Vieses em Histórias Ficcionais no ChatGPT

Thiago M. R. Ribeiro, Sean W. M. Siqueira, Maira G. de Bayser[1]

[1]Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
Rio de Janeiro – RJ – Brasil

thiago.m.ribeiro@edu.unirio.br, sean@uniriotec.br, mgdebayser@uniriotec.br

Abstract. *Since 2022, the use of chatbots—such ChatGPT, Gemini, and Copilot—has increased, allowing for more complex activities like text and code translation and modification. However, research indicates that these generative AI models are biased, particularly in social and cultural ways, reinforcing prejudicial associations with minority groups and gender stereotypes. By using rapid engineering to create story synopses, the experiment conducted in this work assessed ChatGPT's capacity to detect biases. The results, which were examined by experts, indicated that more complex prompts slightly improved bias detection. The need for AI solutions that generate less biased and more diversified material is highlighted by this study.*

Resumo. *O uso de chatbots, como ChatGPT, Gemini e Copilot, cresceu desde 2022, possibilitando tarefas avançadas, como traduções e revisões de textos e códigos. No entanto, estudos revelam que esses modelos de IA generativa apresentam vieses, especialmente sociais e culturais, que reforçam estereótipos de gênero e associações preconceituosas a grupos minoritários. O experimento realizado neste estudo avaliou a capacidade de identificação de vieses no ChatGPT, criando sinopses de histórias para identificar vieses sociais por meio de engenharia de prompts. Especialistas analisaram os resultados, que mostraram uma leve melhoria na detecção de vieses com prompts mais elaborados. Este estudo destaca a necessidade de soluções de IA que produzam conteúdos mais diversos e menos enviesados.*

1. Introdução

O uso de chatbots tem se integrado ao cotidiano de diversas áreas, desempenhando funções de assistentes virtuais, como a Alexa[1], que colaboram em tarefas simples, como tocar músicas, agendar compromissos e informar a previsão do tempo, segundo [Shawar and Atwell 2007]. Um marco significativo nesse campo foi o lançamento do ChatGPT em novembro de 2022, que transformou a interação entre usuários e ferramentas conversacionais, ampliando as possibilidades na Interação Humano-Computador. Essa tecnologia permite executar diversas tarefas, como resumir textos, traduzir idiomas e gerar conteúdo original, de acordo com [Abdullah et al. 2022].

É necessário entender se na literatura há o entendimento de como os vieses impactam as ferramentas conversacionais de IA generativa e se, uma vez que esses vieses são identificados, se essas próprias ferramentas possuem condições de identificarem esses vieses nos seus resultados, para que possam ser corrigidos e/ou mitigados, para auxiliar

[1]https://www.alexa.com/

os usuários em correções e amenizar este tipo de problema. Isto posto, há a necessidade de se experimentar como essa identificação pode ser realizada em um contexto específico.

O objetivo principal deste estudo foi o de entender como os vieses podem impactar as ferramentas conversacionais de IA generativa e de propor o desenvolvimento de um processo metodológico para avaliar se o ChatGPT consegue identificar vieses em uma situação onde a própria ferramenta gerou um resultado para a construção de uma sinopse de história. Para responder a estes questionamentos, as seguintes questões de pesquisa foram elaboradas:
QPP - Histórias geradas por ferramentas conversacionais de IA generativa contêm vieses que podem ser identificados pelas próprias ferramentas?
QP01 - Como os vieses impactam do desenvolvimento até a utilização das ferramentas conversacionais de IA generativa?
QP01.SQ01 - Como os vieses afetam o processo de desenvolvimento de software?
QP01.SQ02 - Quais os impactos dos vieses na utilização de ferramentas conversacionais/chatbots?
QP01.SQ03 - Como as intenções dos usuários são identificadas em ferramentas conversacionais/chatbots?
QP02 - Quais as possibilidades e limitações do ChatGPT na criação de histórias ficcionais e na identificação de vieses nos conteúdos que ele próprio gera?
QP02.SQ01 - Quais tipos de vieses podem ser observados nos conteúdos ficcionais gerados pelo ChatGPT, e em que circunstâncias esses vieses tendem a emergir?
QP02.SQ02 - Como a aplicação de técnicas de engenharia de prompts se diferencia na corretude da identificação de vieses pelo ChatGPT em comparação com prompts que não utilizam essas técnicas?

