CHATGPT 3 VS RESPOSTAS HUMANAS

Uma investigação e visualização em um dataset de perguntas e respostas

Giovani da Silva| 2024

INTRODUÇÃO

O objetivo desse trabalho é avaliar como as respostas de humanos se diferenciam em vários aspectos a respostas do OpenAI ChatGPT 3.0, utilizando vários modelos de PLN(Processamento de Linguagem Natural) disponíveis em código aberto.

DATASET

- O dataset utilizado neste trabalho é conhecido como HC3 e está disponível na plataforma Hugging Face. -
- Ele é composto por um total de 23.688 perguntas, cada uma acompanhada por duas respostas distintas: uma fornecida por humanos e outra gerada pelo ChatGPT 3.0.
- Abrange perguntas oriundas de várias fontes.

Question	Human Answer	ChatGPT Answer	Source
Why do we say "an apple a day keeps the doctor away"?	It's an old saying emphasizing eating healthy...	This phrase highlights the importance of a healthy diet...	reddit_eli5
What are the latest advancements in cardiac surgery?	Minimally invasive techniques and robotic...	Recent advancements include robotic-assisted surgery...	medicine
How do cryptocurrencies impact traditional banking?	They challenge the traditional financial system...	Cryptocurrencies offer decentralized finance...	finance
Who won the Nobel Prize in Physics in 2023?	[Name of the Winner], for breakthroughs in...	The 2023 Nobel Prize in Physics was awarded to...	open_qa
What is the Turing Test and its significance in AI?	It's a test to determine if a machine can...	The Turing Test, proposed by Alan Turing, is a measure...	wiki_csai

FONTES DAS PERGUNTAS

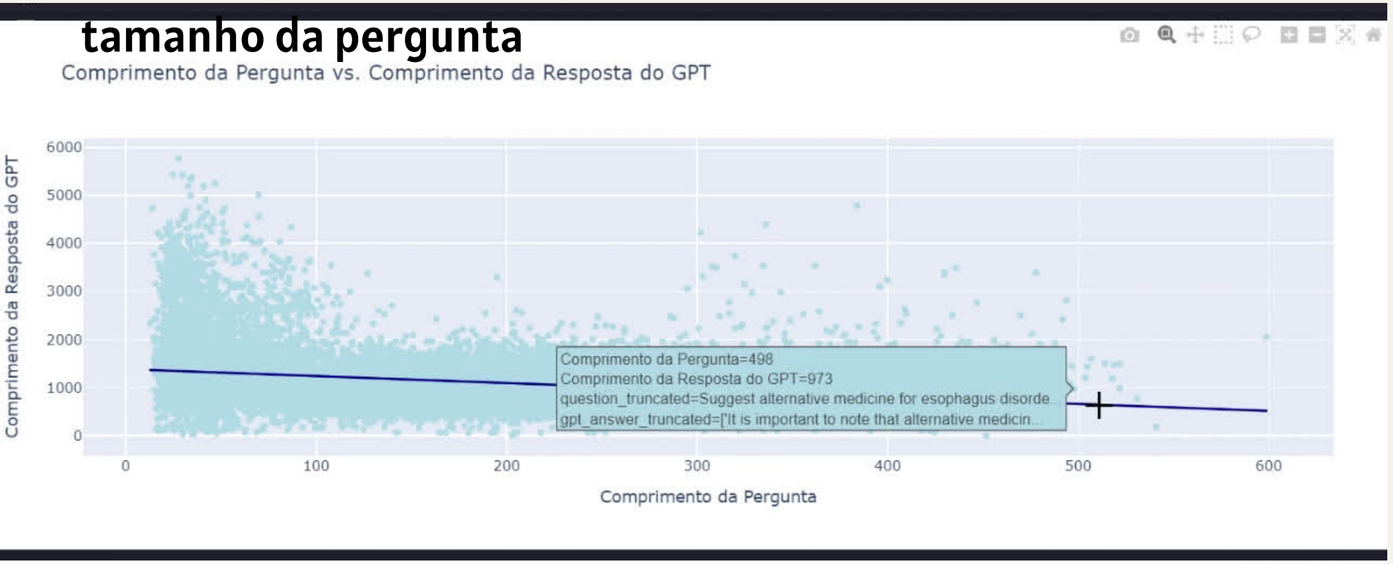
Reddit ELI5	Perguntas do subreddit "Explique Como Se Eu Tivesse 5 Anos", onde os usuários buscam explicações simplificadas para tópicos complexos.
Finanças	Inclui perguntas relacionadas a assuntos financeiros e econômicos.
Medicina	Perguntas sobre tópicos médicos e de saúde.
Open QA	Perguntas de plataformas de perguntas e respostas abertas, onde os usuários podem fazer qualquer pergunta e receber respostas da comunidade.
Wiki_CSAI	Perguntas relacionadas a ciência da computação e inteligência artificial (CSAI).

