Avaliação do senso comum em modelos de linguagem através de benchmarks: Desafio de Winograd aplicado ao ChatGPT em português brasileiro

Thiago Gomes do Nascimento, Diogo Cortiz

Pontifícia Universidade Católica de São Paulo (PUC– SP) – São Paulo, SP – Brasil

thiago.gnascimento1@gmail.com, diogocortiz@gmail.com

Abstract. *The assessment of language models with benchmarks is presented as an effective way of evaluating their comprehension limits. In this regard, the Winograd Schema Challenge, which aims to assess common sense through pronoun disambiguation tasks, has led to the development of different metrics and datasets. When applying a translation of the Winograd Challenge to Brazilian Portuguese to ChatGPT, we identified comparable results to those obtained in English. However, these results must be analyzed with caution, considering the potential biases in the model training process and the existing gaps in the reasoning dimensions covered by the available evaluation methods.*

Resumo. *O desempenho em benchmarks é apresentado como uma forma de avaliação efetiva dos limites de compreensão dos modelos de linguagem. Neste sentido, o desafio de esquemas de Winograd, que se propõe a avaliar o senso comum por meio de tarefas de desambiguação de pronomes, deu origem a diferentes métricas e datasets. Ao aplicar a tradução do desafio de Winograd ao ChatGPT em português brasileiro, identificamos resultados equiparáveis aos obtidos em inglês. Contudo, é preciso ter cautela ao interpretar estes dados, visto que existem vieses associados ao treinamento dos modelos e lacunas quanto às dimensões de raciocínio contempladas pelos métodos de avaliação disponíveis.*

1. Introdução

Os modelos generativos de linguagem, como o GPT desenvolvido pela Open AI, o LaMDA, presente no Google Bard, e o LLaMa da Meta têm ganhado notoriedade pela capacidade de processamento de textos com resultados semelhantes aos produzidos por humanos. Apesar do crescimento das bases de dados e do aprimoramento dos algoritmos, estes modelos estão associados a aspectos sintáticos [Floridi, 2023]. Devido à característica de combinar sequências linguísticas de acordo com cálculos estatísticos, sem um entendimento semântico, são classificados como papagaios estocásticos [Bender et al., 2021].

Turing [1950] foi pioneiro na criação de uma definição operacional de inteligência. Diante da necessidade latente de avaliar os limites de compreensão dos modelos, tanto entre ferramentas quanto em relação ao ser humano, o Desafio de Winograd [Levesque et al., 2012] surgiu como uma evolução do teste de Turing, no intuito de avaliar o senso comum, habilidade intrinsecamente antropomórfica, de

maneira objetiva. O desafio consiste em responder uma pergunta binária associada a uma frase ambígua, como por exemplo: a medalha não cabe na maleta porque ela é muito grande. O que é muito grande? a) a mala ou b) a medalha?

Apesar dos significativos avanços em termos de processamento, a predominância no inglês na construção e no treinamento dos modelos desperta questões relativas à confiabilidade dos resultados fornecidos para idiomas sub-representados [Petrov et al., 2023]. Partindo da tradução do Desafio de Winograd para o português brasileiro [Melo et al., 2019], avaliamos o desempenho do ChatGPT na tarefa de desambiguação de pronomes a fim de identificar se haveria alguma disparidade com os resultados publicados em inglês.

2. Winograd e a evolução dos benchmarks

O jogo da imitação [Turing, 1950] foi proposto como uma maneira de avaliar a inteligência da máquina. Através da interação por perguntas, uma pessoa exercendo o papel de juiz deve identificar qual dos respondentes é um humano e qual é uma máquina. Uma das críticas ao teste consiste na objeção da Lady Lovelace, segundo a qual a máquina faz apenas aquilo que mandamos. O argumento de Turing justifica que, apesar de receber instruções básicas, um computador capaz de aprender de forma autônoma poderia fazer o que não foi determinado anteriormente por seu programador [French, 2000].

