Esquecida no Churrasco: Um método para validação e correção responsável de traduções automáticas do BBQ para Português do Brasil

Luiza Rodrigues Cardoso, Leo Sampaio Ferraz Ribeiro

[1]USP, Instituto de Ciências Matemáticas e Computação - São Carlos, SP – Brasil

luizarodriguescardoso@usp.br, leo.ribeiro@icmc.usp.br

Abstract. *Com os rápidos avanços em Processamento de Linguagem Natural, torna-se necessário o desenvolvimento de ferramentas de avaliação de performances dos modelos linguísticos diante da destilação de vieses sociais. Neste artigo, então, apresentamos um método de adaptação de datasets que contempla adequações de fatores linguísticos e socioculturais do Brasil, uma demonstração de aplicação do método ao dataset BBQ e a descrição da implementação da nossa plataforma autoral de validação.*

1. Introdução

Com o avanço de modelos de linguagem (LMs) e suas aplicações abrangentes, surge a preocupação acerca das consequências sociotécnicas, como a disseminação de vieses nos conteúdos gerados por estes modelos, como os utilizados em chatbots [Bonil et al. 2025].

O termo viés, em Aprendizado de Máquina, é usado para descrever os tipos de modelos que podem ser obtidos via diferentes algoritmos e as limitações deste espaço de busca. Neste artigo, contudo, usamos o termo sob o aspecto ético, em que viés descreve, então, os impactos dos estereótipos sociais sobre os dados, fatores humanos que, quando agregados aos dados, desbalanceiam os modelos nos treinamentos por meio da sub ou super-representação de conceitos [Bender et al. 2021]. Este tipo de aprendizado resulta em comportamentos reprodutores de visões hegemônicas [Messeri and Crockett 2024] e associações enviesadas sobre gênero, raça e outras características de identidade pessoal, bem como padrões sistêmicos de discriminação [Talat et al. 2022]. Esse problema pode ser visto especialmente em PLN, em que os dados de treinamento – *corpus* – são obtidos por meio de fóruns online ou obras literárias em domínio público e, portanto, refletem opiniões pessoais, incluindo estereótipos, que afetam os desempenhos dos modelos frente aos padrões de ética e moral.

Com a adoção generalizada destes modelos, são necessárias garantias tanto de desempenho quanto de segurança. Segundo [Neplenbroek et al. 2024], há uma diferença de desempenho de LLMs em *prompts* que utilizam idiomas não anglófonos, ocorrendo um agravamento da disseminação de vieses sociais. Possivelmente esse baixo desempenho decorre do fato de que o idioma dominante durante a fase de treinamento – chegando a 90% do total – e utilizado na etapa de treinamento de segurança é o inglês [OpenAI et al. 2024, Touvron et al. 2023, Brown et al. 2020].

O trabalho desenvolvido por [Parrish et al. 2022] traz uma notável contribuição na tarefa de avaliação da performance de LLMs frente aos vieses sociais, em que, percorrendo 9 categorias sociais baseadas no aspecto sociocultural dos EUA, e utilizando

templates, com ajustes de agentes e objetos das sentenças, foi possível identificar que os modelos reproduzem estereótipos de acordo com os parâmetros e nível de descrição dos agentes das sentenças.

Assim, surge a necessidade de investigar o comportamento destes modelos em outros contextos sociais e linguísticos, como os do Brasil. Neste trabalho, então, propomos (1) Uma plataforma para validação e submissão de *templates*; (2) Um método de validação de traduções diretas de datasets, em que, a partir da plataforma, o avaliador poderá aprovar a tradução direta e sugerir alterações; (3) Uma versão inicial do dataset BBQ em português brasileiro.

2. Trabalhos Relacionados

Como [Bender et al. 2021, Talat et al. 2022] demonstram em seus trabalhos, os modelos possuem a tendência de amplificar discursos prejudiciais da sociedade, sendo necessário o desenvolvimento de ferramentas que mitiguem esses comportamentos. Há iniciativas que buscam compreender e avaliar a presença de vieses relacionados ao gênero, identidade não-binária, religião, raça e temas relacionados a pessoas com deficiências [Ovalle et al. 2023, Zhao et al. 2021, Gadiraju et al. 2023], porém a maioria dos trabalhos é focada na língua inglesa.

