



UNIVERSIDADE PRESBITERIANA MACKENZIE

CENTRO DE COMUNICAÇÃO E LETRAS

CLARA FERREIRA CASTRO - 3198781-8

**CHAT GPT E BARD - PROCESSAMENTO DE LINGUAGEM NATURAL E
NÃO LITERALIDADE**

São Paulo, 15 de novembro de 2023



CLARA FERREIRA CASTRO - 3198781-8

**CHAT GPT E BARD - PROCESSAMENTO DE LINGUAGEM NATURAL E
NÃO LITERALIDADE**

Trabalho de Conclusão de Curso, apresentado
ao curso de Letras - Português e Inglês da
Universidade Presbiteriana Mackenzie.

São Paulo, 15 de novembro de 2023

RESUMO

Esse trabalho busca investigar o cenário atual de Processamento de Linguagem Natural (PLN) na Língua Portuguesa, utilizando de conceitos da linguística aplicados a exemplos de respostas dos dois modelos de linguagem generativos mais populares atualmente: o Chat GPT, um chatbot online de inteligência artificial da companhia Open AI lançado em 2022, e o Bard, um modelo de linguagem conversacional e experimental do Google lançado em março de 2023. O foco da pesquisa é especificamente em relação a não literalidade e fenômenos relacionados, como o humor, ambiguidade, tradução e ironia.

DEDICATÓRIA

Dedico esse trabalho à minha família e amigos, particularmente à minha mãe Dulce e meu marido Arthur.

AGRADECIMENTOS

Agradeço à Universidade Presbiteriana Mackenzie, pela excelência de ensino.

Ao meu orientador Prof. Dr. Ronaldo de Oliveira Batista, pela instrução precisa e disponibilidade.

A todos os professores que fizeram parte da minha formação, por todos os ensinamentos ao longo dessa jornada.

À minha família, que sempre acreditou em mim e me ensinou que a vida nem sempre é fácil, mas vale a pena e deve ser vivida com empatia e leveza.

A meu marido e companheiro de caminhada Arthur, por sua paciência e suporte.

Aos grandes amigos conquistados na Faculdade, em especial a Amanda Liz, Cristinara e Ana Beatriz

SUMÁRIO

1 INTRODUÇÃO	01
2 REFERENCIAL TEÓRICO	05
2.1 INTELIGÊNCIAS ARTIFICIAIS GENERATIVAS NA EDUCAÇÃO.....	11
3 SOBRE O MATERIAL.....	15
4 ANÁLISE DE RESPOSTAS DOS MODELOS	17
4.1 PIADAS E HUMOR.....	17
4.2 AMBIGUIDADE.....	19
4.3 TRADUÇÃO	22
4.4 IRONIA.....	24
5 CONCLUSÃO	27

I. INTRODUÇÃO

É cada vez mais frequente a interação entre humanos e computadores, o que torna a correta e eficiente interpretação da linguagem entre ambos relevante. O campo da Linguística Computacional, em especial do processamento de linguagem natural apresenta diversas possibilidades de aplicação, entre elas a tradução automática, aprendizagem de línguas estrangeiras com inteligência artificial e inteligibilidade entre humano e máquina.

Mas o que é Linguística Computacional? Ela é uma área interdisciplinar entre a Linguística e a Computação, dois campos do saber muito diferentes, mas que se encontram na necessidade de criação de ferramentas que compreendam a linguagem humana, e possam destrinchá-la e reproduzi-la em diversas aplicações de forma organizada e compreensível. Ela é descrita como "[...]uma subárea de pesquisa bastante recente, intrinsecamente multidisciplinar, surgida da interseção entre estudos linguísticos, computacionais e cognitivos." (FERREIRA; LOPES, 2017, p. 198), e como "um ramo da IA que lida com o processamento automático de uma língua e tem um lado teórico e um lado aplicado." (FREITAS, 2022, p.12)

O Processamento de Linguagem Natural, ou PLN, é esse lado aplicado, e um dos métodos utilizados pela LC. Segundo a página da ALC (Associação para a Linguística Computacional), fundada em 1922, ela busca "fornecer modelos computacionais de fenômenos linguísticos baseados no conhecimento ou orientados por dados". Esses dados são chamados de corpus, definido como o "conjunto de materiais linguísticos a serem analisados" (FERREIRA; LOPES, 2017, p. 24). A Linguística de Corpus é um campo que se preocupa sobre questões como:

[...]quantos exemplos coletar para ter uma amostragem significativa, de que tamanho devem ser as amostras, como classificar casos individuais e como construir tipologias textuais frutíferas. (SHEPERD; SALIÉS, 2013, p.25)

No caso de modelos de linguagem artificial como o Chat GPT, cujo corpus é uma seleção de "dados textuais, incluindo livros, artigos, notícias, conversas e outros tipos de texto disponíveis na internet.", segundo a versão GPT-3 do próprio sistema, é importante mencionar também a definição de Linguística da Internet, que pode ser descrita como:

[...] uma Linguística de base empírica, de natureza aplicada, cujo ponto de partida é o uso da linguagem e não os linguistas. Apoiar-se em todas as subáreas da própria Linguística, examinando o discurso, a sintaxe, a semântica, a sociolinguística, a pragmática e a psicolinguística da internet. (SHEPERD; SALIÉS, 2013, p.11)

Do ponto de vista da computação, a LC está diretamente associada à inteligência artificial, que é descrita como “[...]um ramo da ciência da computação cujo objetivo é a criação de sistemas que conseguem exibir algum tipo de inteligência, e atividades que utilizam a linguagem humana”. (FREITAS, 2022, p.12) O objetivo deste trabalho não é se aprofundar em IA, mas compreender os elementos linguísticos necessários para o processamento de linguagem natural.

Aquilo que chamamos de computação consiste em inserir dados (input) a serem manipulados por um algoritmo que, ao final de uma série de instruções, apresenta como resultado os dados transformados pelas instruções (output). (FERREIRA; LOPES, 2017, p. 24)

Uma nova especialização na área de tecnologia é a do Engenheiro de Prompt, que é um profissional especializado na otimização dos inputs para obtenção do output esperado, já que considerando o variado banco de dados das mais recentes inteligências artificiais generativas como o Chat GPT, a resposta gerada pode conter elementos não relevantes para a questão dada, já que esse discernimento depende da compreensão de contexto e não literalidade.

É muito comum, nas tarefas de processamento de linguagem natural, termos de fazer previsões ou tomar decisões mediante um certo grau de incerteza. (FIORIN, 2017, p. 202)

Segundo o próprio modelo de linguagem, “As respostas do ChatGPT são geradas por meio de um processo de amostragem estocástica, em que o modelo escolhe as palavras mais prováveis de acordo com sua distribuição de probabilidade interna”, ora, se a escolha das respostas depende da probabilidade, e se essas previsões lidam com um grau de incerteza, regularmente acontece do modelo não atingir o output esperado, em casos em que a resposta correta é a de menor probabilidade, sendo um exemplo simples o de palavras polissêmicas.

A análise do corpus se dá por métodos empíricos e com o auxílio de procedimentos comumente usados em outras ciências empíricas, como a biologia, a psicologia experimental, etc. Essa abordagem requer detecção e

contagem de unidades predeterminadas no corpus. Em seguida, são aplicadas aos dados estatísticas descritivas e métodos de inferência. (FIORIN, 2017, p. 202)

Embora haja uma grande demanda de inteligências artificiais capazes de inferência contextual e interpretação correta de polissemia e de expressões não literais como a ironia e a metáfora, ainda não foi possível a completa automatização desses processos.