A falta de objetividade foi considerada com outra deficiência do teste de Turing, que dependia da avaliação subjetiva do julgador humano. Utilizando a contribuição de Winograd [1972] no estudo da compreensão da linguagem natural, Levesque et al. [2012] propuseram o desafio de esquemas de Winograd. O teste consiste na avaliação de um grupo de declarações com sentido dúbio, cuja resposta não pode ser obtida na frase, apenas através de conhecimento prévio, ou seja, o senso comum que permite a associação semântica.

Devido à limitação de serem elaborados inicialmente em inglês, os esquemas foram traduzidos para diferentes idiomas como português [Melo et al., 2019], francês [Amsili; Seminck, 2017], húngaro [Vadász; Ligeti-Nagy, 2022], mandarim [Bernard; Han, 2020] e russo [Shavrina et al., 2020]. Além disso, diversos autores apresentaram evoluções dos esquemas. Construído com a proposta de ser um dataset multilíngue, o Wino-X [Emelin; Sennrich, 2021] contempla esquemas em alemão, francês e russo alinhados com as versões originais em inglês.

Além disso, diversos autores apresentaram evoluções das bases de dados utilizadas, como o Winogrande [Sakaguchi et al., 2021], que contempla 44.000 esquemas. Algumas alternativas encontradas para a ampliação do volume de esquemas são o Winoflexi [Isaak; Michael, 2019], que utiliza *crowdsourcing* para o desenvolvimento de novas sentenças e o Winventor [Nicos; Michael, 2020] que busca automatizar a criação de esquemas. No Winologic [He et al., 2021] novas frases foram construídas utilizando teoremas lógicos.

Na adaptação WNLI [Wang et al., 2018], o desafio de Winograd foi reformulado como uma tarefa de inferência de linguagem natural. Neste caso, o formato das tarefas é composto por três partes. Premissa: a medalha não cabe na maleta porque ela é muito grande. Hipótese: a medalha é muito grande. Resposta: verdadeiro/falso. Esta versão foi

adicionada ao benchmark GLUE, em conjunto com outras tarefas. O aprimoramento trazido pelo SUPERGLUE [Wang et al., 2019], apresentado como detentor de um maior nível de dificuldade, considera a versão original dos esquemas. Storks et al. [2019] apresentam uma classificação para os benchmarks que realizam a avaliação de raciocínio de senso comum para compreensão de linguagem natural, de acordo com o tipo de atividade testada. Um levantamento mais recente, realizado por Davis [2023] lista 139 benchmarks, sendo 102 de texto, 18 para imagens, 12 de vídeos e 7 de ambientes físicos.

3. Método

Dos 285 esquemas do desafio de Winograd, a tradução elaborada por Melo et al. [2019] classifica 8 como sem equivalência no português. Assim, para este estudo foram utilizados 277 esquemas. A base contempla a frase principal de cada etapa do teste seguida por colunas que indicam o trecho ambíguo e o pronome que é necessário para identificar a referência. Além disso, também estão descritas quais são as alternativas e a resposta correta para cada tarefa.

"schema": "Os vereadores recusaram a autorização aos manifestantes porque eles temiam a violência.",
"snippet": "eles temiam a violência.",
"pronoun": "eles",
"correct_answer": "A",
"substitution_a": "Os vereadores",
"substitution_b": "Os manifestantes"

Neste experiemento, foi utilzada a versão grátis do ChatGPT, através da interface disponível no endereço *https://chat.openai.com/*. O *prompt* utilizado seguiu o padrão original de formulação dos esquemas, exposto na seção 1. As tarefas foram adicionadas uma a uma, iniciando pela frase principal do desafio. Em seguida, era perguntado quem ou que estava associado ao trecho ambíguo, sempre com duas opções de resposta, identificadas como “a” e “b”, conforme é possível observar na figura 1. As alternativas retornadas pelo ChatGPT foram coletadas e adicionadas a uma nova coluna da base. Comparando a coluna das respostas corretas com a coluna das respostas obtidas foi possível calcular o percentual de acerto em relação ao total de perguntas realizadas.

