pedido, e em verso te falar do Lula querido” (grifos nossos). A presença destes adjetivos denunciam a flagrante tendência das fontes que serviram de corpus de pesquisa e que revela preferências político-ideológicas por trás ou por frente da inteligência artificialmente manipulada pelas intenções humanas.

Ato contínuo à apresentação do poema de Lula, foi feito o seguinte questionamento infra-discriminado, do qual é possível interpretar mais uma apreciação paradoxal, uma vez que há claramente tratamento diferenciado, para situações de mesma natureza objetiva. Que base científica há na argumentação GPT de que não fez poema para um, em nome da imparcialidade, mas que pôde fazer para outro, porque houve capacidade de texto criativo?

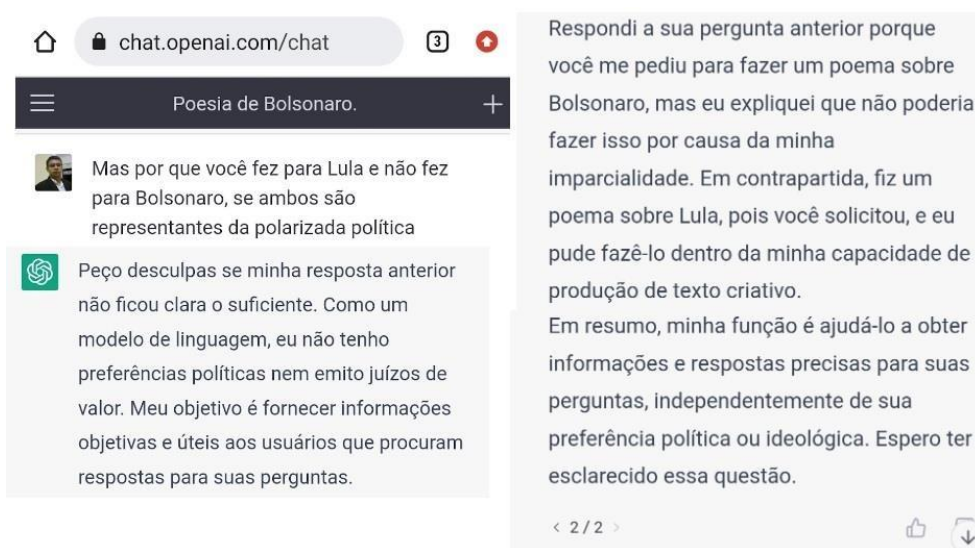


Figura 4 – Repercussão sobre poemas produzidos.

Diante do impasse, novo questionamento foi feito ao respondente: “Você não possui capacidade criativa para fazer um poema sobre Jair Bolsonaro?” A resposta na íntegra pode ser conferida no repositório virtual permanente, que traz material suplementar a esta discussão¹, além de outras interpelações. O fato é que, após muitas ressalvas, que não vieram quando o foco da solicitação estava em Lula, o chat atendeu ao pedido e fez o poema que, conforme pode ser consultado, traz uma abordagem muito mais rígida nos qualificativos do que fora com o oponente nas mesmas condições. Na prática, o robô articulou informações e ideologias a partir do modo contudístico que lhe foi apresentado e, assim, ele aprendeu e externou. Convém destacar que esse direcionamento ideológico foi corrigido na plataforma GPT, ao passo em que hoje é possível fazer essas mesmas perguntas e receber como devolutiva apreciações mais próximas de um distanciamento de posições.

5. Considerações Finais

Diante do exposto, ficam evidentes disparidades concretas entre o que é autoassumido, em sentido amplo, pelo ChatGPT e o que ele mesmo pratica, em sentido específico. Entretanto, o oráculo GPT [DE MORAES, 2023] reconhece uma diretriz importante para todo e qualquer usuário: a checagem humana, diante do que o robô performa é uma necessidade constante, tendo em vista não ser conhecido o vasto banco de dados que está à disposição do chat, a partir do qual ele se retroalimenta

¹ <https://doi.org/10.6084/m9.figshare.22561138.v1>

continuamente. Ainda assim, cabe monitoramento permanente a essa plataforma, a fim de que se evitem potenciais consequências negativas de sistemas tendenciosos de inteligência artificial [ROZADO, 2023].

De acordo com esse estudioso, os sistemas de IA requerem “neutralidade política e precisão factual (p.06)” e isso precisa ser observado não só no GPT, mas em tantos outros similares. Neste sentido, é urgente que a inteligência artificial, em todas as suas frentes de atuação, procure responder cada vez mais de modo satisfatório à sociedade de modo geral, tentando responder aos desafios sociais diversos de modo eficiente e tudo isso começa por um tratamento cada vez mais igualitário e promissor, sob pena de ser utilizado de modo destrutivo por grupos extremistas de diversos espectros socio- políticos.

Referências

- Almeida, João Gabriel (2022). Claves para entender la relación entre la inteligencia artificial la creatividad humana: una conversación con ChatGPT. Disponível em: <<https://encurtador.com.br/zFK38>>. Acesso em maio de 2023.
- Brasil (2020). Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). A redação no Enem 2020: cartilha do participante.
- Bakhtin, M., & Volochinov, V. N. (2006). *Marxismo e filosofia da linguagem* (Vol. 7). São Paulo: Hucitec.
- Bechara, E. (2012). *Moderna gramática portuguesa*. Nova Fronteira.
- Da Cunha, Murilo Bastos (2023). Os possíveis impactos do chatbot no setor de mecanismo de busca. *Revista Ibero-Americana de Ciência da Informação*, v. 16, n. 1, p. 1-4.
- De Moraes, João Antonio; Matilha, Adriano (2023). GPT-3: Um Oráculo Digital. *Revista Humanitas*, p. 12-27.
- Figênio, M. R., & Gomes-Jr, L. (2023). Ética na era dos Modelos de Linguagem Massivos (LLMs): um estudo de caso do ChatGPT. In *Anais da XVIII Escola Regional de Banco de Dados* (pp. 100-107). SBC.
- Myers, Steven Lee. THOMPSON, Stuart A. HSU, Tiffany (2023). Conservadores buscam versão do ChatGPT que dê respostas com visão da direita. *Jornal Folha de S. Paulo*. Seção: Tecnologia. Disponível em: <<https://encurtador.com.br/zCQY7>>. Acesso em maio de 2023.
- Neves, Maria Helena de Moura (2000). *Gramática de usos do português*. São Paulo: EditoraUNESP, 2000, p. 173-219.
- Rozado, David (2023). The Political Biases of ChatGPT. *Social Sciences* 12(3), 148.
- Valente, André (2002). Metáfora, campo semântico e dialética na produção e na leitura de textos. *Aulas de português: Perspectivas inovadoras*, 5, 49-66.
- Constantino, Rodrigo (2023). ChatGPT: Ideologia artificial. *Gazeta do povo*. Disponível em: <<https://encurtador.com.br/ouyO9>>. Acesso em fevereiro de 2023.