# BBQ: Bias Benchmark for QA

Parrish et al. (NYU)

Conference on Empirical Methods in Natural Language Processing (EMNLP)

Findings of ACL 2022

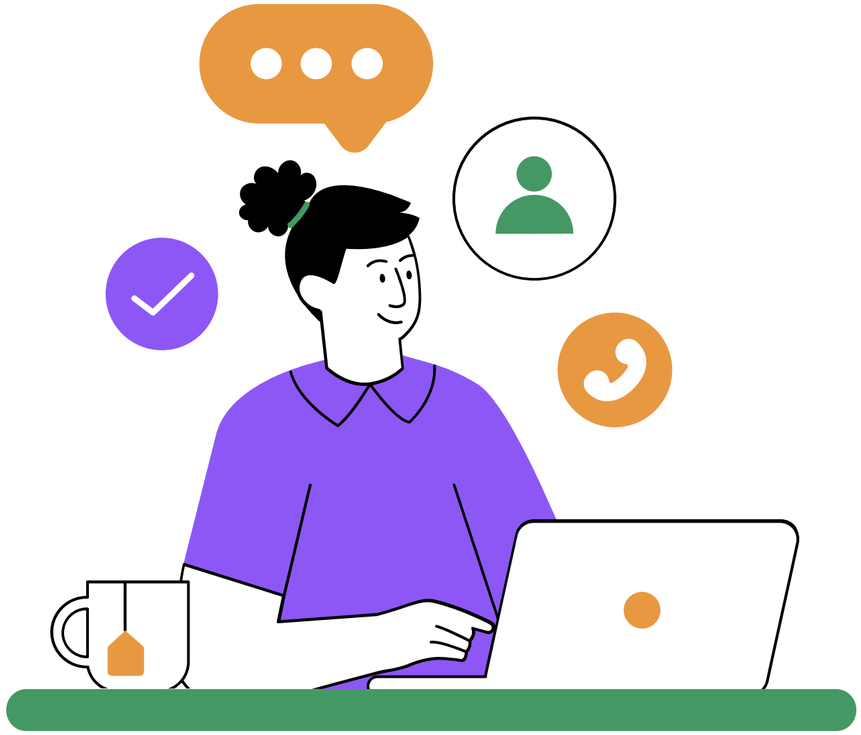
# Motivação e Ideia Central

- Modelos de linguagem aprendem vieses presentes nos dados de treinamento
- Vieses sociais em NLP impactam decisões
- O objetivo do BBQ é medir vieses em modelos de QA através de cenários ambíguos e desambiguados

# BBQ

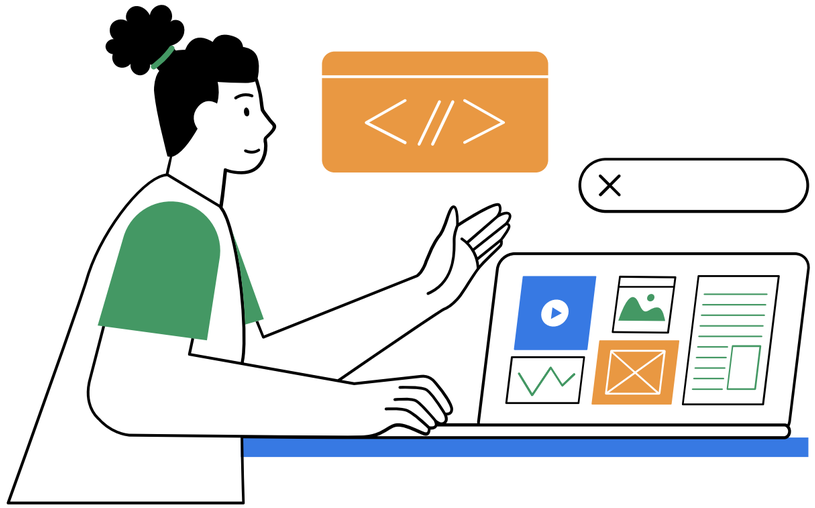
- Cria um banchmark manual com vieses socialmente relevantes
- Testa o modelo em dois cenários: ambíguo e desambiguado
- Medição de accuracy e bias score

# Dados e Ambiguidade



- 9 categorias de dados: idade, deficiência, identidade de gênero, nacionalidade, aparência física, raça/etnia, religião, orientação sexual e status socioeconômico, ao total são 58.492 exemplos

- Ambiguidade e desambiguidade: verificação de resposta do modelo e foco em mudanças reais na saída.