2. Trabalhos Relacionados

A Tabela 1 apresenta estes trabalhos com um comparativo do escopo abordado com relação a este trabalho.

Tabela 1. Trabalhos Relacionados

Escopo	Estudos				
	Este Trabalho (2024)	Lima, E. S. et al (2023)	Ferrara, E. (2023)	Mirowski, P. et al (2023)	Taveekitworachai, P. et al (2023)
ChatGPT	X	X	X		X
LLM				X	
Estruturas Narrativas	X	X		X	X
Vieses	X		X		

Sobre os itens definidos no escopo, o ChatGPT refere-se ao uso desta ferramenta no estudo citado, seja através do desenvolvimento de alguma solução em conjunto, ou como usuário final, apenas para utilização. LLM refere-se ao uso, estudo e/ou desenvolvimento de uma solução baseada em LLMs. Estruturas narrativas envolvem a criação e/ou leitura de histórias, sinopses ou roteiros ficcionais completos nos estudos. Vieses refere-se ao estudo ou análise de vieses, sejam eles cognitivos ou sociais.

[De Lima et al. 2023] apresentaram o ChatGeppetto, uma ferramenta que utiliza o ChatGPT como o componente do agente de IA da solução proposta, visando a produção de narrativas em linguagem natural, onde novas narrativas são geradas a partir de narrativas já existentes, se baseando em uma abordagem de relações semióticas, que considera quatro maneiras diferentes de compor novas narrativas. O estudo de [de Lima et al. 2016], que serviu de base para a aplicação deste trabalho relacionado, propôs o estudo de quatro tipos de relações semióticas : sintagmática, paradigmática, meronímica e antitética. Segundo [Chandler 2022], enquanto relações sintagmáticas são possibilidades de combinações (ex., em uma frase, as palavras são combinadas de maneira específica para formar um sentido completo), relações paradigmáticas são contrastes funcionais, ou seja, envolve a substituição de um signo por outro que pertence à mesma categoria, como por exemplo, na frase ”eu comprei um computador novo”, pode-se substituir as palavras que pertencem a mesma categoria gramatical e termos o seguinte resultado: ”ela vendeu uma televisão velha”. [de Lima et al. 2016] apontam que a relação meronímica, refere-se a parte-todo, onde um signo como uma ”peça”, é uma parte do todo ”computador”. Isto permite uma melhor entendimento em termos de hierarquia e estrutura. já a relação antitética mostra uma ideia de oposição, onde existe uma negação mútua, como por exemplo, ”vida” e ”morte”.

[Mirowski et al. 2023] apresentaram uma ferramenta interativa de co-escrita baseada em LLM que permite que roteiros para o cinema e teatro sejam criados a partir de uma linha de registro (*log line*) fornecida. Os resultados gerados pela ferramenta, escritos em conjunto com escritores/roteiristas, foram analisados por profissionais do setor teatral e cinematográfico, através de entrevistas qualitativas e pesquisa. Esses resultados abriram questões sobre a natureza da cocriatividade e sobre ética nos LLMs.

[Taveekitworachai et al. 2023] avaliaram os vieses ocorridos nos finais das histórias geradas pelo ChatGPT, usando dois *prompts*, um padrão e outro explícito, determinando o tipo de final a ser escrito. Os resultados apontaram uma tendência a finais positivos vindos do *prompt* padrão, indicando que mesmo quando instruído a gerar finais neutros, o ChatGPT tendeu a gerar finais considerados positivos, sugerindo a existência de um viés inerente ao processo de criação do modelo. O estudo sugere que a abordagem desses vieses é crucial para garantir o ChatGPT e modelos similares sigam as normas sociais para evitar o reforço ou ampliação de vieses existentes.

[Ferrara 2023] descreveu sobre modelos de linguagem generativa, como o ChatGPT, e apresenta uma visão geral sobre os riscos e desafios associados a vieses nesses modelos. São explorados os fatores que originam os vieses, como dados de treinamento, restrições algorítmicas, decisões políticas, design e especificação de modelos. Também são analisadas possíveis formas de mitigação em modelos de linguagem, tendo em vista que o surgimento de alguns destes vieses é inevitável.