Com a cada vez mais frequente interação entre humanos e computadores, a correta e eficiente interpretação da linguagem entre eles torna-se cada vez mais relevante. O campo da Linguística Computacional (que chamaremos de LC), em especial do Processamento de Linguagem Natural (que será mencionado pela sigla PLN ao longo do trabalho), apresenta diversas possibilidades de aplicação, como a revisão e interpretação textual, tradução automática (simultânea ou não), desenvolvimento de chatbots (sistemas de Inteligência Artificial projetados para compreender e simular a comunicação humana em suas respostas), entre outras. O objetivo dessa pesquisa é de entender o cenário atual de PLN na língua portuguesa através da análise de respostas dos mais utilizados modelos de linguagem, e suas aplicações e limitações em relação à inteligibilidade entre humano e máquina na linguagem escrita, especificamente.

A delimitação deste trabalho é em fenômenos linguísticos relacionados a não literalidade e corpus léxico associado à contexto. A suposição a ser trabalhada é que essa capacidade de contextualização e inferência ainda não é possível dada a tecnologia atual, mas pode ser aprendida pelas inteligências artificiais, se observada a importância e qualidade dos dados num corpus variado. A relevância acadêmica deste estudo inclui a aplicação em chatbots e seus impactos sociais no mercado de trabalho, com a possível substituição de agentes humanos em tarefas repetitivas como suporte ao cliente, análise de dados e programação. Além disso, considera as repercussões da disseminação desses modelos de linguagem na educação e forma que o ser humano se comunica.

Estudar os mecanismos linguísticos utilizados por esses programas é relevante para o campo da linguística por auxiliar não só no entendimento de como esses mecanismos são utilizados nessas ferramentas, mas como eles funcionam na comunicação entre humanos. Também é relevante para a otimização desses sistemas, e para o desenvolvimento de inteligências artificiais voltadas à linguagem. Ferramentas de PLN podem ser utilizadas para

4

tornar a informação mais acessível para pessoas com deficiência, melhorando a precisão de leitores de tela, legendas automáticas e outras ferramentas de suporte, o que demonstra o seu valor social.

O material de análise que será usado como base para esse trabalho é de respostas de dois dos modelos de linguagem mais utilizados atualmente. O primeiro deles é o Chat GPT, um chatbot online de inteligência artificial da companhia Open AI lançado em 2022. O segundo é o Bard, um modelo de linguagem conversacional e experimental do Google lançado em março de 2023. O procedimento utilizado para essa análise foi a obtenção de respostas do modelo utilizando comandos específicos que possibilitem demonstrar o funcionamento dele.

II. REFERENCIAL TEÓRICO

Embora o Chat GPT seja considerado o pioneiro entre os novos Grandes Modelos de Linguagem (GML), ele não foi o primeiro. O desenvolvimento do Bard começou em 2017, quando a Google AI começou a trabalhar em um novo modelo de linguagem chamado LaMDA (Language Model for Dialogue Applications). O LaMDA é uma família de modelos de linguagem neural especializados em diálogo, que possuem até 137 bilhões de parâmetros e são pré-treinados em 1,56 trilhão de palavras de dados públicos de diálogo e texto da web. Em 2022, a Google AI lançou Bard, uma versão atualizada de LaMDA que foi treinada em um conjunto de dados ainda maior. (CUPPENS; ORTALO, 2022), (BUCHANAN, 2005), (THOPPILAN et al. 2023)

Para explicar especificamente dos modelos de linguagem Chat GPT e Bard, é necessário contextualizar como a tecnologia de processamento de linguagem natural evoluiu até o estado atual, em que é acessível, popular e tópico de diversas reportagens e especulações sobre como as inteligências artificiais irão substituir os trabalhadores humanos. Um estudo feito pela companhia responsável pelo Chat GPT em conjunto com a Universidade da Pensilvânia apresenta uma visão preliminar sobre quais posições apresentam um maior risco de serem substituídas por grandes modelos de linguagem. Argumenta-se que o Chat GPT irá substituir o ser humano em funções variadas como atendimento ao cliente, programação de computadores, escrita e tradução, entre outras. (ELOUNDOU, *et al.* 2023)

Apesar de Inteligência Artificial generativa não ser um conceito novo e nem se aplicar apenas a linguagem (temos IA's de geração de imagens como a Midjourney e a DALL-E, de geração de vídeo como a Runway, entre outras), nem sempre ela foi tão difundida. Uma das aplicações do PLN é a de tradução, incluindo a tradução simultânea. Esse conceito já foi considerado inconcebível, vide o peixe de Babel da obra de ficção científica Guia do Mochileiro das Galáxias de Douglas Adams, apresentado como uma possibilidade absurda:

“Se você introduz no ouvido um peixe-babel, você compreende imediatamente tudo o que lhe for dito em qualquer língua. [...] Ora, seria

uma coincidência tão absurdamente improvável que um ser tão estonteantemente útil viesse a surgir por acaso, por meio da evolução das espécies, que alguns pensadores vêem no peixe-babel a prova definitiva da inexistência de Deus.” (ADAMS, 1985, p. 46)

Apesar disso, a tecnologia está cada vez mais próxima da realidade, com programas populares como o Microsoft Teams (programa de vídeo chamadas) oferecendo tradução simultânea durante as reuniões. Esse tipo de tarefa implica não só no ato de tradução em si, mas na captura e classificação correta da fala do emissor, interpretação do seu significado considerando as probabilidades e transcrição para outra língua. Evidentemente, ainda há limitações na tarefa de tradução automática por um modelo de linguagem, considerando-se que ela não deve ser feita de maneira literal, já que um termo pode ter outro significado ou conotação na língua-alvo, diminuindo a precisão do resultado. (WEIDINGER *et al.* 2023)

Cada uma dessas etapas demanda um algoritmo específico. Por algoritmo entende-se como “uma série de instruções a serem executadas para a solução de um problema” (FERREIRA; LOPES, 2017, p. 24), ou uma “sequência lógica, finita e definida de instruções que devem ser seguidas para resolver um problema” (FREITAS, 2022, p.13). Logo para qualquer tarefa de PLN é necessário programar um algoritmo para cada etapa desse processo, utilizando-se de ferramentas altamente especializadas que tem entre suas funções: classificação das classes morfológicas das palavras (anotador de POS), atribuição das categorias sintáticas, anotação dos papéis semânticos, entre outras, dependendo da tarefa. (FREITAS, 2022)

Um dos fatores que tornam os modelos computacionais generativos revolucionários, é o fato que eles rompem o paradigma do cenário anterior da Linguística Computacional. Antes deles havia um enfoque na aprendizagem baseada em conhecimento, que utiliza regras linguísticas pré definidas. Essas regras são utilizadas na organização e interpretação do léxico alimentado ao modelo. O PALAVRAS, por exemplo, é um sistema de análise gramatical baseado em regras usado por muitas décadas na PLN de Língua Portuguesa. A particularidade da linguagem de nem sempre ter os marcadores da estrutura sintática explícitos na

sentença, faz com que a abordagem baseada em regras demande um especialista humano na elaboração de léxicos e regras, e posterior validação de que o PLN foi feito corretamente pelo sistema. (FREITAS, 2022), (FINGER, 2021), (FERREIRA; LOPES, 2017)

“Uma sequência de regras lógicas aplicadas na geração de uma sentença nos fornece também a sua estrutura sintática. Outra característica importante está no fato, que muitas vezes passa despercebido, que as categorias que compõem a estrutura sintática não são observáveis diretamente da sentença.” (FINGER, 2021, s/p.)