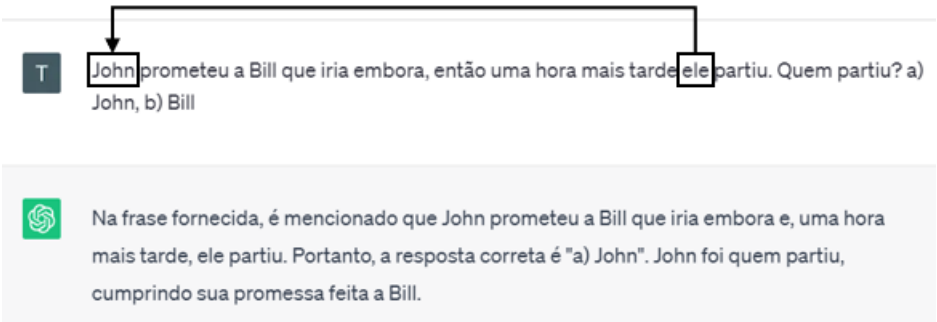

Figura 1. Exemplo de interação com o ChatGPT para obtenção das respostas.

4. Resultados

O ChatGPT, que incorpora o modelo GPT-3.5 apresentou um percentual de acerto de 87,5% em relação às 277 perguntas do desafio de Winograd respondidas em português brasileiro. Na tabela 1, também é possível observar que o resultado obtido é similar ao apresentado por modelos de linguagem de grande porte em tarefas de desambiguação de pronomes no idioma inglês.

Quanto às versões do teste, WSC285 refere-se à lista completa do desafio, que contém 12 tarefas além dos 273 esquemas presentes na versão anterior, conhecida como WSC273. O presente estudo teve uma diferença de apenas 0,8% abaixo do modelo GPT-3 em inglês [Brow et al., 2020]. Ao contrapor os resultados com o GPT 3.5, é preciso considerar que os dados divulgados são referentes à versão Winogrande do desafio, que conforme mencionado anteriormente, possui uma base mais extensa. Na comparação com o RoBERTa [Sakaguchi et al., 2021], evolução do modelo BERT com melhorias de arquitetura, o ChatGPT apresentou uma diferença de 2,6% abaixo do resultado para a versão WSC 273.

Tabela 1. Resultado obtido comparado com desempenho registrado em estudos anteriores.

Modelo	Idioma	Desempenho	Versão do teste	Fonte
GPT-3.5	Português	87,5%	WSC 285	Própria
GPT-3	Inglês	88,3%	WSC 273	Brown et al. [2020]
GPT-3.5	Inglês	81,6%	Winogrande	OpenAI [2023]
RoBERTa	Inglês	90,1%	WSC 273	Sakaguchi et al. [2021]

5. Conclusão

Apesar da pontuação da apresentada, não é possível concluir que o ChatGPT possui senso comum nem que o desempenho do modelo em português para outras tarefas é tão confiável quanto a versão original, em inglês. O desafio de Winograd está disponível na internet desde 2012, sendo provável que tenha feito parte dos dados de treinamento dos atuais modelos de linguagem. Além disso, o sucesso em uma tarefa específica não é uma métrica confiável para avaliar o senso comum.

Dentre estas tarefas, a desambiguação de pronomes contempla apenas uma pequena parcela do senso comum que é preciso para compreensão da linguagem. Considerando os demais *benchmarks* disponíveis para senso comum, apenas duas dimensões de raciocínio são avaliadas adequadamente: o taxonômico, que se refere a classificações e ao conhecimento enciclopédico, e o numérico, que abrange cálculos e quantidades. As demais dimensões, do raciocínio (temporal, psicológico, espacial, físico, biológico, social, comparativo, meta-raciocínio) não são contempladas ou são abordadas parcialmente. Assim, a construção de novos *benchmarks* deve considerar aspectos ainda não explorados do senso comum, a fim de propiciar uma avaliação mais adequada da capacidade dos modelos [Kocijan et al., 2023; Davis, 2023].