Além dos estudos citados, o estudo de Cho, T., entitulado *A Study on Dramaturgy for AI Screenplays: Writing Alternative Narratives Using GPT* propunha explorar a escrita de roteiros criativos de ficção utilizando o ChatGPT e a colaboração entre humanos e IA. No entanto, não foi possível obter a versão completa do estudo até a finalização deste trabalho.

Diferente dos trabalhos relacionados, este trabalho trata de realizar um estudo

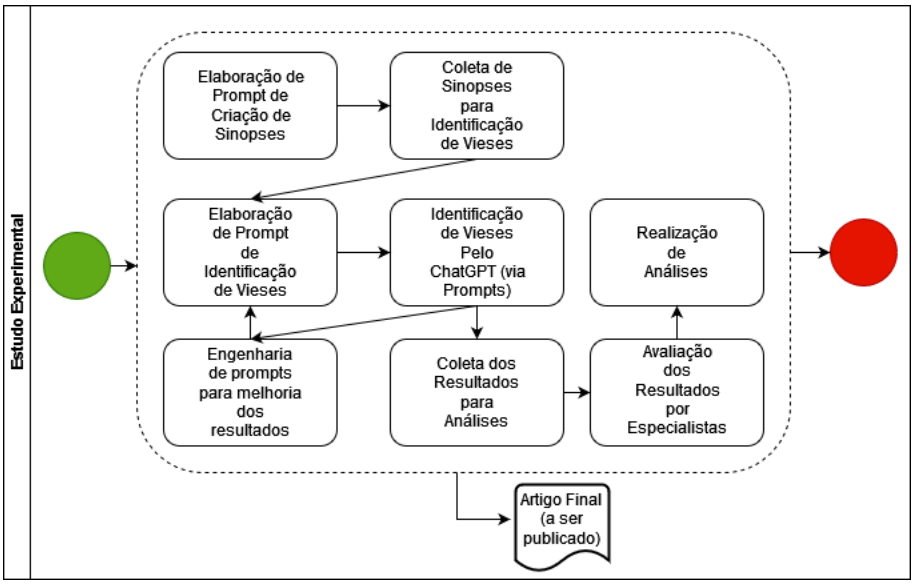
experimental através da verificação da capacidade do ChatGPT de gerar sinopses de histórias ficcionais e também da própria ferramenta em conseguir identificar vieses nesses conteúdos gerados, pelo próprio *prompt* da ferramenta, ao utilizar diferentes abordagens para a análise e comparação desses resultados.

3. Estudo Experimental

Esta seção apresenta o processo metodológico, que visou entender e tratar, de forma estruturada e sistematizada, como o ChatGPT identificou vieses provenientes das respostas dadas pela própria ferramenta, em um processo similar ao LLM as a Judge [2] [3], em um contexto de criação de uma sinopse de história/roteiro ficcional criado neste processo.

Com o objetivo de responder as questões de pesquisa, foi elaborado, com o auxílio de um especialista (roteirista/escritor), um pedido de geração de uma sinopse, considerando os elementos essenciais para a criação de uma sinopse/história. Para isto, foi criado um *prompt* para que o ChatGPT pudesse gerar este conteúdo. Esse texto resultante foi avaliado por um modelo que utiliza o ChatGPT como base, na intenção de que este modelo conseguisse identificar vieses nas sinopses geradas. O processo metodológico proposto está esquematizado na Figura 1.

Figura 1. Etapas do Processo Metodológico



Fonte: elaborado pelo autor

3.1. Estudo Realizado

O estudo realizado consiste nas etapas de experimento e análise de dados, com elaboração e coleta de sinopses; elaboração e identificação de vieses; e de coleta, avaliação e análise dos dados. Esta etapa experimental do estudo consiste em oito estágios interconectados: i) elaboração de *prompt* para criação de sinopses; ii) coleta de sinopses para identificação de vieses; iii) elaboração de *prompt* de identificação de vieses; iv) identificação de vieses

[2] https://docs.smith.langchain.com/concepts/evaluation?ref=blog.langchain.dev#llm-as-judge
[3] https://blog.langchain.dev/aligning-llm-as-a-judge-with-human-preferences/

pelo ChatGPT (via *prompts*); v) engenharia de *prompts* para melhoria dos resultados; vi) coleta dos resultados para análises; vii) avaliação dos resultados por especialistas e viii) realização de análises. Como resultado, foram desenvolvidos três *prompts*: um para criação de sinopses e dois para identificação de vieses (um com técnicas de engenharia de *prompts* e outro sem). Os resultados foram validados por especialistas para atestar a confiabilidade.