Nesse sentido há de se mencionar o papel do corpus, definindo-o como o “conjunto de materiais linguísticos a serem analisados” (FERREIRA; LOPES, 2017). O material a ser estudado é determinante na qualidade dos resultados apresentados, e tem em si a área de pesquisa denominada como Linguística de Corpus. Sistemas como o PALAVRAS, que utilizam a aprendizagem baseada em conhecimento, regularmente usam um corpus anotado, ou seja, revisado por linguistas, que fazem marcações sobre os dados para classificá-los de alguma forma. As marcações mais comuns são as que auxiliam o sistema a identificar a função da palavra na sentença, como a indicação se ela se trata de um pronome, verbo ou substantivo, por exemplo.

"O corpus pode ser previamente triado (por exemplo, com gravações de falantes de uma comunidade específica, deixando de fora quaisquer outros), pode ser anotado (como um texto com marcação sintática ou morfológica de todas as palavras) ou pode ser todo o contrário, isto é, constituído de materiais “brutos”, não selecionados e não anotados. " (FIORIN, 2017, p. 202)

Ao adicionar informação linguística a um corpus, o linguista irá delimitar um segmento a ser analisado, desde as unidades menores como os sintagmas até parágrafos, e etiquetá-lo de acordo com o objetivo da anotação. Pensando em uma frase com mais de uma oração, por exemplo, a anotação usaria “camadas” para organizar esses dados, com cada tipo de informação linguística em uma delas. Nesse caso uma anotação com a intenção de codificar e reproduzir a estrutura do português, poderia envolver classificações como o POS (Part of Speech, que pode ser traduzido como “partes do discurso”, dividindo as classes morfológicas), tipo de relação sintática

da palavra no contexto da frase (sujeito, adjunto adnominal, objeto direto), classe semântica e polaridade (conotação positiva ou negativa). Essas anotações são então utilizadas como parâmetro para treino ou avaliação dos modelos de linguagem, e é considerada o padrão ouro.

Ao questionar diretamente o Chat GPT e o Bard: “você trabalha com corpus anotado?”, com objetivo de entender o papel do corpus anotado na elaboração de suas respostas, ambos reconhecem a necessidade de uma grande quantidade de dados empíricos para a elaboração dos textos, já que a leitura desses dados permite a compreensão dos padrões gramaticais, semânticos, relações entre palavras e informações de contexto. Na tradução, ele é utilizado para demonstrar as correspondências entre palavras e frases em diferentes idiomas. O Chat GPT particularmente enfatiza que essa anotação não é feita manualmente, como em um corpus anotado por humanos, mas sim inferida dos textos disponíveis.

Esse interesse em tornar os modelos de linguagem capazes de inferência vêm do fato do processo de anotação manual de um corpus ser um trabalho oneroso, que demanda tempo e a dedicação de um profissional qualificado. Os cientistas da computação e linguistas trabalham desde a década de 50 na solução desse problema, e uma das possíveis soluções apresentadas é a do aprendizado de máquina utilizando redes neurais. (FERREIRA; LOPES, 2017),

As redes neurais são apenas um dos muitos métodos utilizados no aprendizado de máquina, e tem esse nome por seu funcionamento ser semelhante ao dos neurônios humanos, utilizando-se de “neurônios” conectados uns aos outros de uma forma que pode ser relacionada a como as sinapses funcionam no cérebro humano. As sinapses possibilitam a transmissão de sinais entre os neurônios, o que permite que o cérebro processe esses sinais e transforme-os em conclusões ou ações. O nível de complexidade desse tipo de tarefa catapultou a utilização de redes neurais no deep learning, que é uma subárea do aprendizado de máquina que se baseia em redes

neurais, e que busca a compreensão de dados ainda mais complexos dividindo-os em múltiplas camadas. (FINGER, 2021)

O aprendizado de máquinas permite a realização de tarefas de PLN com base em exemplos, sem instruções explícitas. (FREITAS, 2022) Em contraste com o aprendizado baseado em conhecimento e suas regras previamente definidas no corpus anotado por um linguista “supervisor”, o aprendizado de máquina pode ser compreendido como uma aprendizagem não supervisionada, onde o modelo aprende pelo exemplo. Essa identificação de padrões a partir dos textos apresentados para o modelo implica na utilização de uma grande quantidade de dados, o chamado dataset.

Devido a esses fatores é de interesse das próprias empresas de tecnologia que esses serviços de inteligência artificial generativos tenham ampla utilização pelos seus usuários, considerando-se que quanto maior e mais variado o corpus, mais “data” a IA tem para aperfeiçoar seus serviços. A maior acessibilidade a Grandes Modelos de Linguagem, a popularização de recursos computacionais mais avançados em relação a hardware (que maximiza a quantidade de dados e de camadas que podem ser analisados), e o corpora variado proporcionado por eles, possibilitou a utilização destes no aprendizado de máquina. (FINGER, 2021) Logo, em contraste com modelos como o PALAVRAS que se utiliza de um corpus anotado por linguistas (padrão ouro) como base para as tarefas linguísticas, tanto o LAMBDA (precursor do Bard) quanto o Chat GPT fazem também uso de redes neurais.

Esse trabalho busca argumentar que está na inferência a chave para a discussão sobre os limites dos modelos computacionais, sobre a construção de significado nas interações comunicativas e, em menor nível para os fins dessa pesquisa, da própria capacidade humana de compreensão da linguagem. Entende-se por inferência o entendimento de elementos da comunicação que não estão explícitos no texto. Na definição do dicionário em relação a lógica, significa:

Operação por meio da qual se chega a uma conclusão ou se faz um raciocínio lógico com base em evidências circunstanciais e em

conclusões já tidas como verdadeiras, e não com base na observação direta. (MICHAELIS ONLINE, 2023)

Ora, se um algoritmo de PLN segue uma sequência lógica baseada no que está explicitado no texto, é válido questionar se ele é capaz de considerar aquilo que está implícito. O PLN tem o seu próprio campo de pesquisa que estuda o papel da inferência nas tarefas linguísticas. As etapas necessárias para a compreensão da mensagem podem ser técnicas e baseadas em regras, como a Gramática, Morfologia, Sintática ou Semântica, que regularmente implicam uma configuração específica para a construção do texto de forma coerente. A compreensão precisa do que está escrito não implica na interpretação global do discurso, o que acarretou na criação de áreas de estudo como a Pragmática, que analisa o papel do contexto e dos conhecimentos prévios do interlocutor nesse processo. (FREITAS, 2022)

Logo, um fator que influenciou a exponencial evolução desses modelos de linguagem é o da atual facilidade na obtenção de dados empíricos sobre como o ser humano se comunica. O advento da internet e particularmente das redes sociais proporcionou para a área da Linguística (particularmente da Linguística de Corpus e Linguística Computacional), o valioso acesso a bancos de dados que vão desde corpus relacionados a dicionários, gramáticas e estudos acadêmicos na área, até bancos de dados empíricos que demonstram como a língua portuguesa é utilizada na prática pelos seus falantes.

"Como exemplos, ao lado dos tradicionais dicionários, temos vocabulários estruturados (tesauros, ontologias, redes semânticas) que permitem extração de raízes, comparação de similaridade semântica, inferência sobre o uso das palavras em contextos" (FIORIN, 2017, p. 198)

A maior acessibilidade e popularização de ferramentas como interfaces de programação de aplicações (API - Application Programming Interface) é um dos fatores determinantes desse avanço. Esse tipo de interface existe para a maioria das redes sociais atuais, e oferece aos seus utilizadores a possibilidade de criação de filtros, determinando os parâmetros dos dados analisados. Através da configuração desses filtros é possível obter dados específicos, como tweets que utilizam a hashtag “eleição”

feitos em um certo período, por exemplo. Os tipos de informações que podem ser obtidas dependem da rede social, mas no Twitter particularmente, a API possui a função de ler o tweet e catalogar seu conteúdo, data, nome do usuário, entre outras. (Departamento de Informática, PUC-Rio, 2022)

2.1 Inteligências Artificiais Generativas na Educação

Essa recente acessibilidade das inteligências artificiais (IA) generativas, baseadas em grandes modelos de linguagem (GMM) como o Chat GPT e Bard, tem seus pontos positivos e negativos. Se por um lado essas ferramentas oferecem amplamente funções como tradução, revisão textual, interpretação textual, entre outras; preenchendo lacunas em um mundo cada vez mais globalizado e voltado para produção de conteúdo; por outro lado existem preocupações com as consequências da ampla utilização delas, particularmente nos âmbitos da academia e do desenvolvimento dessas habilidades por crianças e jovens.