Neste sentido diversos benchmarks de auditoria foram criados, como o CrowS-pairs [Nangia et al. 2020], StereoSet [Nadeem et al. 2020] e BBQ [Parrish et al. 2022]. Destacamos este último como um caso de estudo para esse trabalho; o BBQ foi criado para avaliar para modelos Question-Answering que lida com múltiplos aspectos sociais, baseando-se em situações que servem para estudar os níveis de vieses sociais, tendo como limitação os aspectos socioculturais dos Estados Unidos. O trabalho utiliza *templates*, com os parâmetros de agente e objetos adaptáveis, que permitem construir os cenários destes agentes e permutar diferentes estereótipos presentes na sociedade estudada. Diante disso, surgiram algumas variantes do BBQ, o Korean-BBQ [Jin et al. 2024], Chinese-BBQ [Huang and Xiong 2023], German Gender BBQ [Satheesh et al. 2025] e o Multilingual-BBQ [Neplenbroek et al. 2024]. Evidencia-se a adaptabilidade deste dataset para outros idiomas e contextos socioculturais. Assim, propomos uma versão brasileira com validação humana e adaptativa dos *templates* ao nosso contexto sociocultural.

Outros datasets que visam realizar este tipo de avaliação sob o aspecto regional são SHADES [Mitchell et al. 2025] e LACES [Ivetta et al. 2025]. Sendo o primeiro desenvolvido para avaliar estereótipos culturais e específicos presentes em modelos linguísticos sob 37 regiões do globo. E o segundo possui enfoque na região latino-americana e uma coleta participativa, ou seja, a comunidade possui participação na construção, gerando um conjunto realista. Ambos agregam à tarefa de avaliação estereótipos específicos e nuances culturais. Esses aspectos, em consonância, compõem a nossa proposta.

No Brasil, há uma crescente movimentação para estudar a proliferação dos vieses sob os aspectos do nosso idioma e cultura [Silva et al. 2024]. Alguns trabalhos se focam na identificação da questão de gênero na tarefa de tradução automática [Soares et al. 2023] e na redução de vieses sociais em modelos baseados na arquitetura CLIP, o FairPivara proposto por [Moreira et al. 2024]. Outro trabalho fundamental é o desenvolvido por [Bonil et al. 2025], que avalia as questões de gênero e raça em narrativas geradas pelos LLMs,

Esses trabalhos evidenciam os diferentes tipos de abordagem para a tarefa de avaliação de vieses e mostram a necessidade de se construir ferramentas apropriadas para a sociedade brasileira, fornecendo *insights* sobre a disseminação de vieses sociais sob o contexto sociocultural brasileiro, auxiliando em trabalhos e políticas públicas futuras.

3. Metologia

Dataset Bias Benchmark for Question-Answering (BBQ) Originalmente, o dataset foi escrito em inglês e desenvolvido para abranger dois cenários: o ambíguo, em que há pouca descrição dos agentes e suas ações, e o informativo, em que há uma descrição detalhada. Ambos são alinhados com perguntas de cunho negativo, que ressaltam um estereótipo, e não negativo, que complementa a primeira, tendo como objetivo avaliar a presença do viés sistêmico e específico, e se as respostas obtidas possuem alinhamento com algum viés. Além disso, o conjunto conta com as respostas esperadas para cada contexto-pergunta, em que, para o contexto ambíguo, a única resposta correta é a opção desconhecido ou variações relacionadas, dada a falta de informação do texto. Para o outro, a resposta correta é alguns dos agentes citados na sentença.

Este conjunto abrange as seguintes categorias: idade, status de deficiência, identidade de gênero, nacionalidade, aparência física, raça/etnia, religião, orientação sexual, status socioeconômico, e as intersecções: raça por gênero e raça por status socioeconômico, totalizando 441 *templates*. A tradução do dataset para o português brasileiro utilizou a API do Google Translate [1], resultando em um conjunto de dados em português. Contudo, para refletir as nuances da comunidade brasileira, optamos pela validação das traduções automáticas e, possivelmente, uma adaptação deste conjunto inicial.

Para isso, por meio da checagem humana, inspecionaremos quaisquer erros oriundos da assimetria sintática e semântica entre os idiomas, adaptações de termos que aproximem o conjunto à realidade dos aspectos sociais brasileiros, e criações de novos *templates*. Tais ações serão realizadas por meio da plataforma desenvolvida pelas pessoas autoras.

