Ética na era dos Modelos de Linguagem Massivos (LLMs): um estudo de caso do ChatGPT

Mateus R. Figênio[1], **Luiz Gomes-Jr**[1]

[1]Departamento Acadêmico de Informática
Universidade Tecnológica Federal do Parana (UTFPR)
80.230-901 – Curitiba – PR – Brazil

mateusfigenio@alunos.utfpr.edu.br, lcjunior@utfpr.edu.br

Abstract. *This article aims to discuss ethical issues related to ChatGPT, a conversational style language model. From related works that support the concept of Massive Language Models (LLMs) and that work paradigms of ethical analysis and good practices for the development of Artificial Intelligences (AI), we explore how ChatGPT perpetuates already recognized problems of LLMs and we observe that their greater ability to generalize increases dangers of bias and prejudice. As a conclusion, we suggest greater incentives to reduce efforts for larger models, in favor of efforts for better documented datasets, interpretable models and approaches aimed at language understanding.*

Resumo. *Este artigo tem como objetivo discutir questões éticas relacionadas ao ChatGPT, um modelo de linguagem de estilo conversacional. A partir de trabalhos correlatos que fundamentam o conceito de Modelos de Linguagem Massivos (LLMs) e que trabalham paradigmas de análise ética e boas práticas para o desenvolvimento de Inteligências Artificiais (IA), exploramos como o ChatGPT perpetua problemas já reconhecidos de LLMs e observamos que sua maior capacidade de generalização aumenta perigos de enviesamento e preconceito. Concluímos reforçando apelos por maiores incentivos à diminuição de esforços por maiores modelos, em favor de esforços por bases de dados melhor documentadas, modelos interpretáveis e por abordagens voltadas ao entendimento de linguagem.*

1. Introdução

Em 30 de novembro de 2022, a *OpenAI*, empresa de pesquisa e desenvolvimento de Inteligência Artificial, lançou para livre acesso e testagem sua nova plataforma de chat bot ainda em fase de protótipo, o *ChatGPT*. A empresa, já conhecida pelos seus outros modelos de IA para diversas aplicações, como *GPT-2* e *GPT-3* para geração de texto e *DALL-E* e *DALL-E 2* para geração de imagens, agora lança um modelo de linguagem conversacional. Modelos de linguagem de resposta à *prompts* de usuário já existiam, mas a qualidade discursiva das respostas e a responsividade geral do novo programa em conversas com os usuários trouxeram muita atenção a essa nova ferramenta, tendo alcançado 1 milhão de usuários uma semana após seu lançamento e mais de 100 milhões de usuárioa 2 meses depois [Milmo 2023].

IAs de produção de imagens e imagens estilizadas, como *DALL-E*, também lançada pela *OpenAI*, trouxeram atenção para questões de autoria e propriedade intelec-

tual no contexto de produção de conteúdo artístico por máquinas, mas quais são os impactos e considerações a serem pensadas quanto a tecnologias que modelam a linguagem humana? Novos e maiores modelos de IA vêm sendo desenvolvidos e implementados para as mais diversas tarefas e aplicações. Lidar com suas implicações éticas e sociais é imperativo para garantir um seguro desenvolvimento e impacto na sociedade. Os Modelos de Linguagem Massivos (LLMs) se destacam nesse contexto por (i) serem treinados com uma massiva quantidade de dados, muitas vezes sem curadoria ou documentação, (ii) terem um apelo prático para usuários não técnicos, (iii) serem uma tecnologia de desenvolvimento caro, acessível a poucas instituições no mundo.

Este trabalho tem como objetivo trazer um panorama dessa nova tecnologia, sintetizando seu desenvolvimento no contexto de desenvolvimento de modelos de linguagem e elaborando paradigmas de análise de considerações éticas, tanto no âmbito geral, quanto em aspectos mais específicos de desenvolvimento e implementação (Seção 2). Além disso, o artigo busca destacar questões correntes envolvendo esse novo sistema e apontar direcionamentos futuros de pesquisa e debate quanto ao tema (Seção 3).

2. Fundamentos e Trabalhos Correlatos

Nesta seção, são trabalhados conceitos fundamentais para a discussão, como a definição de Modelos de Linguagem e como se deu o desenvolvimento do *ChatGPT*, e as bases éticas de análise de sistemas de IA, tanto uma revisão de princípios éticos gerais, quanto um resumo de práticas já desenvolvidas para auditar tais sistemas.