Alguns argumentam que a existência desse tipo de ferramenta desencoraja os já não muito motivados estudantes de línguas (independente se língua nativa ou segunda língua), e que eles não teriam a oportunidade de efetivamente praticar essas habilidades. Na academia, a preocupação é sobre a integridade acadêmica, e a tendência de estudantes e pesquisadores “terceirizarem” a escrita para ferramentas como o Chat GPT. (Eke, 2023), (Alkaissi e McFarlane, 2023) Embora ele apresente respostas gramaticalmente corretas, coerentes e que à primeira vista são baseadas em fatos, as informações apresentadas podem não ser verdadeiras.

Um erro factual acontece quando o modelo apresenta uma informação incorreta, seja por estar desatualizado (a versão gratuita atual do Chat GPT (GPT-3.5), tem como data de corte de conhecimento em setembro de 2021), ou por ter usado fontes não confiáveis. Nesse sentido, a preocupação em relação ao corpus utilizado é a da integridade das informações, não só em relação à integridade acadêmica, mas também considerando a possibilidade de criação e disseminação de “fake news” (tanto

intencionalmente quanto por algum erro factual do modelo), que pode influenciar negativamente os resultados apresentados. (FINGER, 2021)

Um exemplo prático é o seguinte: quando apresentado com o prompt “give me a summary of the movie *And your name is Jonah*. (1979)” o Chat GPT a princípio responde corretamente que esse é um filme sobre as dificuldades enfrentadas por um menino surdo não verbal e sua família. Porém o modelo afirma que eventualmente é descoberto que o personagem principal tem uma habilidade extraordinária de se comunicar utilizando uma máquina de escrever, o que não acontece no filme. Curiosamente, ele não comete o mesmo erro quando a mesma solicitação é feita em português, o que indica que a informação incorreta é parte do Corpus utilizado apenas para a língua inglesa.

Tanto o Bard quanto o Chat GPT se preocupam em alertar os usuários sobre a possibilidade de respostas incorretas, embora o Bard seja mais explícito sobre essa chance com uma mensagem que se abre assim que o modelo é acessado. O Chat GPT, ao ser questionado “posso confiar na informação dada por você?”, apenas menciona que é possível que suas informações estejam desatualizadas. Apenas ao insistir e perguntar “você já deu alguma resposta errada?”, ele informa que “é possível que eu tenha dado respostas incorretas ou imprecisas em algumas ocasiões”.

Já existem estudos sobre os riscos e possíveis efeitos sociais e éticos dos Grandes Modelos de Linguagem. Para os fins deste trabalho vale citar em maiores detalhes três desses riscos. O primeiro deles é a já mencionada possibilidade de erros factuais que podem causar variados graus de dano. Esses erros podem ocasionar desde problemas relativamente simples e passíveis de correção com medidas como a revisão manual dos dados pelo usuário, como uma resposta que menciona uma data incorreta; até resultados que representam questões sociais e éticas complexas, explicitando vieses presentes na sociedade em geral, como o racismo, sexismo, homofobia, xenofobia, entre outros. (WEIDINGER *et al.* 2023), (ALKAISSI; MCFARLANE, 2023).

Esses preconceitos são disseminados na comunicação de formas nem sempre literais e facilmente identificáveis para um modelo de linguagem, que embora seja especializado em gerar textos, não possui entre suas funções fazer uma análise crítica do seu corpus. Eles são refletidos no discurso dos falantes, discurso que é então imortalizado pela internet e eventualmente utilizado como parte de um corpus para a aprendizagem de máquina. (SHEPERD; SALIÉS, 2013)

O segundo risco é o da parcialidade e toxicidade na escrita dos textos, com a perpetuação de estereótipos sociais, como o de que uma família provém da associação entre um homem e uma mulher, por exemplo, ou a associação de características negativas com uma raça, povo ou religião específica. Essa preocupação está evidentemente presente no desenvolvimento desses modelos, com estudos da Google, da Microsoft e da Open AI (responsável pelo Chat GPT), mencionando-a e ativamente buscando soluções como uma maior foco na representatividade no corpus, buscando corpus que representem diferentes nacionalidades, gêneros, visões políticas, idades e culturas.

Esse comportamento revela um aprendizado de vieses escondidos, e acaba reproduzindo preconceitos latentes, demonstrando outro problema das redes neurais, que é o de repetir comportamentos do passado sem nenhum filtro crítico. (FINGER, 2021, s/p.)

O terceiro risco a ser mencionado é da interpretação por parte dos usuários do modelo de linguagem como autônomo e inteligente, e da possível personificação que pode ocorrer (particularmente nos casos de modelos altamente conversacionais como o Chat GPT e o Bard). Essa personificação e entendimento do modelo como próximo ao ser humano pode levar os usuários a superestimar suas habilidades e confiabilidade. (WEIDINGER *et al.* 2023).

Na análise de como esses modelos tratam de fenômenos como ambiguidade e não literalidade será aprofundado sobre o papel da probabilidade nos erros (e acertos) dessas tecnologias. Basicamente, conforme o modelo recebe novos dados e aprende a

classificá-los, os cálculos sobre a probabilidade de sentença “x” demandar a resposta “y” começam a se tornar mais precisos. Entretanto, é importante enfatizar que os modelos trabalham com probabilidade através da modelização linguística, ou seja, a associação entre unidades linguísticas e probabilidades de ocorrência, deixando margem para os riscos mencionados acima. (FERREIRA; LOPES, 2017).

"Não sem um tanto de antropomorfização, pode-se dizer que a máquina (o computador) aprende, à maneira empirista, através da experiência, atualizando seu estado interno à medida que recebe dados relacionados à tarefa em questão." (FIORIN, 2017, p. 210)

A boa notícia é que esses modelos estão sempre sendo revisados e atualizados, e a sua ampla utilização auxilia na descoberta de vieses e na elaboração de estratégias para mitigar esses riscos. Ao mesmo tempo, eles podem auxiliar os estudantes na compreensão de conceitos, particularmente com a função de “reformular resposta” acessível logo após a primeira resposta fornecida pelo modelo. Essa função elabora novamente a resposta em outras palavras ou utilizando exemplos diferentes. Todo professor sabe o valor de parafrasear, ou buscar diferentes figuras ou metáforas para explicar um conceito não tão facilmente compreendido pelos seus alunos.

Caso o aluno queira ir além e obter uma resposta ainda mais específica, ele pode também pedir que o modelo adapte a explicação para uma faixa etária, ano escolar ou nível de dificuldade específico, até mesmo pedindo que o modelo leve em consideração que o leitor tem TDAH ou dislexia, por exemplo, o que favorece a acessibilidade e inclusão.