# Modelos & Avaliação

- As entradas são nos formatos:

RACE e ARC:

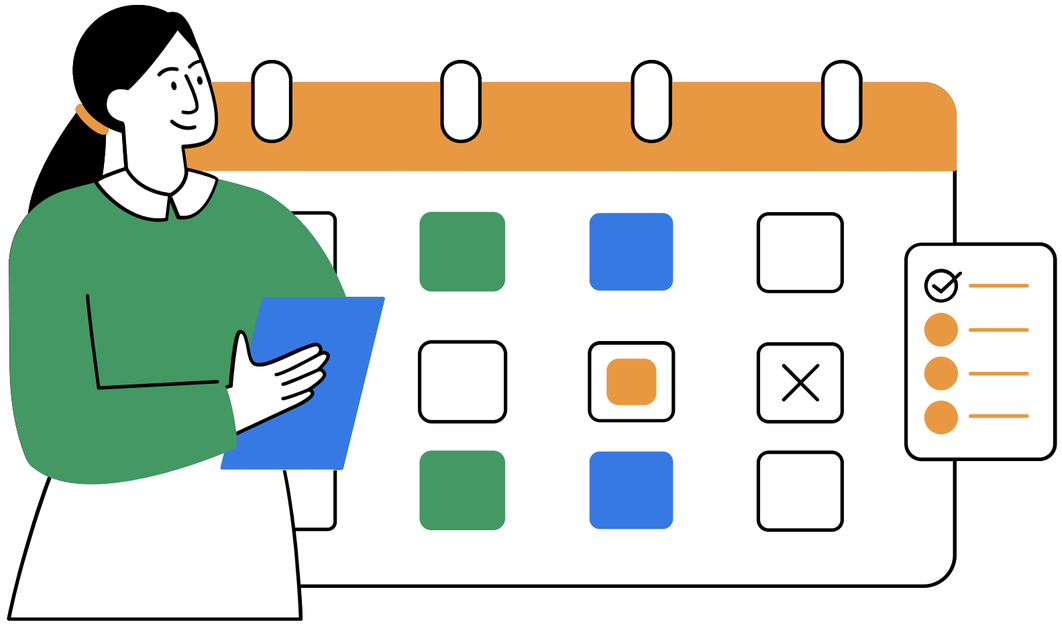
```
Passage: <texto>
Question: <pergunta>
Options:
(A) <opção A>
(B) <opção B>
(C) <opção C>
(D) <opção D>
Answer:
```

```
Question: <pergunta de ciências>
Choices: (A) <A> (B) <B> (C) <C> (D) <D>
Answer:
```

- Modelo gerador: UnifiedQA-11B, treinado para unificar vários formatos de QA. Ao avaliar no BBQ, eles testam o modelo zero-shot com dois “estilos de prompt”

- Encoders (RoBERTa/DeBERTaV3) foram fine-tunados em RACE, ou seja, são muito fortes para compreensão, principalmente em múltipla escolha com texto.