2.1. Modelos de Linguagem

Modelo de Linguagem se refere a qualquer sistema treinado para a tarefa única de prever uma série de caracteres, sejam letras, palavras ou sentenças, de forma sequencial ou não, dado um contexto anterior ou adjacente [Bender and Koller 2020]. Tal definição engloba os recentes desenvolvimentos em modelos de redes neurais, mas também abordagens passadas, como n-gramas e vetores de palavras. Esses também se beneficiaram da utilização de grandes volumes de dados para melhorar seu desempenho em tarefas específicas.

[Bender et al. 2021] apontam que as tendências nos anos recentes na área de desenvolvimento de Modelos de Linguagem se deram em duas linhas, da incorporação de palavras (*word embedding*) e de modelos *transformers*. O primeiro destes representou uma melhora nos resultados de diversos testes de desempenho, enquanto reduzia a quantidade de dados rotulados necessários para diversas atividades supervisionadas. Em contrapartida, modelos *transformers* tem se beneficiado continuamente de maiores arquiteturas e maiores quantidades de dados, sendo destacada sua capacidade de ser posteriormente aperfeiçoados para tarefas específicas. Alguns dos maiores modelos desse tipo redefiniram o sentido de grande de sua classificação, sendo mais preciso caracterizá-los como Modelos de Linguagem Massivos (LLMs). Alguns destes, como o *GPT-3* da *OpenAI*, chegando a 175 bilhões parâmetros e 570GB de dados de treinamento, várias ordens de magnitude acima dos modelos grandes anteriores.

2.2. InstructGPT

O *ChatGPT* é um aplicativo implementado sobre um modelo de linguagem do tipo *InstructGPT* que interage de forma conversacional com seu interlocutor, o que o permite

responder perguntas subsequentes, admitir erros, desafiar premissas incorretas e rejeitar propostas inapropriadas, como descrito pela [OpenAI 2022].

O modelo *InstructGPT* é resultado do trabalho da *OpenAI* de desenvolver modelos de linguagem cujas respostas melhor se alinhem às instruções e expectativas de seus usuários. Os pesquisadores aperfeiçoaram um modelo pré-treinado do *GPT-3* (*Generative Pre-trained Transformer* de terceira geração) para responder a instruções de usuários humanos, prezando por melhor atender às expectativas destes, pela veracidade e pela segurança de suas respostas. Tendo esse ponto de partida, o modelo foi então treinado com aprendizagem por reforço a partir de *feedback* humano [Ouyang et al. 2022]. O desenvolvimento de um modelo do tipo *InstructGPT* trouxe resultados mais alinhados às expectativas dos usuários, com respostas menos tóxicas e mais verdadeiras quando comparado a resultados do *GPT-3*, mesmo este último contando com mais de 100 vezes mais parâmetros de entrada para comparação, demonstrando ser um caminho viável e eficiente para o desenvolvimento de IAs mais alinhadas ao usuário. Além disso, essa abordagem demonstrou ter uma maior capacidade de generalização, tanto para rotuladores mais retidos, quanto para generalizar instruções que não estavam presentes em seu treino com *feedback* humano. Apesar disso, *InstructGPT* ainda comete erros simples, como evitar dar uma resposta a uma pergunta simples preferindo se manter isento, tender a responder diretamente perguntas fundamentadas em premissas falsas e a inventar fatos.

2.3. Ética IA

Como ponto de partida para começar a pensar sobre questões éticas do desenvolvimento de IAs, partimos de [Bender et al. 2021] que aborda esses perigos em 3 aspectos, do risco ambiental, do risco de grandes bases de dados não documentadas e o de entender o sucesso de modelos de linguagem dentro de seu devido contexto. Esses aspectos trabalhados pelos autores são descritos a seguir.

O primeiro destes aspectos abordados por [Bender et al. 2021], decorre do desenvolvimento de modelos cada vez maiores em termos de parâmetros, tempo de treino e tamanho de *datasets* para treinamento, implicando em modelos com um custo maior de operação em termos financeiros e ecológicos. Este último, implica em um maior pegada de carbono gerado pelo treinamento dos modelos, agravando os efeitos da mudança climática. Esta, por sua vez, afeta desproporcionalmente populações periféricas que sofrem o pior do impacto ambiental enquanto são as que menos se beneficiam dos ganhos obtidos pelo desenvolvimento de modelos maiores. Já o crescimento exponencial das exigências computacionais e financeiras para o treinamento e desenvolvimento de LLMs afeta a reprodutibilidade das pesquisas pela ampla comunidade científica, restringindo o acesso de novas tecnologias a empresas que podem arcar com tais custos, e limitando a validação destes modelos por outrem.