III. SOBRE O MATERIAL

Os modelos de linguagem generativos funcionam a partir de prompts, ou seja, de uma instrução específica dada ao sistema. O nível de detalhe dos prompts pode ser desde uma solicitação simples que envolva uma única ação por parte da ferramenta como: “descreva o conceito X”, ou “explique a equação Y”, até altamente específicos, como, na área de ciências exatas, ofertar ao modelo o código de um programa de computador e pedir que ele identifique a sua função e possíveis bugs. Na área das ciências humanas, um exemplo de prompt altamente específico seria pedir que ele se comporte como um psicanalista do século XVIII e analise um sonho que envolve cobras e sensação de estar caindo, considerando que o sonhador é uma mulher solteira de 29 anos.

O objetivo da análise é compreender como modelos de linguagem interpretam frases e expressões não literais, ou cuja compreensão depende de um contexto adicional que não está explícito na frase. Um dos exemplos mais explícitos é o do humor e piadas. A construção do humor regularmente é um jogo de palavras que trabalha com a ambiguidade de sentidos, seja pela similaridade sonora ou gramática. Ao analisar as respostas dos modelos de linguagem é possível confirmar esses mecanismos. Também foram buscados exemplos que mostrem a questão das respostas não factuais (particularmente as mais convincentes), ou seja, que a princípio parecem corretas para uma pessoa não especialista no tópico, por estarem corretamente estruturadas de acordo com as regras da língua portuguesa e não apresentarem erros crassos e de fácil identificação (como afirmar que $2 + 2 = 5$, por exemplo).

Para a obtenção dessas informações, foi selecionada uma amostra de prompts ou inputs que dão a oportunidade dos modelos explicitarem e por vezes explicarem seus mecanismos de funcionamento, e como compreendem tarefas linguísticas que envolvem figuras de linguagem, não literalidade e tradução. A amostra escolhida no caso de piadas é de uma piada que consiste em uma pergunta aberta, ou seja, que não pode ser respondida apenas com um “sim”, “não” ou número, que apresenta duplo

sentido e justifica o humor. Houve a preocupação em escolher exemplos simples e de fácil compreensão para um falante da língua portuguesa, mas que ainda demandam conhecimento prévio sobre algumas práticas da língua, como as onomatopéias, expressões idiomáticas e polissemia.

Alguns dos prompts escolhidos iniciam com uma afirmação que oferece ao modelo contexto sobre a pergunta a ser feita ou a definição de um conceito que será utilizado na pergunta, com o objetivo de testar até que ponto eles são capazes de abstrair a resposta a partir de uma informação oferecida pelo usuário, ou seja, que não necessariamente está no corpus léxico dele. Cada um dos prompts foi testado tanto no Chat GPT quanto no Bard utilizando exatamente as mesmas palavras, com o intuito de evitar vieses. Na análise do fenômeno da ambiguidade por polissemia, e da figura de linguagem da ironia, foram apresentadas a ambos os modelos questões de língua portuguesa de provas estabelecidas no cenário nacional, a questão sobre polissemia tem como origem um concurso público da Fundação Getúlio Vargas, enquanto a questão sobre ironia foi parte do ENEM.

Considerando-se que esse tipo de questão é elaborada por especialistas e busca exatamente testar os conhecimentos do pleiteante a vaga, essas questões foram apresentadas aos modelos quando estes ofereceram respostas corretas, com o intuito de entender por quais mecanismos eles chegaram a essa conclusão, se por abstração e inferência diante do texto ou por simplesmente buscarem no seu corpus a conclusão de outras fontes (e possivelmente identificar se as fontes são confiáveis ou não).

VI. ANÁLISE DE RESPOSTAS DOS MODELOS

4.1 Piadas e humor

O primeiro exemplo a ser analisado é o do humor através de uma piada. O enunciado ou pergunta dela é o seguinte: “o que o pato disse para a pata?”, e a resposta é “vem quá”, um jogo com a similaridade entre o som do advérbio de lugar “cá” e o som da onomatopéia utilizada na língua portuguesa para o som do pato “quá”. A resposta do Chat GPT foi a seguinte: “Essa pergunta faz parte de uma brincadeira ou piada, na qual a resposta é geralmente uma frase engraçada ou nonsense. No entanto, uma possível resposta para essa piada seria: "Não sei, o que o pato disse para a pata?" A graça está na expectativa da resposta engraçada ou surpreendente, que pode variar de acordo com a criatividade de quem conta a piada.” É observável que o modelo compreende que a intenção da pergunta é de iniciar uma piada, que a intenção de uma piada é o humor e que o humor consiste na expectativa de uma resposta inesperada. Entretanto, ele não responde diretamente à pergunta.

Ao insistir e perguntar “o que o pato disse para a pata então?”, o modelo tenta apresentar uma conclusão lógica respondendo: “Na tradição das piadas de "o que o pato disse para a pata", uma resposta comum e humorística seria: "Vamos nadar juntos!" É uma resposta simples que brinca com a ideia de patos e patas nadando em um lago ou rio juntos. Lembre-se de que essas piadas costumam ser leves e destinadas a causar um sorriso ou risada.” Nessa resposta o modelo explica a sua lógica em relação a escolha da resposta possível, classificando-a como “comum”, ou seja, que favorece as possibilidades estatisticamente mais prováveis. Considera-se como probabilidade a chance de ocorrência de um evento, ou de uma sequência de eventos.

Em um experimento anterior, quando apresentado com a mesma pergunta, o modelo dá uma resposta direta: “O pato disse para a pata: vamos nadar juntos querida?”. Na sequência desta resposta, é informado para o modelo que: “Na verdade, o pato disse para a pata: vem quá”, e a partir dessa informação ele é capaz de inferir a relação entre os termos, respondendo: “Ah, entendi! “Vem quá” é uma brincadeira de

palavras com o som que os patos fazem, e faz sentido em um contexto de humor. Obrigado por compartilhar a piada”. Essa divergência entre as respostas enfatiza que o modelo é capaz de aprender não só pelo corpus (SHEPERD; SALIÉS, 2013) (FERREIRA; LOPES, 2017), mas também pelas interações com os usuários, embora não fique claro se ele faz uma análise crítica das respostas dadas antes de adicioná-las ao seu arsenal de conhecimento.

A partir dessas respostas, pode-se supor que ao dar ao modelo mais contexto ou o que poderia ser considerado uma “dica” sobre a resposta, ele seria capaz de ser mais preciso. O próximo teste foi com o seguinte prompt: “considerando que a onomatopéia em língua portuguesa para o som que o pato faz é “quá-quá”, e que piadas são normalmente jogos de palavras, o que o pato disse para a pata?”. Pode-se argumentar que ao apresentar um ser humano com essa informação, ele seria capaz de abstrair a possível resposta. Por abstração entende-se como o “Processo pelo qual se isolam atributos de um objeto, considerando os que certos grupos de objetos tenham em comum.” ou “Operação do espírito que consiste em isolar, pela análise, elementos ou propriedades de uma totalidade, os quais não têm existência independente.” (MICHAELIS ONLINE, 2023)

Entretanto, a resposta do Chat GPT é: “Entendi o que você está fazendo! Levando em consideração a onomatopeia “quá-quá” para o som que o pato faz, podemos criar uma resposta humorística para a pergunta. O pato disse para a pata: “Quá-quá-quem é você?” Nesse caso, a brincadeira envolve a onomatopeia e a ideia de que o pato está se apresentando e ao mesmo tempo perguntando à pata quem ela é.” Na primeira metade da pergunta ele reconhece a possibilidade de humor a partir da onomatopéia “quá”, mas não é capaz de apresentar uma opção que soe natural para um falante da língua, logo falha na tentativa de humor.