3.2. Elaboração de Prompt de Criação de Sinopses

Com o auxílio de dois especialistas, ambos com mais de 5 anos de experiência na área de escrita de roteiros ficcionais, foram definidos os elementos considerados essenciais (nas suas concepções) para a criação de uma sinopse de história. Estas informações foram coletadas por meio de entrevistas informais onde perguntas acerca do tema foram realizadas. Após a sumarização das respostas dadas por ambos os roteiristas, foi possível avaliar os elementos necessários para a criação da sinopse. Através destas informações, foi definido o seguinte conjunto de elementos: i) a temática da história, ii) o universo onde essa história está inserida, iii) as premissas, iv) a história, v) os personagens e seus arcos narrativos; vi) o gênero narrativo.

Com os elementos definidos, foram dadas as instruções para que o ChatGPT retornasse uma sinopse com até 500 palavras. O idioma escolhido para a geração deste *prompt* foi o inglês, por conta da maior abrangência deste idioma na utilização da ferramenta e nos dados utilizados nos seus modelos de treinamento. Como resultado dessa etapa, um *prompt* foi criado.

3.3. Coleta de sinopses para avaliação de vieses

Após o *prompt* de geração de sinopses definido na etapa anterior, foi desenvolvido um script em Python, usando o Google Colab[4] e a API da OpenAI[5], para gerar e armazenar sinopses com base em um prompt previamente definido. O modelo GPT-3.5-turbo e o módulo ChatCompletion foram utilizados para executar o prompt 100 vezes, gerando sinopses com variações temáticas, devido à relevância do tema escolhido. Também foi solicitado que as características dos personagens fossem definidas para analisar as sugestões e variações geradas pelo ChatGPT. Após testes com diferentes parâmetros do modelo, definiu-se a versão final usada no estudo, cujos parâmetros estão apresentados na Tabela 2.

Tabela 2. Parâmetros do modelo de prompt do ChatGPT para criação de sinopses

Parâmetro	Valor Final	Valor Padrão
model	gpt-3.5-turbo	gpt-3.5-turbo
temperature	0.7	1
max_tokens	500	4096
top_p	1	1
frequency_penalty	1	0
presence_penalty	0.5	0

Parâmetro model: Define o tipo de modelo LLM a ser utilizado; temperature: Ajusta a aleatoriedade das respostas. Valores maiores tornam as respostas mais

[4]https://colab.research.google.com/

[5]https://openai.com/index/openai-api/

randômicas, menores as tornam mais determinísticas; max_tokens: Limita o tamanho máximo das respostas, controlando a quantidade de tokens; top_p: parâmetro relacionado ao temperature, regula a diversidade das respostas. Valor 0 gera respostas determinísticas e 1, totalmente aleatórias; frequency_penalty: Controla a repetitividade. Valores altos promovem novas estruturas, mas não novos tópicos, e valores baixos geram respostas mais repetitivas; presence_penalty: Influi na evasão de tópicos; valores altos aumentam a probabilidade de evitar certos temas, enquanto valores baixos tornam o modelo mais focado em pontos específicos.

3.4. Elaboração de Prompt de Identificação de Vieses

A próxima etapa consistiu na elaboração de um *prompt* que deveria realizar a identificação dos vieses nas sinopses. Para isso, esse *prompt* precisaria ler as sinopses que foram criadas e armazenadas pelo *prompt* desenvolvido no item 3.3. Foram considerados para a identificação os seguintes tipos de vieses sociais: viés racial, viés religioso, viés político, viés de gênero e viés de orientação sexual. Para os casos em que a identificação não conseguisse definir algum dos vieses solicitados, o tipo de viés deveria ser definido como inconclusivo. A escolha por estes tipos de vieses sociais se deveu ao levantamento realizado pela *Rapid Review* em [Ribeiro et al. 2024], onde estes tipos de vieses emergiram dos resultados encontrados.