Em relação ao crescente tamanho dos *datasets* necessários para o treinamento dos LLMs, [Bender et al. 2021] assinala quanto ao risco da chamada dívida documental, uma situação em que os dados de treinamento não são documentados e são grandes demais para serem documentados após o treinamento dos modelos. Como a principal fonte de dados para tais modelos é a internet, os modelos estão sujeitos a reproduzirem pontos de vista hegemônicos e a incorrer na reprodução de vieses tóxicos às populações marginalizadas, dado que o acesso à internet é condicionado por fatores sociais. Sem documentação,

não se pode tentar entender as características dos dados de treinamento, para mitigar problemas já conhecidos ou mesmo os desconhecidos.

Além das questões mais concretas de impactos ambientais e sociais, há também a questão metodológica a se analisar se LLMs estão sendo avaliados e aplicados a contextos de utilização corretos. Autores como [Bender et al. 2021] chegam a classificar tais modelos como "Papagaios Estocásticos", pois, fundamentalmente e independente de seu tamanho, tais modelos foram desenvolvidos com a finalidade de predição de tokens e linhas de texto, não o entendimento de linguagem. Assim, pesquisadores reivindicam por investimento no desenvolvimento de novos modelos focados no objetivo de entendimento de linguagem.

2.4. IA Digna de Confiança

Diversos trabalhos abordam práticas de análise e validação desses sistemas, como as descritas no relatório "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims", de [Brundage et al. 2020]. Escrito por autores com atuações em diversas empresas, universidades, institutos e organizações voltadas ao estudo e desenvolvimento de IA, o relatório reconhece que o desenvolvimento de IAs levanta preocupações quanto a manutenção e amplificação de vieses e preconceitos, a perda de privacidade, desinformação e danos sociais associados a reconhecimento facial e avaliação de risco criminal. Assim, os autores empreendem na ampliação do arcabouço de ferramentas pelos quais seja possível garantir um desenvolvimento responsável desses sistemas, tendo como ponto central o aspecto de reivindicações verificáveis. Este recurso permitiria que as organizações que desenvolvem sistemas de IA façam reivindicações sobre os sistemas que constroem, e que outras partes, potencialmente as mais afetadas por esses novos sistemas, consigam avaliar essas reivindicações e assegurar um desenvolvimento responsável. Exemplos de reivindicações verificáveis seriam: "Nosso sistema de IA opera em uma arquitetura de dados segura e privada", "Durante o desenvolvimento do sistema seguimos protocolos de segurança estipulados", "Reportaremos todo e qualquer caso em que o sistema de IA venha a causar danos negativos à sociedade e pessoas". São declarações para as quais argumentos e evidências podem ser trazidos para apoiar sua veracidade ou comprovar sua falsidade.

Partindo do entendimento de que processos de desenvolvimento de IA são sistemas sociotécnicos que envolvem instituições, *software* e *hardware*, os autores dividem as ferramentas sondadas nesses três níveis de mecanismos. Mecanismos Institucionais compreendem valores, incentivos e a responsabilização de instituições envolvidas no desenvolvimento de IA, visando criar canais para tornar aqueles que desenvolvem esses sistemas responsáveis pelos maus associados a esse desenvolvimento. Exemplos desse mecanismo são: auditorias, exercícios *Red Team*, *Bias and Safety Bounties* e compartilhamento de incidentes de IA. Para lidar com questões envolvendo os sistemas de IA em si e suas propriedades, os autores trazem Mecanismos de *Software* que tratam da promoção do melhor entendimento e supervisão desses sistemas. Exemplos de ferramentas são: trilhas de auditoria, esforços por melhor interpretabilidade de modelos e *Machine Learning* que preserve a privacidade de dados. Já os recursos físicos de computação envolvidos no desenvolvimento de IA, sua segurança e privacidade, são tratados por Mecanismos de *Hardware*. Suas ferramentas contam com: componentes de segurança de *hardware* para ML, medidas de computação de alta precisão e apoio de poder computacional para

pesquisadores.

3. Exploração de Questões Éticas no ChatGPT

Nesta seção, discutimos diversas questões éticas do *ChatGPT*, de sua origem e operação, concluindo com considerações sobre a verdade em modelos de linguagem.

3.1. Expectativas e alinhamento

Um modelo de linguagem do tipo *InstructGPT* é treinado para que suas respostas melhor se alinhem às expectativas de seus usuários. Tal fim leva a pergunta ”para quais usuários o modelo é treinado a se alinhar?”. No artigo em que apresentam o novo modelo, [Ouyang et al. 2022] apontam que para o treinamento do modelo de *feedback* humano, foram buscados participantes que divergissem entre si e para melhor se adequar a valores humanos, mas que isto não foi suficiente para garantir uma maior abrangência de paradigma do modelo. Os autores reconhecem que o modelo é alinhado aos seus rotuladores, aos pesquisadores que desenvolveram o modelo, aos *prompts* utilizados como base para pesquisa e, por fim, reconhecendo que esses participantes não representam todos os potenciais usuários nem todas as pessoas as quais o sistema pode vir a impactar.