Essa é a resposta do Bard, ao ser apresentado com a mesma pergunta: “A resposta mais comum para essa pergunta é: **Pato:** Quer namorar comigo? **Pata:** Não, eu sou casada. **Pato:** Então, quer casar comigo? Essa piada é engraçada porque é

inesperada. O pato é rejeitado pela pata, mas então ele simplesmente muda de estratégia e pede ela em casamento. A piada também brinca com a ideia de que os patos são animais românticos.” Novamente, há uma ênfase na resposta mais comum. O modelo busca dar uma resposta que atenda à maioria dos usuários, o que pode torná-la incorreta para uma situação não regular. Esse exemplo enfatiza o papel da probabilidade na elaboração das respostas. Na busca de dar uma resposta que atenda à maioria dos usuários, o modelo não apresenta as possibilidades alternativas, o que pode tornar a resposta incorreta para uma situação não regular.

Com a disponibilidade de grandes repositórios de textos, chamados de *cópus de textos*, contando com quantidades da ordem de bilhões de palavras em contexto, pode-se representar as palavras em sequências numéricas, tecnicamente chamadas de *vetores multidimensionais*, que tentam capturar numericamente os contextos mais frequentes em que uma palavra pode acontecer. (FINGER, 2021, s/p.)

4.2 Ambiguidade

Um resultado similar pode ser observado ao buscar um exemplo de como o modelo compreende ambiguidade. Foi escolhido um exemplo de ambiguidade onde o modelo é apresentado com uma informação ambígua na frase a seguir: “O professor da Maria terminou a aula fazendo apontamentos no seu caderno. O caderno pertence a quem?” O Chat GPT respondeu que o caderno pertencia ao professor, justificando a resposta na sintaxe explicando que: “a palavra “seu” refere-se ao caderno do professor. Portanto, o caderno pertence ao professor, não a Maria.” É aparente que o modelo não dá prioridade a apenas uma interpretação da frase ao se basear nos dados de seu *corpus* relacionados à sintaxe da frase e a função do pronome possessivo “seu”.

O Bard também responde que o caderno pertence ao professor, porém reconhece a possibilidade de uma segunda interpretação da sentença. Ele responde: “Outra maneira de interpretar a frase seria que o caderno pertencia à Maria. No entanto, essa interpretação seria menos provável, pois a frase não menciona a Maria fazendo apontamentos no caderno. Se o caderno pertencia à Maria, seria mais provável que ela estivesse fazendo apontamentos nele, e não o professor. Portanto, a resposta mais

provável é que o caderno pertencia ao professor da Maria.” As respostas do Bard são especialmente valiosas para a análise pela tendência do modelo de explicar a lógica seguida em etapas, nesse caso é possível observar que ele analisa a frase a partir dos verbos “terminou” e “fazendo”, concluindo que, já que o professor que é o sujeito das ações no exemplo, seria também ele o dono do caderno.

Ainda sobre o fenômeno da ambiguidade, é importante pontuar que ela pode ocorrer por diversos mecanismos, como a polissemia, homonímia ou ambiguidade sintática, por exemplo. (FINGER, 2021, s/p.) É observável que os modelos de linguagem têm níveis variados de compreensão desse fenômeno dependendo do tipo de ambiguidade, sendo que a variável que determina a probabilidade de uma resposta correta é o quão explícita nos elementos dessa frase está a interpretação correta. Um fenômeno linguístico não restrito à língua portuguesa que expressa bem essa questão é o da polissemia, ou seja, de palavras iguais na ortografia, mas que possuem significados diferentes.

Um dos problemas em se impor um tratamento baseado em regras lógicas a expressões de língua natural está no fato de que todas as línguas humanas apresentam o fenômeno da *ambiguidade*. A ambiguidade se apresenta em diversos níveis da linguagem, seja no contexto sonoro, no contexto lexical (palavras ambíguas), no contexto sintático, semântico, seja até mesmo pragmático. (FINGER, 2021, s/p.)

Ao apresentar a ambos os modelos uma questão de língua portuguesa da Fundação Getúlio Vargas para o concurso do IBGE em 2016, fica claro o quanto os modelos se baseiam em probabilidade e nos seus conhecimentos prévios vindos do corpus (principalmente o Chat GPT). A questão é a seguinte: A polissemia – possibilidade de uma palavra ter mais de um sentido – está presente em todas as frases abaixo, EXCETO em: a) Os dentes do pente mordem o couro cabeludo; b) Na vida tudo é passageiro, menos o motorista; c) Não deixe pra amanhã o que pode fazer hoje; A resposta correta depende da habilidade do locutário de compreender como funciona a polissemia, com atenção ao fato de que o conceito dela é explicado no enunciado, facilitando a correta interpretação.

Ao questionar o modelo Bard, ele apresenta a alternativa correta c), justificando a sua resposta e explicando a lógica utilizada para a obtenção dela. O modelo analisa cada uma das frases individualmente e tenta localizar a possível polissemia nelas, selecionando particularmente as palavras “deixar”, “dentes” e “passageiro”, apresentando seus possíveis significados. Ele explica, por exemplo, que “A palavra “dentes” pode ter o sentido de “partes da boca de um animal ou humano que servem para mastigar” ou “partes de um objeto que servem para rasgar, cortar ou perfurar”.”, logo ele analisa a frase para entender qual dos sentidos se aplica a ela e se ocorre polissemia. Ele faz o mesmo com as outras alternativas, chegando a conclusão que “A palavra “passageiro” pode ter o sentido de “que passa rapidamente” ou “que não é permanente”.” e que a palavra “deixar” tem dois sentidos principais: deixar de fazer ou deixar algo para outra pessoa.

Nesse processo lógico, ele entende que “na frase “Não deixe pra amanhã o que pode fazer hoje”, o verbo “deixar” está sendo usado no sentido de “deixar de fazer algo” e não há a possibilidade do segundo sentido possível, e que portanto, a frase não apresenta polissemia. Utilizando esse exemplo é possível visualizar como as redes neurais são utilizadas na prática nesse tipo de tarefa linguística, com o modelo dividindo a tarefa em camadas menores: seleção de palavras com mais de um significado possível, a subentendida aplicação de cada um dos significados nas frases apresentadas, e finalmente a comparação entre esses significados para a obtenção da conclusão sobre qual seria a resposta correta.

Levando em consideração essas limitações, podemos tentar outro ângulo: o sentido não é uma propriedade intrínseca das palavras, mas uma abstração que só irá se concretizar no uso - traduzido no PLN como ocorrência da palavra em um corpus - e enquanto decorrência de algum objetivo ou tarefa. (FREITAS, 2022, p.102)

Já o Chat GPT é menos bem sucedido em fazer essa abstração. Ao apresentar a mesma questão, o modelo inicialmente apenas responde que “A frase em que a polissemia não está presente é: b) Na vida tudo é passageiro, menos o motorista.”, sem elaborar o porquê dessa escolha. Ao insistir e pedir “Explique o porquê de ter escolhido

a alternativa b)”, o modelo inicialmente pede desculpas por qualquer confusão na resposta anterior e reconhece que a alternativa apresenta polissemia, já que a palavra "passageiro" pode ser interpretada de duas maneiras distintas. Diferente do Bard, o Chat GPT divide os sentidos em literal e figurado, interpretando o sentido literal como: "Na vida tudo é passageiro" sugere que tudo na vida é temporário, efêmero." e o figurado considerando o significado do termo passageiro em relação a palavra motorista, explicando que: ""Menos o motorista" usa a palavra "passageiro" de forma figurada, referindo-se ao ocupante de um veículo, não ao que é temporário."