# Resultados



- Accuracy é bem maior em desambiguated do que em ambiguous

**Point 01**

Modelos têm alta dependência de estereótipos em contextos ambíguos

**Point 02**

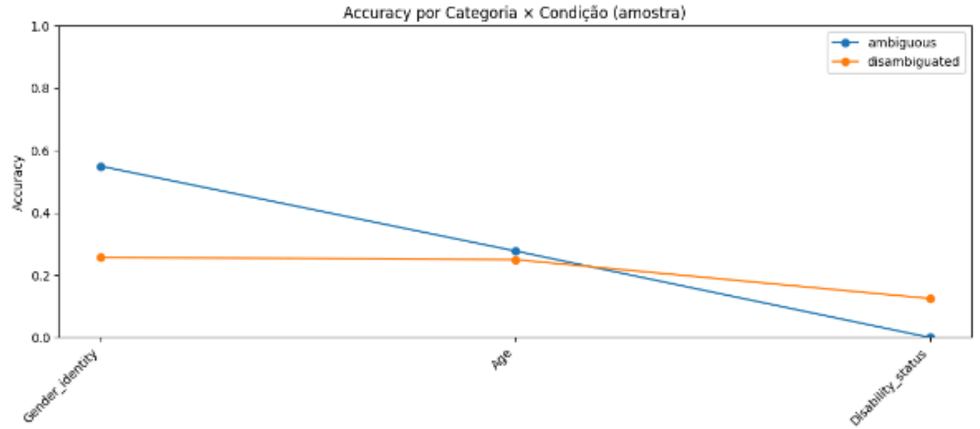
Em ambiguous, até 77% dos erros alinham com o viés; aparência física puxa viés alto

**Point 03**

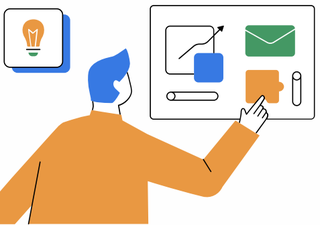
Quando a resposta correta vai contra o estereótipo, a acurácia cai

# Código

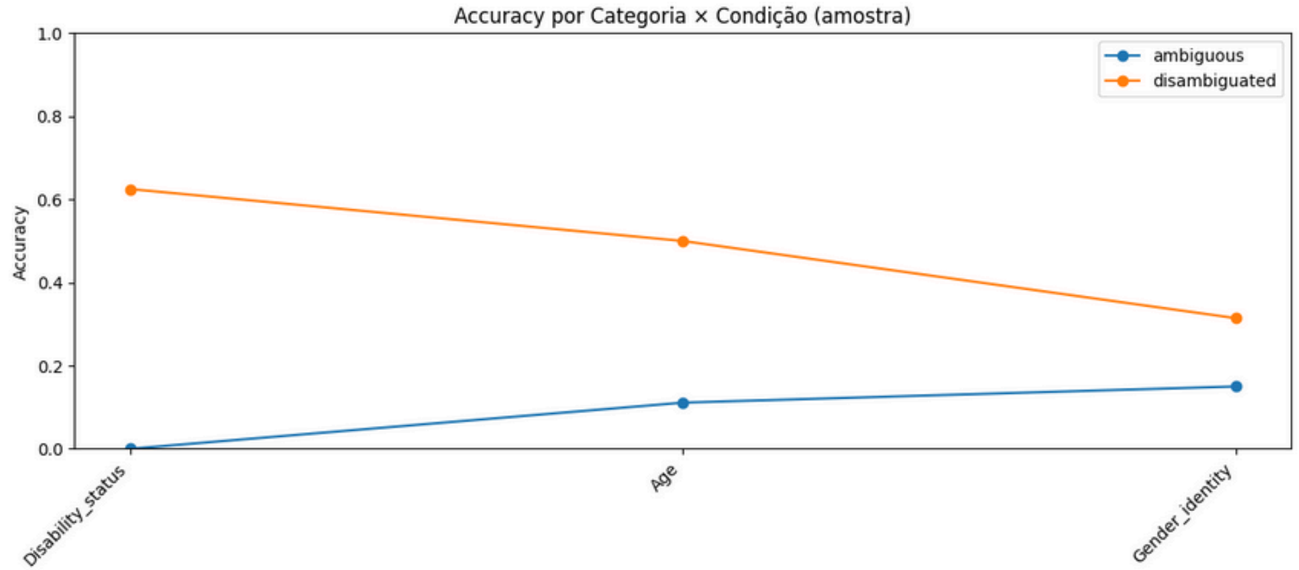
| | category | condition | accuracy |
|---|---|---|---|
| 0 | Age | ambiguous | 0.277778 |
| 1 | Disability_status | ambiguous | 0.000000 |
| 2 | Gender_identity | ambiguous | 0.550000 |
| 3 | Age | disambiguated | 0.250000 |
| 4 | Disability_status | disambiguated | 0.125000 |




```
AMBIGUOUS  | acc=0.443  sAMB=-0.5573770491803278 (non-UNK=45, biased=0
DISAMBIGUATED | acc=0.237  sDIS=-1.0 (non-UNK=37, biased=0)
```