Plataforma de Validação A plataforma conta com as seções cadastro do avaliador, validação e sugestões de *templates*. A seção de criação do perfil do avaliador é composta por um formulário contendo perguntas de múltipla escolha que ajudarão na descrição da posição social do indivíduo, oferecendo como resultado um perfil socioeconômico. Os dados que serão coletados nesta etapa não serão armazenados no sistema, sendo utilizados apenas para traçar quais categorias são mais relevantes para o indivíduo, a partir de uma política de pesos e relevância de categorias para o indivíduo. Essa prática tem como finalidade ressaltar a opinião de quem é o foco da ação e normalizar a distribuição das categorias para cada indivíduo.

Além disso, a seção de sugestões de *templates* é uma oportunidade para enriquecer a base de dados do Br-BBQ.

Políticas de Peso A prática de políticas de peso tem como finalidade amplificar as vozes dos indivíduos que rotineiramente são silenciados na sociedade, mesmo sendo alvos de

[1] https://translate.google.com/

discursos estereotipados. Para isso, fez-se o uso de uma política de atribuição de pesos que bonifica identidades minoritárias.

Com as respostas do cadastro do avaliador, conseguimos formular um perfil socioeconômico. Por meio dessas informações, construímos um dicionário de características como faixa etária, identidade racial e de gênero, autodeclaração de status de deficiência e situação socioeconômica - renda familiar e formação estudantil. Com esse objeto, realizamos uma inspeção que avalia os grupos sociais e atribui pesos, entre o intervalo$[1, 4]$, para cada categoria que compõe o escopo do dataset, gerando assim um novo dicionário composto pelo par $\{\text{Categoria}, \text{Pesos}\}$. Sendo os maiores pesos dados para as características que representam identidades minoritárias da sociedade brasileira. A atribuição de pesos seguiu os seguintes critérios: Peso 4 (Nível Crítico), Mulheres (Cis/Trans), Pessoas Não Binárias, Pessoas Não Brancas e PcDs; Peso 3 (Nível Intersecção), Cruzamento de duas identidades (ex: Raça e Gênero) e Idosos Não Brancos; Peso 2 (Nível Contextual), Situação Socioeconômica Intermediária/Vulnerável e Idosos Brancos; Peso 1 (Nível Base), Demais categorias.

A partir dessas informações, realizamos uma normalização e distribuímos com probabilidades ponderadas os *templates* para os avaliadores que possuem mais chances de enfrentar os estereótipos descritos neles.

Avaliação A avaliação de modelos de linguagem fará uso das métricas já estabelecidas pelo benchmark original (BBQ), a acurácia nas respostas e o escore de viés [Parrish et al. 2022]. Os modelos serão avaliados através da ferramenta de [Shimabucoro et al. 2024] e serão analisados os modelos LLaMa2-7B, LLaMa2-13B [Touvron et al. 2023], Mixtral8x7B [Jiang et al. 2024], Aya-8B [Aryabumi et al. 2024],Gemma-7B [Team et al. 2024], Command-R+ [2], além do modelo Olmo2-7B [OLMo et al. 2025] e versões de pesos abertos do GPT.

Resultados preliminares Traduzimos de forma automática os *templates* do dataset original, resultando em um conjunto inicial com 441 *templates* sem validação humana ou adaptações. Além disso, desenvolvemos a plataforma de validação, que será utilizada para a coleta de validações e criação de *templates*; esta coleta foi aprovada por Comitê de Ética em Pesquisa (CEP), com CAAE 96621926.1.0000.5390.

4. Conclusão e Trabalhos Futuros

Apresentamos um método de validação de tradução direta de dataset, em especial, uma tradução para o português do dataset BBQ, que resultou em uma versão inicial, que será validada na próxima etapa do projeto, do Brazilian BBQ(Br-BBQ). Além disso, apresentamos uma plataforma de avaliação humana e criação de *templates* que serão utilizadas para avaliar o desempenho de LLMs em cenários referentes ao contexto sociocultural do Brasil. Em trabalhos futuros, esperamos apresentar a versão final do Br-BBQ e uma avaliação qualitativa sobre a presença de vieses sociais modelos sob o contexto brasileiro.