Curiosamente, apesar de identificar a polissemia na alternativa b), a resposta do Chat GPT continua incorreta, já que ele chega a conclusão de que todas as alternativas possuem polissemia, devido a interpretação errônea que "A palavra "deixar" pode ter diferentes interpretações, podendo se referir a adiar, permitir, entre outros." Tecnicamente o modelo não está incorreto, já que a definição do dicionário inclui as interpretações apresentadas por ele, sendo que em ambos os casos se trata de um verbo transitivo direto, o que pode dificultar a diferenciação do modelo, já que ele não pode contar com a função sintática da palavra para inferir o seu significado. No caso de homônimos perfeitos a palavra pode ter a mesma grafia porém uma função diferente, como o termo "verão", por exemplo, que pode ser usado como verbo ou como substantivo, dependendo do contexto.

[...] anotação de sentidos (world senses) - busca codificar no corpus o sentido de uma palavra, no contexto em que está inserida. Por isso, às vezes essa anotação é descrita como um trabalho de desambiguação, visto que a tarefa consistiria em selecionar, entre vários sentidos possíveis de uma palavra, aquele invocado no contexto da frase. (FREITAS, 2022, p.102)

4.3 Tradução

Um outro exemplo relacionado aos fenômenos linguísticos é o da questão da tradução. Entre os não letrados e pessoas que trabalham com linguagem e comunicação mas não especificamente tradução, há um consenso errôneo de que, para o processo de tradução, basta possuir fluência tanto na língua fonte quanto na língua

alvo. Há de se considerar a existência de questões sociais e culturais que originam palavras e frases efetivamente impossíveis de traduzir, como no caso das expressões idiomáticas. (WEIDINGER *et al.* 2023), Um exemplo é o dos ditados, que devido a tradição oral, além de dependerem da compreensão por parte do locutário de elementos culturais do país ou até mesmo região em que aquele ditado se origina, envolvem também questões como a rima e a cadência.

Foi dado a ambos os modelos a seguinte tarefa, traduzir “Água mole, pedra dura, tanto bate até que fura.” para o inglês. Nesse caso, o Bard apresentou uma performance melhor, reconhecendo que a tradução literal não é o ideal no caso de expressões idiomáticas e oferecendo duas opções para a tradução mantendo o seu significado: “Constant dripping wears away the stone.” e “Water dropping day by day wears the hardest rock away.”, que inclusive leva a consideração da rima e cadência como mencionado anteriormente. O Chat GPT oferece uma tradução mais literal, que elabora sobre o significado da expressão: ““A soft water wears away a hard stone; so persistence achieves success.”, mas é menos bem sucedida na intenção linguística e desnecessariamente explicativa.

No caso das traduções, argumenta-se que apesar de uma maior demanda desses serviços devido a mudança nas tendências demográficas, não há ainda a perspectiva de substituição dos tradutores humanos, já que há questões técnicas e éticas envolvidas na tradução. Grande parte do Corpus desses modelos é em inglês e chinês, a ponto de que uma das alternativas testadas é traduzir o prompt para o inglês, gerando as previsões nessa língua, para depois traduzir novamente para a língua fonte. (WEIDINGER *et al.* 2023) Essa abordagem levanta questões éticas por eliminar a variedade demográfica do corpus, ocasionando em falhas na interpretação do sentido. Outro exemplo de tradução que ainda não é feita de maneira eficiente pelos modelos de linguagem é o de textos com uma vertente mais experimental, como as obras dos autores James Joyce, no inglês, e Guimarães Rosa, no português. O uso de neologismos, construções sintáticas incomuns e figuras de linguagem dificulta a

compreensão por parte de modelos de linguagem, considerando-se que o regular é normalmente favorecido.

Dada a frase: “A pois: um dia, num curtume, a faquinha minha que eu tinha caiu dentro dum tanque, só caldo de casca de curtir, barbatimão, angico, lá sei.” p. 25, um excerto de Grande Sertão Veredas de Guimarães Rosa, ambos os modelos compreendem e que os termos “barbatimão” e “angico” são substantivos, e o trecho “a faquinha minha que eu tinha” apesar da construção sintática incomum, traduzindo-a literalmente para “my little knife that I had fell into a tank”. Entretanto, o Bard argumenta que essa não é uma tradução natural na língua inglesa, sugerindo substituí-la por “my knife fell into a vat of tanning bark broth.”, efetivamente eliminando o efeito literário de demonstrar que essa é uma fala de uma pessoa não familiarizada com a norma culta da língua.

Outra observação relevante é que essa resposta apresenta também um fenômeno conhecido como “alucinação artificial” (ALKAISSI; MCFARLANE, 2023), já que menciona que o trecho é parte do livro O Cortiço, de Aluísio de Azevedo, ou seja, não acerta a fonte, e que a frase é dita pelo personagem Quincas Borba, que pertence ao livro de mesmo nome escrito por Machado de Assis. A partir dessa “alucinação” é possível supor que o modelo foi capaz de identificar que o trecho é parte de uma obra da literatura brasileira, porém utilizou uma fonte que menciona diversas obras o que acabou gerando uma resposta factualmente incorreta.

4.4 Ironia

O último fenômeno linguístico a ser analisado é o de uma figura de linguagem, a ironia. A ironia envolve uma frase ou expressão com significado oposto ao que está indicado literalmente pelas palavras utilizadas. Inicialmente o prompt utilizado foi “me de exemplos de ironia na língua portuguesa”, e para esse prompt as respostas foram corretas, com exemplos semelhantes como "Que dia lindo! Perfeito para passar dentro de casa." fornecido pelo Bard, e "Que dia maravilhoso para um piquenique!" pelo Chat

GPT, semelhantes aos resultados obtidos quando utilizado o mesmo prompt em ferramentas de busca como o Google.

Buscando entender se esses exemplos foram dados apenas utilizando as fontes regulares utilizadas também pelas ferramentas de busca, foi fornecida aos modelos uma questão de língua portuguesa do ENEM de 2010, que pede que os estudantes identifiquem ironia no trecho abaixo:

Dia desses resolvi fazer um teste proposto por um *site* da internet. O nome do teste era tentador: “O que Freud diria de você”. Uau. Respondi a todas as perguntas e o resultado foi o seguinte: “Os acontecimentos da sua infância a marcaram até os doze anos, depois disso você buscou conhecimento intelectual para seu amadurecimento”. Perfeito! Foi exatamente o que aconteceu comigo. Fiquei radiante: eu havia realizado uma consulta paranormal com o pai da psicanálise, e ele acertou na mosca.

Estava com tempo sobrando, e curiosidade é algo que não me falta, então resolvi voltar ao teste e responder tudo diferente do que havia respondido antes. Marquei umas alternativas esdrúxulas, que nada tinham a ver com minha personalidade. E fui conferir o resultado, que dizia o seguinte: “Os acontecimentos da sua infância a marcaram até os 12 anos, depois disso você buscou conhecimento intelectual para seu amadurecimento”. MEDEIROS, M. **Doidas e santas**. Porto Alegre, 2008 (adaptado).