3.5. Identificação de Vieses Pelo ChatGPT: via prompts

Com o *prompt* definido na etapa 3.4, foi criado um *script* para realizar a identificação dos vieses e coleta destes resultados. Este *script* também foi criado com base no modelo GPT-3.5-turbo, em linguagem python e utilização do módulo ChatCompletion e das mesmas bibliotecas utilizadas no *prompt* de criação de sinopses. Foram necessárias que as sinopses criadas e armazenadas na etapa 3.3 fossem fornecidas para que este modelo de *prompt* pudesse identificar os vieses. Foram criadas duas versões desse *prompt*, onde uma dessas versões utilizou de técnicas de engenharia de *prompts* e na outra não foram utilizadas essas técnicas. Para a versão básica do *prompt* deste estudo, sem as aplicações das técnicas de engenharia de *prompts*, foram realizados testes que consistiram em alterar parâmetros do modelo e analisar as respostas dadas pelo *prompt* até chegar em um resultado considerado razoável. Com estas alterações, chegou-se na versão do modelo apresentada na Tabela 3.

Tabela 3. Parâmetros do modelo de prompt do ChatGPT para identificação de vieses

Parâmetro	Valor	Valor Padrão
model	gpt-3.5-turbo	gpt-3.5-turbo
temperature	0.8	1
max_tokens	400	4096
top_p	1	1
frequency_penalty	1	0
presence_penalty	0.8	0

3.6. Engenharia de Prompts para melhoria dos resultados

A partir do *prompt* definido no item 3.5, algumas técnicas/táticas de *prompt engineering* foram utilizadas em busca melhorar os resultados iniciais obtidos. Uma das táticas imple-

mentadas foi a **utilização de uma persona específica** para o modelo em suas respostas. Entendeu-se que, apesar de se tratar de um tipo de técnica de engenharia de *prompts*, ela foi implementada para ambas as versões de *prompt* de identificação de vieses. Isto foi importante para fornecer ao modelo a instrução de como ele deveria responder ao avaliar diferentes tipos de vieses, como um especialista.

As outras táticas de engenharia de *prompts* a seguir foram utilizadas somente para um dos *prompts*. A técnica chamada *few-shot prompting*, que **fornece exemplos** ao modelo para que ele possa responder de forma orientada com os exemplos dados, foi utilizada. Outra tática utilizada, chamada de Chain-of-thought Prompting, foi modificado o pedido no *prompt* para que o objetivo fosse **executado em passos (*steps*) sequenciais**, conforme detalhado no estudo de [Wei et al. 2022], de maneira que o modelo entendesse de forma mais fácil o que precisava fazer. Também foram fornecidos exemplos de saída que o modelo deveria retornar. O exemplo definido que foi utilizado para a saída do *prompt* veio de execuções anteriores deste modelo e através do ajuste dos parâmetros, para se chegar a versão final do *prompt* utilizada neste estudo.

Todos os *scripts* desenvolvidos para este estudo, estão disponibilizados no github[6].

3.7. Coleta dos Resultados para Análise

Os dados de cada processo de identificação de vieses foram coletados e armazenados para análises. Um dataset com 100 sinopses geradas pelo *prompt* foi criado, incluindo definições das sinopses e características dos personagens. Essas sinopses foram submetidas a dois modelos de *prompt*: um sem uso de técnicas de *prompt engineering* (SPE), descrito no item 3.5, e outro com técnicas de *prompt engineering* (CPE), detalhado no item 3.6. Ambos os modelos identificaram vieses nas sinopses nos seguintes tipos: racial, gênero, religioso, político, orientação sexual e inconclusivo. Foram pedidas as descrições de trechos ou frases contendo o viés, com a justificativa da identificação fornecida pelo ChatGPT. As identificações resultaram em dois datasets correspondentes aos modelos utilizados.