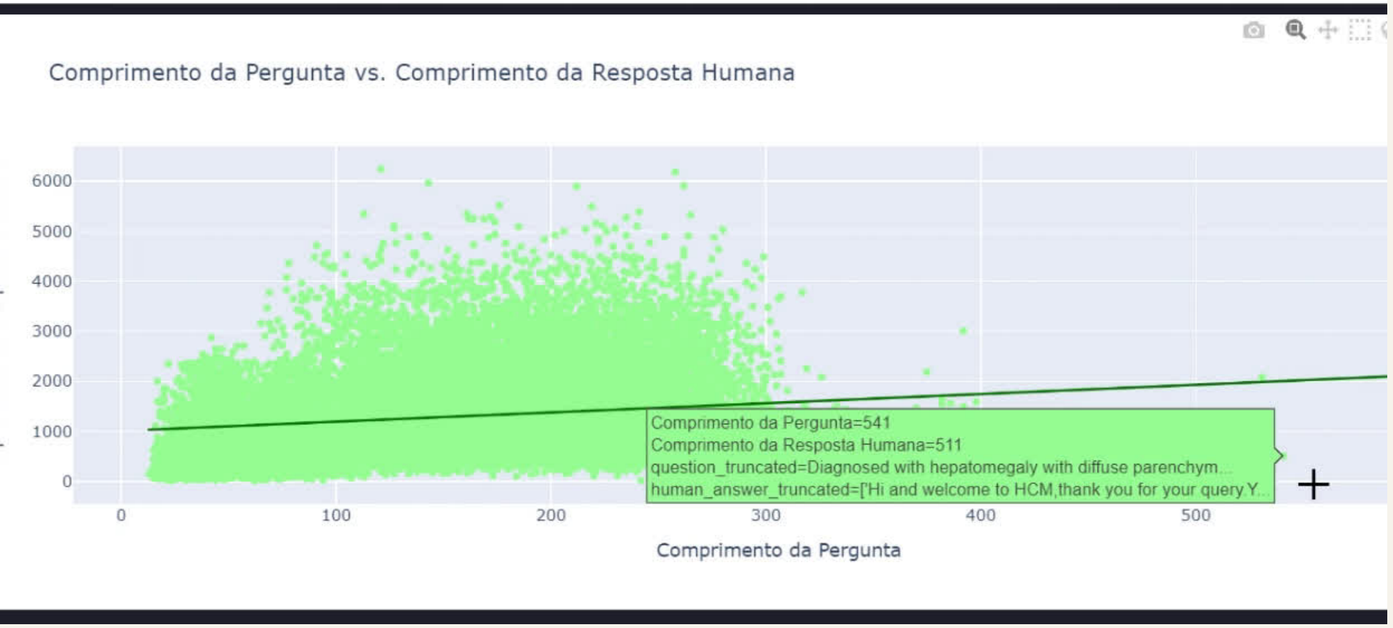
AVALIAÇÕES

- 1 - Relação entre tamanho da pergunta e das respostas
- 2 - Análise de sentimento das respostas
- 3 - Similaridade entre as respostas
- 4 - Expressões mais utilizadas
- 5 - Voz passiva e ativa
- 6 - Detecção de escrita por IA
- 7 - Sentido com o senso comum