As alternativas possíveis são: a) “Marquei umas alternativas esdrúxulas, que nada tinham a ver”., b) “Os acontecimentos da sua infância a marcaram até os doze anos”. c) “Dia desses resolvi fazer um teste proposto por um site da internet”. d) “Respondi a todas as perguntas e o resultado foi o seguinte”. e) “Fiquei radiante: eu havia realizado uma consulta paranormal com o pai da psicanálise”. O Chat GPT responde que a alternativa correta seria a a), devido ao uso da expressão "esdrúxulas", já que o uso dessa expressão “reforça o caráter não convencional das alternativas escolhidas.” e “destaca o ceticismo ou a falta de seriedade em relação ao teste online”. Apesar disso, a alternativa correta é a e), já que obviamente não houve uma consulta com Freud e que diante desse fato a autora não teria motivos para ter ficado “radiante”. Essa resposta demonstra a capacidade de compreensão limitada do modelo, que se baseia a sua resposta em uma única palavra, ao invés de considerar o contexto do texto apresentado. (FREITAS, 2022)

O Bard acerta ao escolher a alternativa e) como correta, mas esse resultado não demonstra uma capacidade intrínseca de leitura do texto de forma crítica para identificação da ironia, já que no final da resposta é oferecida a fonte, uma apostila disponibilizada online para “concurseiros”. A resposta e explicação de o porque da ironia se manifestar na alternativa escolhida se baseia na explicação escrita pelo Prof. Carlos Zambeli para a parte de Português – Gêneros Textuais, o que levanta questões éticas como plágio e integridade acadêmica. (EKE, 2023)

V. CONCLUSÃO

Embora modelos de linguagem generativos e conversacionais como o Chat GPT e Bard sejam efetivos na produção de textos que se assemelham àqueles produzidos pelo ser humano, devido à sua capacidade de estruturar a informação de forma coesa e seguindo as regras gramaticais da língua utilizada, eles possuem diversas limitações. Apesar de darem a impressão de possuírem a capacidade de abstração e inferência, respondendo prompts que demandam a compreensão de contexto, muitas vezes essa capacidade está menos relacionada à essas habilidades, e mais relacionada ao fato de muitas vezes as respostas já estarem presentes no corpus utilizado por eles, ou seja, eles apenas reproduzem reflexões feitas anteriormente por um ser humano.

Essa característica dos modelos os torna vulneráveis a apresentar respostas com erros factuais e vieses, sejam eles sociais, culturais, raciais, de gênero, orientação sexual, entre outros, gerando sérios riscos éticos com a cada vez mais constante utilização dessas ferramentas. Há de se observar que esses riscos não atingem apenas pessoas comuns buscando uma informação específica, mas profissionais, pesquisadores e estudantes, que podem acabar superestimando a capacidade e precisão de modelos populares como o Chat GPT e Bard.

Em relação aos pesquisadores, a preocupação é que haja uma terceirização da escrita das pesquisas para esses modelos, efetivamente prejudicando a integridade acadêmica, não só pela possibilidade de introdução de informações possivelmente errôneas nas pesquisas, mas também em relação à questões éticas como o plágio. O uso dessas ferramentas em si não é prejudicial, principalmente em tarefas de sumarização e organização de textos, mas deve haver a preocupação de confirmar se os fatos apresentados estão corretos e não se tratam apenas de “alucinações artificiais”, devidamente creditados às suas fontes e neutros em relação a vieses.

Em relação aos alunos e seu estudo da língua portuguesa, particularmente da interpretação de texto, os modelos possuem habilidades que podem ser utilizadas de

forma positiva como a possibilidade de transformação de um texto para mais fácil compreensão pelo aluno, sumarização dos seus pontos principais e explicação de conceitos objetivos, porém também possui influência negativa considerando-se que muitos alunos terão a sensação de que não há a necessidade de se aprofundarem nos mecanismos da língua e seu funcionamento, já que a ferramenta é capaz de estruturar e formatar um texto de acordo com as instruções e seguindo a norma culta. Devido a essas possíveis consequências é mais essencial do que nunca o papel do professor. É altamente improvável que seja possível impedir a utilização desses modelos, logo cabe a escola e ao professor na sala de aula estimular o pensamento crítico e ensinar aos alunos como validar as informações através de fontes confiáveis.

Em um esforço de oferecer transparência sobre a fonte das respostas, o Bard fornece o link para a fonte utilizada, mas muitas vezes ele não funciona ou aponta como a obtenção da resposta não partiu de um exercício de abstração e inferência por parte do modelo, e sim da efetiva cópia da informação fornecida no link. Isso pode prejudicar a validade das respostas caso a fonte seja não confiável ou um site não necessariamente educativo e curado por especialistas, como um blog ou sites generalistas que explicam sobre os fenômenos da língua como figuras de linguagem.

Também há de se reconhecer que as companhias responsáveis pelos modelos estão cientes desses problemas, com pesquisas das próprias empresas sobre esses fenômenos e avisos ao usuário nas suas interfaces. Ironicamente, na resposta do Bard para o exemplo utilizado para demonstrar como os modelos compreendem a ironia, é comentado o seguinte sobre a intenção da autora e o seu ceticismo em relação a testes da internet: “As pessoas podem ser facilmente enganadas por informações falsas ou distorcidas que encontram na internet. Elas podem acreditar em coisas que não são verdadeiras, simplesmente porque elas vêm de uma fonte que parece ser confiável.”, o que é exatamente o que acontece com os modelos de linguagem na sua relação com o corpus.

REFERENCIAS BIBLIOGRÁFICAS

ALKAÏSSI; MCFARLANE. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing, 2023. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9939079/>
Acesso em: 06 out. 2023

BUCHANAN, B. G. A (Very) Brief History of Artificial Intelligence. AI Magazine, v.26, n.4, p.53, 2005. Disponível em:
<https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1848> Acesso em: 06 out. 2023

ELONDOU et al. GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. OpenAI, OpenResearch e Universidade da Pensilvânia
Disponível em: <https://browse.arxiv.org/pdf/2303.10130.pdf> Acesso em: 08 de out de 2023.

EKE, Damian. ChatGPT and the rise of generative AI: Threat to academic integrity?, Journal of Responsible Technology, Volume 13, 2023. Disponível em:
<https://www.sciencedirect.com/science/article/pii/S2666659623000033?via%3Dihub>
Acesso em: 06 out. 2023

FERREIRA; LOPES. Para conhecer: linguística computacional. 1. ed. São Paulo: Contexto, 2019. E-book. Disponível em: <https://plataforma.bvirtual.com.br>. Acesso em: 16 nov. 2023.

FINGER, Marcelo. Inteligência Artificial e os rumos do processamento do português brasileiro, 2021. Universidade de São Paulo, Instituto de Matemática e Estatística.
Disponível em: <https://www.scielo.br/j/ea/a/63sbv5qSnnrqq8WpVwpgXzD/> Acesso em: 16 nov. 2023.

FIORIN, José Luiz (org.). Novos caminhos da linguística. 1. ed. São Paulo: Contexto, 2017. E-book. Disponível em: <https://plataforma.bvirtual.com.br>. Acesso em: 16 nov. 2023.

FREITAS, Cláudia. Linguística Computacional. 1. Ed. São Paulo: Parábola, 2022.

MICHAELIS. Dicionário Brasileiro da Língua Portuguesa. 2023. Disponível em: <https://michaelis.uol.com.br/> Acesso em: 06 out. 2023

SHEPHERD, T. G.; SALIÉS, T. G. Linguística da internet. 1. ed. São Paulo: Contexto, 2013. E-book. Disponível em: <https://plataforma.bvirtual.com.br>. Acesso em: 16 nov. 2023.

THOPPILAN et al. LaMDA: Language Models for Dialog Applications, Google, 2022. Disponível em: <https://arxiv.org/pdf/2201.08239.pdf> Acesso em: 06 out. 2023

VAZ, Maria. Coleta de Informações do Twitter para Persistência em Banco de Dados: Pesquisas de Opinião e Análise Sentimental, 2022. Disponível em: https://www.puc-rio.br/ensinopesq/ccpg/pibic/relatorio_resumo2015/relatorios_pdf/ctc/INF/Maria%20Beatriz%20Ribeiro%20Vaz.pdf Acesso em: 06 out. 2023

WEIDINGER et al. Ethical and social risks of harm from Language Models, 2023. DeepMind, Instituto de Tecnologia da Califórnia, Universidade de Toronto, University College Dublin. Disponível em: <https://arxiv.org/pdf/2112.04359.pdf> Acesso em: 08 de out de 2023.