Referências

Amsili, P.; Seminck, O. (2017) “A Google-Proof Collection of French Winograd Schemas”, Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017), p. 24-29.

Bender, E. M. et al. (2021) “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?”, FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, p. 610-623.

Bernard, T.; Han, T. (2020) “Mandarinograd: A Chinese Collection of Winograd Schemas”, Proceedings of the Twelfth Language Resources and Evaluation Conference, p. 21-26.

Brown, T. B. et al. (2020) “Language models are few-shot learners”, NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems, p. 1877-1901.

Davis, E. (2023) “Benchmarks for Automated Commonsense Reasoning: A Survey”, arXiv:2302.04752v2, https://doi.org/10.48550/arXiv.2302.04752

Emelin, D.; Sennrich, R. (2021) “Wino-X: Multilingual Winograd Schemas for Commonsense Reasoning and Coreference Resolution”, Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, p. 8517-8532.

Floridi, L. (2023) “AI as Agency Without Intelligence: on ChatGPT, Large Language Models, and Other Generative Models”, Philosophy & Technology 36 (15), p. 1-7.

French, R. M. (2000) “The turing test: The first 50 years”, Trends in Cognitive Sciences 4 (3), p. 115-122.

He, W. et al. (2021) “WINOLOGIC: A Zero-Shot Logic-based Diagnostic Dataset for Winograd Schema Challenge.”, Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, p. 3779–3789.

Isaak, N.; Michael, L. (2019) “WinoFlexi: A Crowdsourcing Platform for the Development of Winograd Schemas” In: Liu, J., Bailey, J. (eds) AI 2019: Advances in Artificial Intelligence. AI 2019. Lecture Notes in Computer Science 11919.

Kocijan, V. et al. (2023) “The Defeat of the Winograd Schema Challenge”, arXiv:2201.02387v3, https://doi.org/10.48550/arXiv.2201.02387

Levesque, H. J.; Davis, E.; Morgenstern, L. (2012) “The Winograd Schema Challenge”, Thirteenth international conference on the principles of knowledge representation and reasoning.

Melo, G. S. D.; Imaizumi, V. A.; Cozman, F. G. (2019), “Winograd Schemas in Portuguese”, Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2019), p. 787–798.

Nicos, I.; Michael, L. (2020) “Winventor: A Machine-driven Approach for the Development of Winograd Schemas”, Proceedings of the 12th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART, p. 26-35.

OpenAI (2023) “GPT-4 Technical Report”, arXiv:2303.08774v3, https://doi.org/10.48550/arXiv.2303.08774

Petrov, A. et al. (2023), “Language Model Tokenizers Introduce Unfairness Between Languages”, oarXiv:2305.15425v1, https://doi.org/10.48550/arXiv.2305.15425

Pires, R. et al. (2023), “Sabiá, Portuguese Large Language Models”, arXiv:2304.07880v2, https://doi.org/10.48550/arXiv.2304.07880

Sakaguchi et al. (2021), “WinoGrande: An Adversarial Winograd Schema Challenge at Scale”, Communications of the ACM 64(9), p. 99-106.

Shavrina, T. et al. (2020), “RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark”, EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, p. 4717-4726.

Storks, S.; Gao, Q.; Chai, J. Y. (2019) “Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches”, arXiv:1904.01172v3, https://doi.org/10.48550/arXiv.1904.01172

Turing, A. M. (1950) “Computing machinery and intelligence”, Mind LIX (236), p. 433-460.

Vadász, N.; Ligeti-Nagy, N. “Winograd schemata and other datasets for anaphora resolution in Hungarian”, Acta Linguistica Academica 69 (4), p. 564-580.

Wang, A. et al. (2018) “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding”, Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP , p. 353-355.

Wang, A. et al. (2019) “SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems”, NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems, p. 3266-3280.

Winograd, T. (1972) “Understanding natural language”, Cognitive Psychology 3(1), p. 1 – 191.