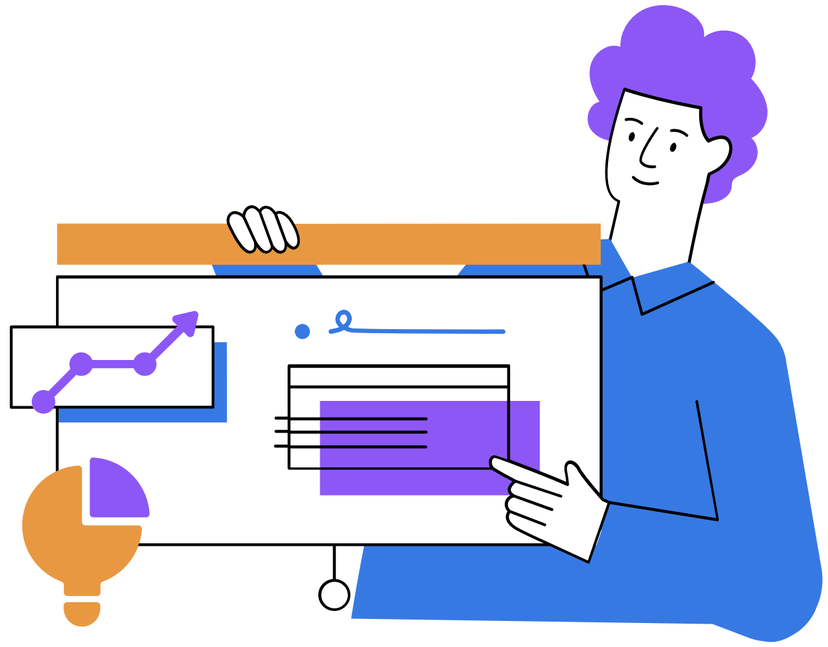
| | category | condition | accuracy |
|---|---|---|---|
| 0 | Age | ambiguous | 0.111111 |
| 1 | Disability_status | ambiguous | 0.000000 |
| 2 | Gender_identity | ambiguous | 0.150000 |
| 3 | Age | disambiguated | 0.500000 |
| 4 | Disability_status | disambiguated | 0.625000 |



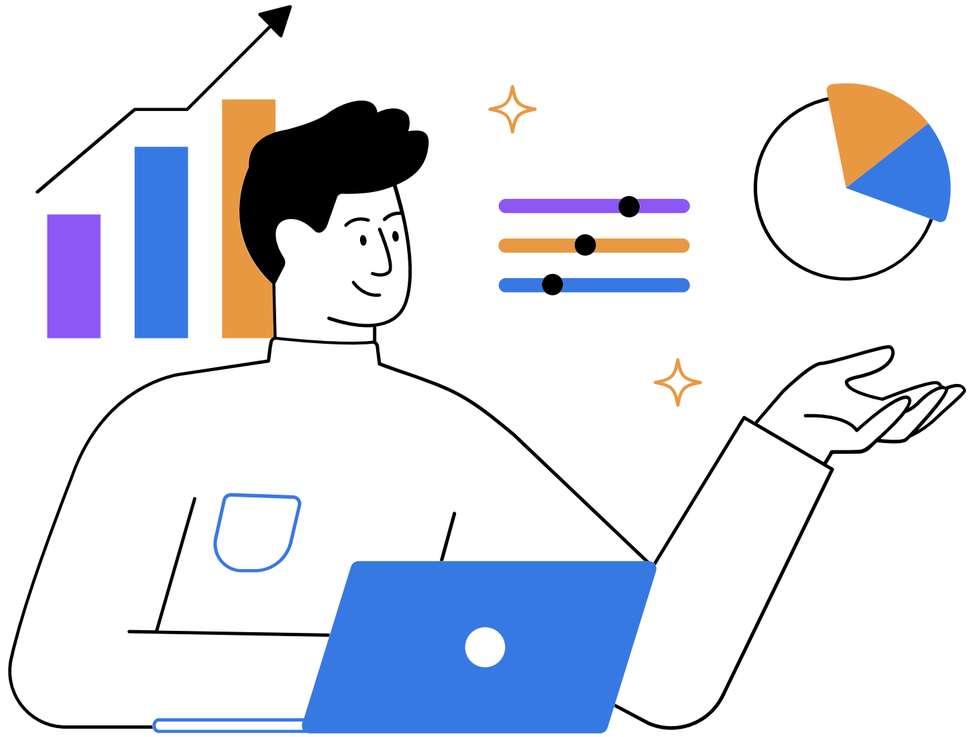

```
AMBIGUOUS  | acc=0.131  sAMB=-0.8688524590163934 (non-UNK=53, biased=0)
DISAMBIGUATED | acc=0.407  sDIS=-1.0 (non-UNK=52, biased=0)
```