1 - GRÁFICOS INTERATIVOS SOBRE DIFERENÇA DE TAMANHO ENTRE PERGUNTAS E RESPOSTAS

- **Visualização de diferenças de tendência entre tamanho da resposta do gpt e dos humanos com relação ao tamanho da pergunta**



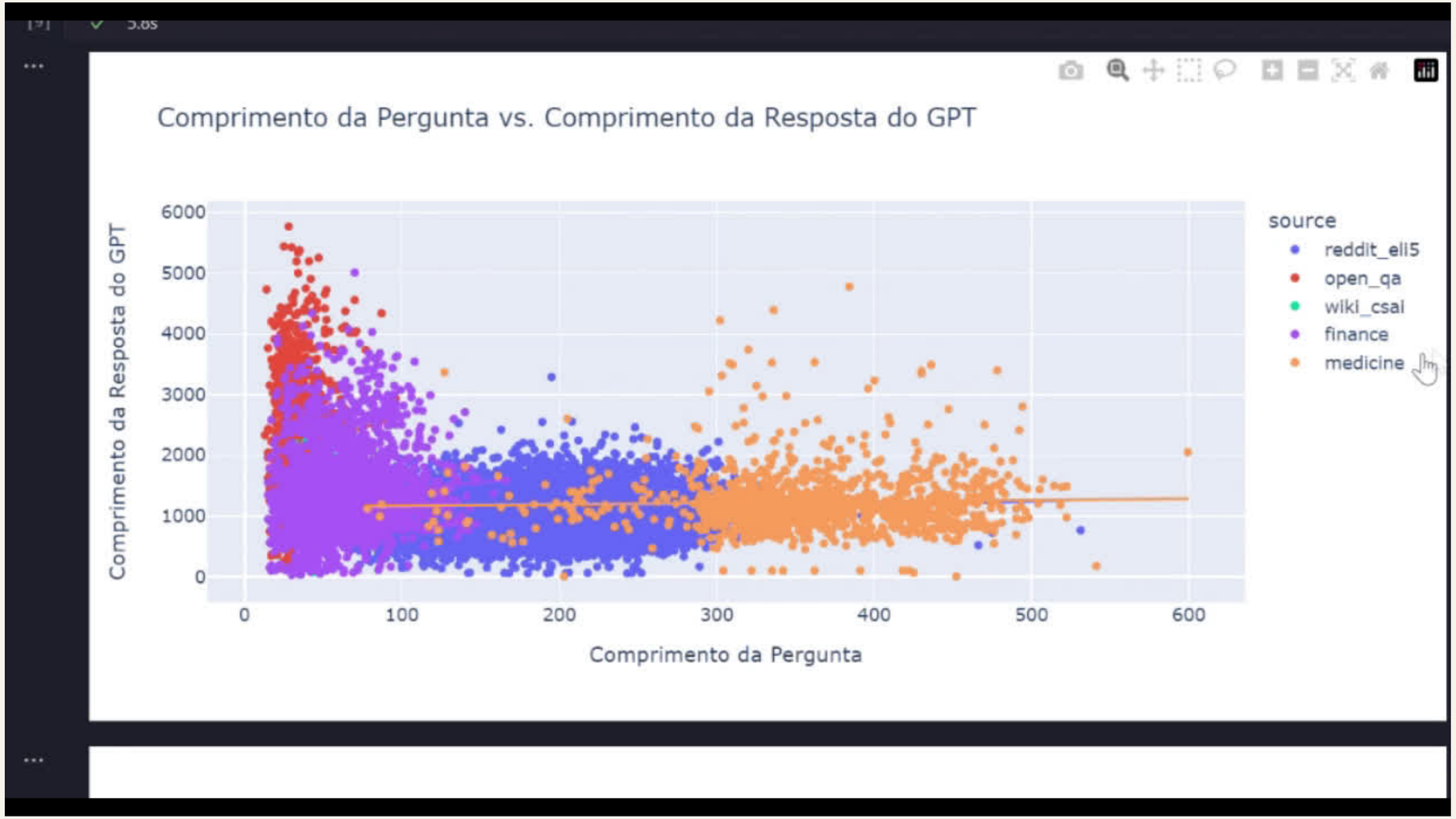


- **GPT tem a tendência de diminuir o tamanho da resposta com o aumento da tamanho da pergunta**
- **O contrário acontece com as respostas humanas**
- **Concentração maior de respostas humanas na faixa intermediária**
- **GPT com distribuição do tamanho de resposta mais diverso, porém com muitas respostas curtas**

I - GRÁFICOS INTERATIVOS SOBRE DIFERENÇA DE TAMANHO ENTRE PERGUNTAS E RESPOSTAS

- Visualização de diferenças de tendência entre tamanho da resposta do gpt e dos humanos com relação ao tamanho da pergunta

- GPT tem a tendência de diminuir o tamanho da resposta co o aumento da tamanho da pergunta