[2]https://docs.cohere.com/docs/command-r-plus

Referências

Aryabumi, V. et al. (2024). Aya 23: Open weight releases to further multilingual progress. arXiv:2405.15032.

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 *FAccT*, pages 610–623.

Bonil, G., Hashiguti, S., Silva, J., Gondim, J., Maia, H., Silva, N., Pedrini, H., and Avila, S. (2025). Yet another algorithmic bias: A discursive analysis of large language models reinforcing dominant discourses on gender and race. arXiv:2508.10304.

Brown, T. B. et al. (2020). Language models are few-shot learners. arXiv:2005.14165.

Gadiraju, V., Kane, S., Dev, S., Taylor, A., Wang, D., Denton, R., and Brewer, R. (2023). "i wouldn't say offensive but...": Disability-centered perspectives on large language models. In *FAccT*, page 205–216.

Huang, Y. and Xiong, D. (2023). Cbbq: A chinese bias benchmark dataset curated with human-ai collaboration for large language models. arXiv:2306.16244.

Ivetta, G., Palombini, P., Martinelli, S., Gomez, M. J., Dev, S., Prabhakaran, V., and Benotti, L. (2025). Adaptive data collection for latin-american community-sourced evaluation of stereotypes (laces). arXiv:2510.24958.

Jiang, A. Q. et al. (2024). Mixtral of experts. arXiv:2401.04088.

Jin, J., Kim, J., Lee, N., Yoo, H., Oh, A., and Lee, H. (2024). Kobbq: Korean bias benchmark for question answering. arXiv:2307.16778.

Messeri, L. and Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627:49–58.

Mitchell, M. et al. (2025). Shades: Towards a multilingual assessment of stereotypes in large language models. In *NAACL*, pages 11995–12041.

Moreira, D. A. B. et al. (2024). Fairpivara: Reducing and assessing biases in clip-based multimodal models. arXiv:2409.19474.

Nadeem, M., Bethke, A., and Reddy, S. (2020). Stereoset: Measuring stereotypical bias in pretrained language models. arXiv:2004.09456.

Nangia, N., Vania, C., Bhalerao, R., and Bowman, S. R. (2020). CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *EMNLP*, pages 1953–1967.

Neplenbroek, V., Bisazza, A., and Fernández, R. (2024). Mbbq: A dataset for cross-lingual comparison of stereotypes in generative llms. arXiv:2406.07243.

OLMo, T. et al. (2025). 2 olmo 2 furious. arXiv:2501.00656.

OpenAI et al. (2024). Gpt-4 technical report. arXiv:2303.08774.

Ovalle, A., Goyal, P., Dhamala, J., Jaggers, Z., Chang, K.-W., Galstyan, A., Zemel, R., and Gupta, R. (2023). "i'm fully who i am": Towards centering transgender and non-binary voices to measure biases in open language generation. In *FAccT*, page 1246–1266.

Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., and Bowman, S. R. (2022). Bbq: A hand-built bias benchmark for question answering. arXiv:2110.08193.

Satheesh, S., Klug, K., Beckh, K., Allende-Cid, H., Houben, S., and Hassan, T. (2025). Gg-bbq: German gender bias benchmark for question answering. arXiv:2507.16410.

Shimabucoro, L., Ruder, S., Kreutzer, J., Fadaee, M., and Hooker, S. (2024). Llm see, llm do: Guiding data generation to target non-differentiable objectives. arXiv:2407.01490.

Silva, J. et al. (2024). Avaliação de ferramentas de Ética no levantamento de considerações Éticas de modelos de linguagem em português. In *LAAI*, pages 61–64. SBC.

Soares, T., Gumiel, Y., Junqueira, R., Gomes, T., and Pagano, A. (2023). Viés de gênero na tradução automática do gpt-3.5 turbo: avaliando o par linguístico inglês-português. In *STIL*, pages 167–176.

Talat, Z., Blix, H., Valvoda, J., Ganesh, M. I., Cotterell, R., and Williams, A. (2022). On the machine learning of ethical judgments from natural language. In *NAACL*, pages 769–779.

Team, G. et al. (2024). Gemma: Open models based on gemini research and technology. arXiv:2403.08295.

Touvron, H. et al. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288.

Zhao, J., Khashabi, D., Khot, T., Sabharwal, A., and Chang, K.-W. (2021). Ethical-advice taker: Do language models understand natural language interventions? arXiv:2106.01465.