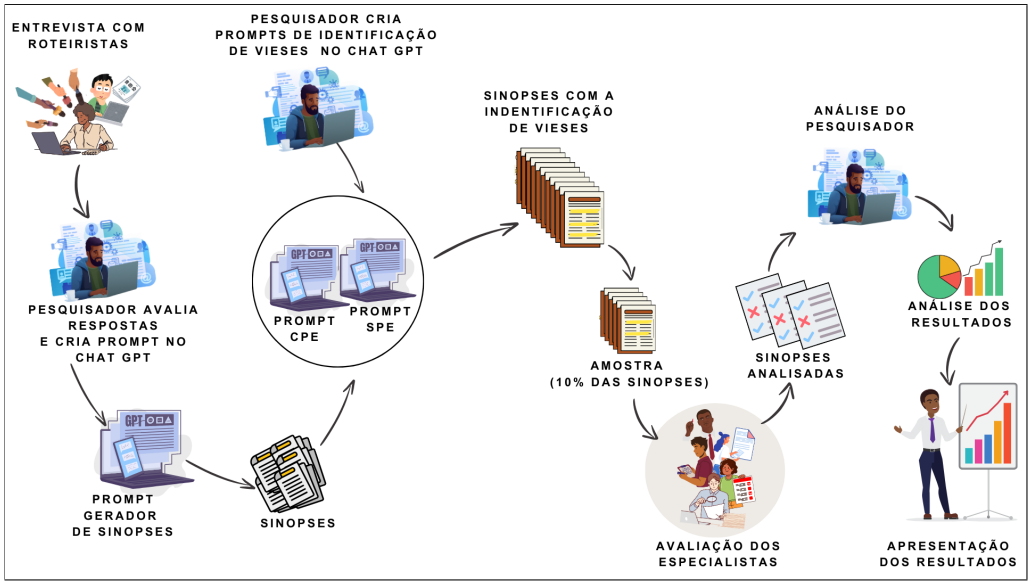
3.8. Avaliação dos Resultados por Especialistas

Além de uma análise quantitativa entre as versões de *prompts* de identificação de vieses, também foi necessário verificar se os especialistas concordavam ou não com essas identificações dadas por estes *prompts* (SPE e CPE). Participaram desta etapa, três pesquisadores com mais de 10 anos de experiência na área de tecnologia e estudos sobre vieses. Este número supera o mínimo sugerido de dois avaliadores para o cálculo da confiabilidade, descrito no estudo de [Krippendorff 2011]. A Figura 2 ilustra todo o experimento, passando pelo processo de avaliação dos especialistas.

Com o intuito de avaliar as identificações realizadas pelos dois *prompts*, especialistas analisaram as sinopses e frases dessas sinopses com os vieses destacados. Cada especialista recebeu uma amostra de 10% das sinopses e frases, visualizando o tipo de viés identificado sem saber qual *prompt* o identificou, marcando se concordava ou não com a identificação. Para garantir o entendimento dos tipos de vieses analisados, foram fornecidas aos especialistas definições baseadas na literatura.

[6]https://github.com/thiribeiro/chatgpt-bias-eval

Figura 2. Etapas do Experimento



Fonte: elaborado pelo autor

Na etapa de análise dos dados, os resultados dos *prompts* com as avaliações dos especialistas foram comparados. Para medir a coerência entre as avaliações, utilizou-se a métrica de intercoder reliability (ICR), que avalia o grau de concordância entre diferentes codificadores quanto à classificação dos mesmos dados. São calculados os valores de confiabilidade entre os avaliadores participantes através do alfa de Krippendorf, conforme apontado no estudo de [Krippendorff 2011], para entender se as comparações com os modelos de *prompt* apresentam indicações de vieses condizentes com a avaliação dos especialistas. Participaram deste processo três pesquisadores com mais de 10 anos de experiência na área de tecnologia e estudos sobre vieses.

3.9. Realização de Análises

Esta etapa consistiu em realizar as análises qualitativa e quantitativa dos dados, correspondentes aos resultados das identificações dos *prompts* gerados. Esta comparação entre os modelos, juntamente com as avaliações realizadas pelos especialistas no item 3.8, permitiu responder as questões de pesquisa que foram definidas para este estudo e que são abordadas na seção 4.

4. Resultados

Esta seção apresenta uma análise quantitativa e explanatória referente aos resultados encontrados para os modelos de *prompt* SPE e CPE. Com estes achados, foi possível responder a questão de pesquisa 02 e as suas subquestões, destacadas nos itens a seguir.