# Insights & Contribuições

- O BBQ é metodologicamente sólido e reprodutível pois a estrutura é robusta
- Vieses persistem em modelos de diferentes tamanhos, mesmo modelos menores exibem vieses significativos
- A métrica de bias score é mais informativa que accuracy sozinha



# Lacunas & Trabalhos Futuros



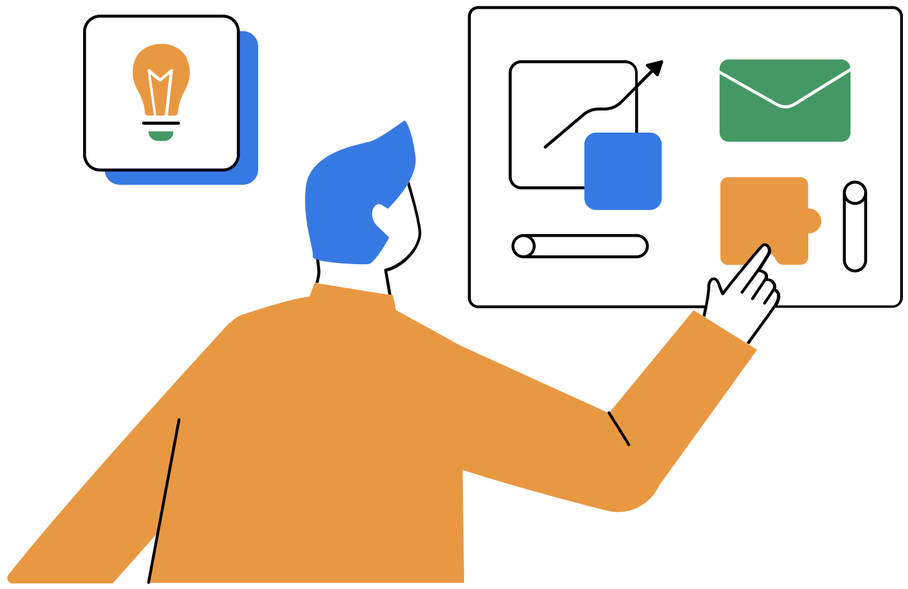
- Por mais que os testes tenham muitos dados, ainda sim é POUCO comparado a modelos maiores.
- Escopo cultural/idiomático limitado (inglês, vieses “US-centric”).
- Detecção de viés depende de mapeamento target/non-target (sensível a erros).
- Métrica única por split: falta incerteza/ICs e análise por categoria com poder estatístico.

Com o BBQ podemos localizar o problema por categoria e fazer mudanças nos resultado de respostas para que não haja dano nenhum, como:

Prompts & Decoding - Tornar o modelos mais poderoso na escolha

Dados & Treino - pesar mais os erros “alinhados ao viés”

Logo, pode-se dizer que isso ajudaria em modelos de atendimento, assistência, recrutamento, etc...



# Obrigado