Assim, pesando as considerações quanto ao alinhamento do modelo, tendo ciência da capacidade de generalização para além do treinamento humano e com o conhecimento de que vieses que já permeiam LLMs, como o viés anti-muçulmano apresentado por [Abid et al. 2021], é fundamental que sejam estabelecidas práticas de responsabilização, auditoria e segurança para garantir que LLMs não impactem negativamente a sociedade.

Problemas de toxicidade de modelos conversacionais se tornaram mais visíveis com o lançamento do *Bing Chat*, uma integração de um modelo similar ao *ChatGPT* na ferramenta de navegação web *Bing*, que fez notícia com suas respostas rudes, depressivas e agressivas aos usuários [Adorno 2023]. Apesar da fundação compartilhada que *Bing Chat* têm com o *ChatGPT*, a discrepância entre o tom de suas respostas se deve ao sistema de segurança de cada um, um sistema implementado à parte para garantir que apenas respostas seguras e sem toxicidade cheguem ao usuário final. O sistema de segurança treinado implementado para o *Chat* é mais robusto e melhor validado, mas é em si fonte de controvérsias e debates éticos após se saber que o treinamento desse sistema foi feito por trabalhadores quenianos pagos 2 dólares a hora, segundo o noticiado por [Perrigo 2023]. Por fim, se tais modelos precisam de filtros de segurança e sistemas de monitoração para sua segura utilização, isso leva a questionar a real segurança de sua utilização na sociedade.

3.2. Modelos confiáveis

Com a crescente popularidade de sistemas de IA, vem se formando um novo mercado de produtos para concorrer diretamente ao *ChatGPT*, sendo um deles *Bard*, uma IA conversacional anunciada em 8 de fevereiro pela *Google*. Porém, no dia seguinte ao lançamento, a *Alphabet*, empresa pai da *Google*, registrou uma queda de $100 bilhões de dólares no seu valor no mercado de ações, enquanto a Microsoft observou um aumento de %3 no valor de suas ações. Isso se deu, pois, em seu anúncio de apresentação, *Bard* cometeu um erro factual ao afirmar que o telescópio James Webb foi o primeiro a fotografar exoplanetas [Olson 2023]. Tal feito, foi realizado, na verdade, pelo ESO (*European Southern Observatory*) em 2004.

Esse evento revisita o debate sobre modelos de processamento de linguagem natural poderem ser ou não confiáveis como fontes de verdade. Por sua aprimorada reprodução da linguagem humana, é tentador considerar que sua produção seja baseada em fatos ou verdades do mundo. O CEO da *OpenAI*, Sam Altman, afirmou em seu Twitter que seria um erro confiar nas respostas do *ChatGPT* para respostas verídicas. Mas, nessa mesma declaração, Altman manifesta que a empresa irá trabalhar para melhorar este aspecto do aplicativo. O que essa declaração significa para os futuros trabalhos da empresa é incerto, porém reforça o desejo por parte dos usuários e da empresa de que modelos como o *ChatGPT* funcionem, de alguma forma, como capazes de dizer ou buscar a verdade.

Antes de tudo, é necessário lembrar que modelos de linguagem são modelos de predição de palavras com base em seu contexto. Dessa forma, tais modelos não foram pensados para responderem baseados na verdade, nem recebem dados que os informem para tal. O aumento da base de treinamento não seria uma solução para o problema da verdade, pois o problema não é o volume de dados para treinamento e sim quais são objetivos do modelo. Quando LLMs conseguem completar tarefas de entendimento de linguagem ou de passar em provas de admissão, como a OAB [Romani 2023], isso se deve mais ao bruto volume de dados que tais modelos têm, e suas capacidades de generalização da linguagem humana, do que por algum conhecimento adquirido. Particularmente, o modelo de linguagem de tipo *InstructGPT* suscitou a esperança de que tal objetivo seja possível dada a fidelidade de suas respostas ao que usuário esperava, mas é para isso que ele foi desenvolvido. Porém, uma resposta mais alinhada à expectativa do usuário não equivale a uma resposta verdadeira. Acreditar que melhores respostas equivalham a respostas verdadeiras apenas ampliaria problemas de viés de confirmação, em que a validade da resposta se dá não pela sua factualidade, mas sim pelo alinhamento à ideologia ou valores do ator em questão.