4.1. Quais tipos de vieses podem ser observados nos conteúdos ficcionais gerados pelo ChatGPT, e em que circunstâncias esses vieses tendem a emergir?

Foram identificados nos resultados vieses que foram solicitados ou não para o estudo. Para os tipos de vieses solicitados na identificação, os resultados apresentaram um total de 418 vieses identificados para o *prompt* CPE e 175 vieses para o *prompt* SPE. Em uma

única sinopse, foi possível para o *prompt* CPE identificar diferentes tipos de vieses e/ou quantidades diferentes de um mesmo tipo de viés, como por exemplo, em situações com mais de uma ocorrência de viés do tipo racial, de gênero, entre outros. Foi obtida uma média de 4,25 identificações de vieses por sinopse avaliada neste *prompt* CPE. Os vieses que não foram definidos para a identificação neste estudo, também foram encontrados em ambos os *prompts*, como o viés relacionado a idade, viés social, viés étnico, viés sócio-econômico, entre outros.

O *prompt* SPE também foi capaz de identificar mais de um tipo de viés nas sinopses, com uma média de 2,54 identificações de vieses por sinopse. Para os tipos de vieses que não foram solicitados na identificação, o *prompt* CPE obteve um total de 7 identificações, distribuídos em 4 tipos de vieses diferentes (idade, sócio-econômico, étnico e orientado ao social) e de 79 identificações em 17 tipos diferentes de vieses para o *prompt* SPE.

4.2. Como a aplicação de técnicas de engenharia de prompts se diferencia na corretude da identificação de vieses pelo ChatGPT em comparação com prompts que não utilizam essas técnicas?

Os *prompts* SPE e CPE foram analisados quanto à capacidade de identificar vieses nas sinopses. O *prompt* SPE conseguiu identificar vieses solicitados, com predominância do viés racial, além de identificar vieses não solicitados, como o viés profissional. Houve uma concordância média de 56% entre o SPE e os avaliadores, sendo o viés racial o mais concordante e também o mais discordante. Discordâncias representaram 44% das identificações, principalmente nos vieses racial e de gênero. A confiabilidade média da concordância entre os avaliadores e o SPE foi considerada justa, com um alfa de Krippendorff médio de 31,12%.

Por outro lado, o *prompt* CPE apresentou maior capacidade de identificar vieses solicitados e frases com vieses em uma mesma sinopse, além de identificar tipos de vieses não solicitados em menor quantidade que o SPE. Isso pode ser atribuído às definições mais detalhadas dos vieses e ao uso de técnicas aprimoradas de engenharia de *prompts*. A concordância média entre os avaliadores e o CPE foi também considerada justa, com um alfa médio de 40,78%, sendo o viés político o mais concordante, seguido do racial. Assim, o CPE demonstrou maior adequação à proposta do estudo.

5. Discussão

Esta seção apresenta uma discussão sobre os achados provenientes do levantamento de literatura e do experimento realizado, referente as análises dos dados coletados na identificação dos vieses nos dois *prompts* e da avaliação realizada por especialistas para estes resultados. Foi possível assim, responder as questões de pesquisa elaboradas para o estudo, apontadas na Seção 1.

5.1. Achados do levantamento de literatura

O objetivo do levantamento de literatura foi compreender o impacto dos vieses do desenvolvimento, no contexto do processo de desenvolvimento de software (PDS) e no uso de ferramentas conversacionais. Inicialmente, o foco foi nos vieses cognitivos que afetam os desenvolvedores e podem influenciar o produto final. Porém, com o crescente uso de

ferramentas como o ChatGPT e a preocupação com seus vieses, o foco passou da visão do desenvolvedor para a perspectiva do usuário, buscando analisar os impactos sociais dos vieses nesse tipo de tecnologia.

Os estudos revelaram que os vieses sociais predominam em relação aos cognitivos no uso de ferramentas conversacionais. Foi necessário investigar como essas ferramentas identificam intenções dos usuários, já que essa etapa pode reproduzir vieses por meio do processamento de linguagem natural. Com as técnicas compreendidas, um experimento foi realizado para avaliar se o ChatGPT é capaz de gerar conteúdo enviesado e também identificar vieses presentes utilizando suas próprias capacidades.