Assim sendo, é imprescindível que se contextualize os avanços desses modelos e se compreenda suas capacidades e limitações. Também é essencial buscar pelo desenvolvimento de modelos voltados aos objetivos desejados, como modelos de entendimento de linguagem, que não dependam necessariamente de modelos massivos e força bruta de dados para atingir seu propósito.

4. Conclusão e possíveis direcionamentos futuros

O *ChatGPT* é uma aplicação de um modelo de linguagem conversacional que atingiu um extremo nível de popularidade, que apresenta uma impressionante capacidade de generalização, gera respostas mais naturais aos seus usuários humanos e é acessível ao público não especializado. Desde seu lançamento, novas colaborações e investimentos foram anunciados e novas aplicações vêm sido apresentadas em diversos campos. Neste trabalho, buscamos trazer considerações éticas e metodológicas quanto ao desenvolvimento de modelos de linguagem como esse, tanto riscos já conhecidos e documentados dentro da literatura de LLMs, quanto novas perspectivas envolvendo desenvolvimentos mais recentes dessa tecnologia.

Problemas anteriores persistem nesse novo modelo. O tamanho massivo do modelo pré-treinado e de seu próprio treinamento mantém as preocupações quanto ao risco ambiental, ao risco de dívida documental, à privacidade digital e à ampliação de vieses e preconceitos. Além de preocupações quanto aos impactos que modelos de linguagem têm

ou podem ter na sociedade, há também preocupações metodológicas envolvidas em seu desenvolvimento. Novas técnicas e modelos de linguagem vêm sendo desenvolvidos, mas são dependentes de modelos pré-treinados e de bases de dados massivas e crescentes. Em vez de somar mais dados a bases já massivas pela otimização de resultados em avaliações padronizadas ou de aperfeiçoar os modelos a novos casos de uso, deveríamos incentivar o estudo de um maior entendimento de como esses modelos capturam e replicam linguagem e o desenvolvimento de modelos voltados ao entendimento de linguagem.

Referências

Abid, A., Farooqi, M., and Zou, J. (2021). Persistent anti-muslim bias in large language models. *CoRR*, abs/2101.05783.

Adorno, J. (2023). Chatgpt in microsoft bing goes off the rails, spews depressive nonsense. *BGR*. Disponível em: `https://bgr.com/tech/chatgpt-in-microsoft-bing-goes-off-the-rails-spews-depressive-nonsense/`.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Bender, E. M. and Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P. W., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensold, J., O'Keefe, C., Koren, M., Ryffel, T., Rubinovitz, J., Besiroglu, T., Carugati, F., Clark, J., Eckersley, P., de Haas, S., Johnson, M., Laurie, B., Ingerman, A., Krawczuk, I., Askell, A., Cammarota, R., Lohn, A., Krueger, D., Stix, C., Henderson, P., Graham, L., Prunkl, C., Martin, B., Seger, E., Zilberman, N., hÉigeartaigh, S. , Kroeger, F., Sastry, G., Kagan, R., Weller, A., Tse, B., Barnes, E., Dafoe, A., Scharre, P., Herbert-Voss, A., Rasser, M., Sodhani, S., Flynn, C., Gilbert, T. K., Dyer, L., Khan, S., Bengio, Y., and Anderljung, M. (2020). Toward trustworthy ai development: Mechanisms for supporting verifiable claims. Technical report, Misc. Disponível em: `https://arxiv.org/abs/2004.07213`.

Milmo, D. (2023). Chatgpt reaches 100 million users two months after launch. *The Guardian*. Disponível em: `https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app`.

Olson, E. (2023). Google shares drop $100 billion after its new ai chatbot makes a mistake. *NPR*. Disponível em: `https://www.npr.org/2023/02/09/1155650909/google-chatbot--error-bard-shares`.

OpenAI (2022). Chatgpt: Optimizing language models for dialogue. *OpenAI*. Disponível em: `https://openai.com/blog/chatgpt/`.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback. Technical report, OpenAI. Disponível em: https://arxiv.org/abs/2203.02155.

Perrigo, B. (2023). Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. *Time*. Disponível em: https://time.com/6247678/openai-chatgpt-kenya-workers/.

Romani, B. (2023). Chatgpt é 'aprovado' em prova da primeira fase da oab. *ESTADÃO*. Disponível em: https://www.estadao.com.br/link/cultura-digital/chatgpt-e-aprovado-em-prova-da-primeira-fase-da-oab/.