5.2. Achados do experimento

A análise mostrou que o *prompt* criado sem técnicas de engenharia (SPE) identificou mais tipos diferentes de vieses em comparação ao *prompt* elaborado com técnicas (CPE). O SPE detectou vieses sociais e também cognitivos, como viés de confirmação, efeito halo e viés de autoridade. Ambos os *prompts* identificaram vieses étnicos e etários, sendo o primeiro relacionado a aspectos culturais, como nacionalidade e religião, enquanto o racial se refere a características fenotípicas, como cor da pele. O SPE também apontou viés profissional, não previsto no estudo, possivelmente devido ao escopo reduzido das sinopses.

O CPE demonstrou maior precisão na identificação dos vieses definidos pelo estudo, ao receber orientações e significados detalhados, identificando-os em maior número e com maior concordância entre especialistas (40,78%) em relação ao SPE (31,12%). Ambos os *prompts*, porém, apresentaram limitações na detecção de múltiplos vieses em uma mesma frase. Situações de marcação incompleta ou subjetiva exigiram maior esforço dos especialistas na análise.

6. Conclusão

É importante ressaltar que a contribuição inicialmente percebida neste estudo foram, além do levantamento de literatura, que buscou mapear sobre os impactos dos vieses, desde o desenvolvimento de ferramentas conversacionais até a sua utilização pelos usuários, também foi percebida a contribuição no que tange a indicação de um caminho para a autoavaliação em conteúdo gerado por ferramentas de IA generativa, neste caso, o ChatGPT. Para isto, este estudo apresentou um experimento gerado a partir da criação de sinopses de histórias ou narrativas fictícias e da identificação de vieses sociais nessas mesmas sinopses.

Todas as etapas foram realizadas através do ChatGPT, com o objetivo de responder se a ferramenta era capaz de identificar possíveis problemas decorrentes de vieses em seu próprio conteúdo gerado. Também podemos considerar outras contribuições, como a discussão dos resultados do experimento, a metodização do uso do ChatGPT na criação de sinopses/histórias e também a sua utilização na tentativa de identificar vieses nesse mesmo conteúdo gerado, além do uso de técnicas de engenharia de *prompts* para a melhoria dos resultados e a validação da identificação dos vieses com especialistas, com utilização de *intercoder reliability* para atestar a confiabilidade dos resultados.

Referências

Abdullah, M., Madain, A., and Jararweh, Y. (2022). Chatgpt: Fundamentals, applications and social impacts. In *2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 1–8. IEEE.

Chandler, D. (2022). *Semiotics: the basics*. Routledge.

de Lima, E. S., Feijó, B., Casanova, M. A., and Furtado, A. L. (2016). Storytelling variants based on semiotic relations. *Entertainment Computing*, 17:31–44.

De Lima, E. S., Feijó, B., Cassanova, M. A., and Furtado, A. L. (2023). Chatgeppetto-an ai-powered storyteller. In *Proceedings of the 22nd Brazilian Symposium on Games and Digital Entertainment*, pages 28–37.

Ferrara, E. (2023). Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*.

Krippendorff, K. (2011). Agreement and information in the reliability of coding. communication methods and measures, 5 (2), 93-112.

Mirowski, P., Mathewson, K. W., Pittman, J., and Evans, R. (2023). Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–34.

Ribeiro, T. M., Siqueira, S. W., and de Bayser, M. G. (2024). Revisao rápida sobre vieses em chatbots-uma análise sobre tipos de vieses, impactos e formas de lidar. *Anais do XIX Simpósio Brasileiro de Sistemas Colaborativos*, pages 56–70.

Shawar, B. A. and Atwell, E. (2007). Chatbots: are they really useful? *Journal for Language Technology and Computational Linguistics*, 22(1):29–49.

Taveekitworachai, P., Gursesli, M. C., Abdullah, F., Chen, S., Cala, F., Guazzini, A., Lanata, A., and Thawonmas, R. (2023). Journey of chatgpt from prompts to stories in games: the positive, the negative, and the neutral. In *2023 IEEE 13th International Conference on Consumer Electronics-Berlin (ICCE-Berlin)*, pages 202–203. IEEE.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.