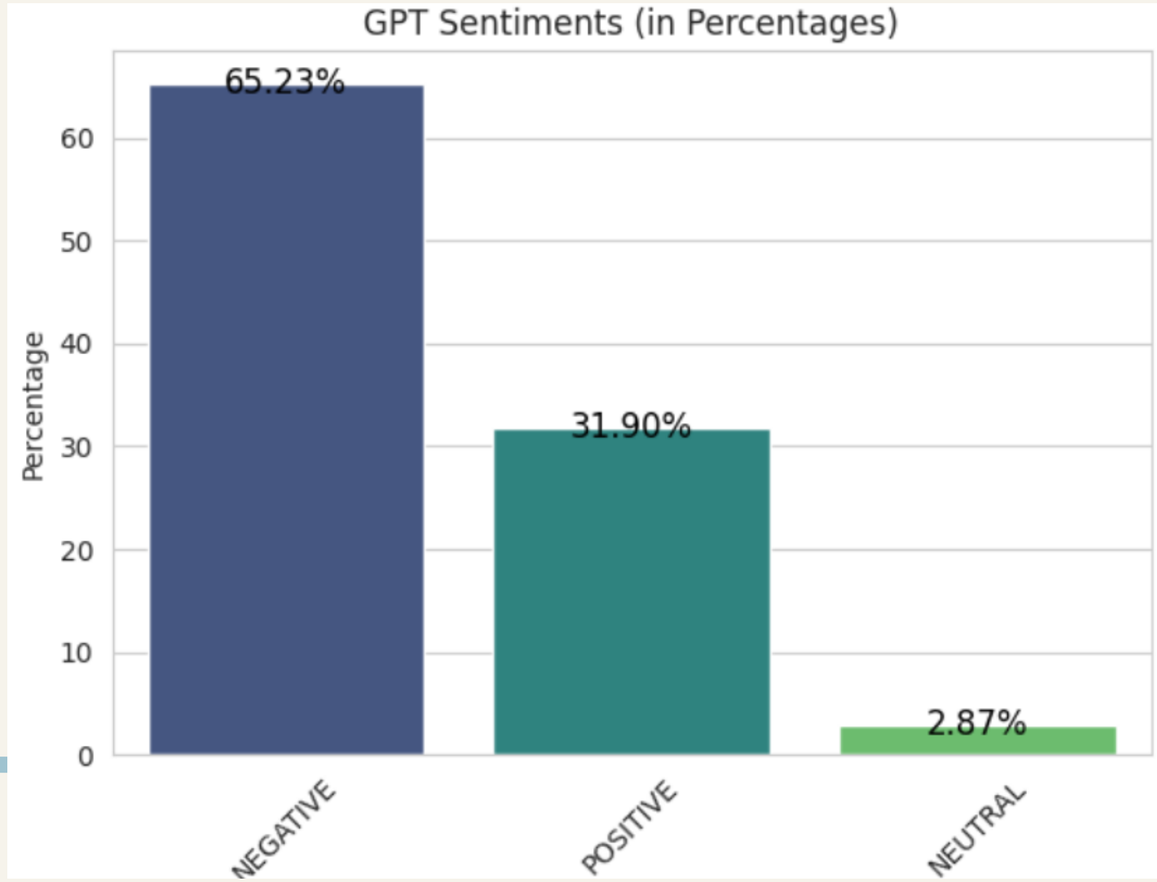


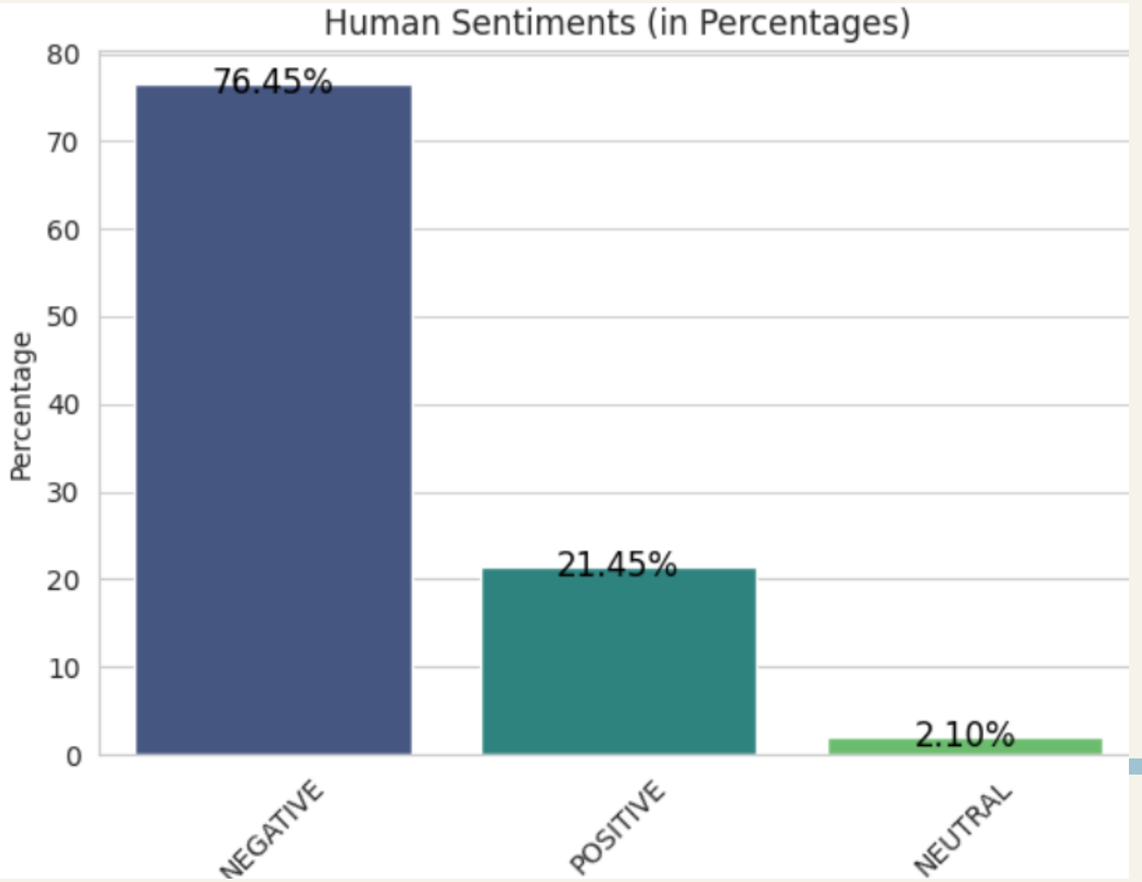
I - GRÁFICOS INTERATIVOS SOBRE DIFERENÇA DE TAMANHO ENTRE PERGUNTAS E RESPOSTAS

- **Visualização de diferenças de tendência entre tamanho da resposta do GPT e dos humanos com relação ao tamanho da pergunta**

2 - ANÁLISE DE SENTIMENTO

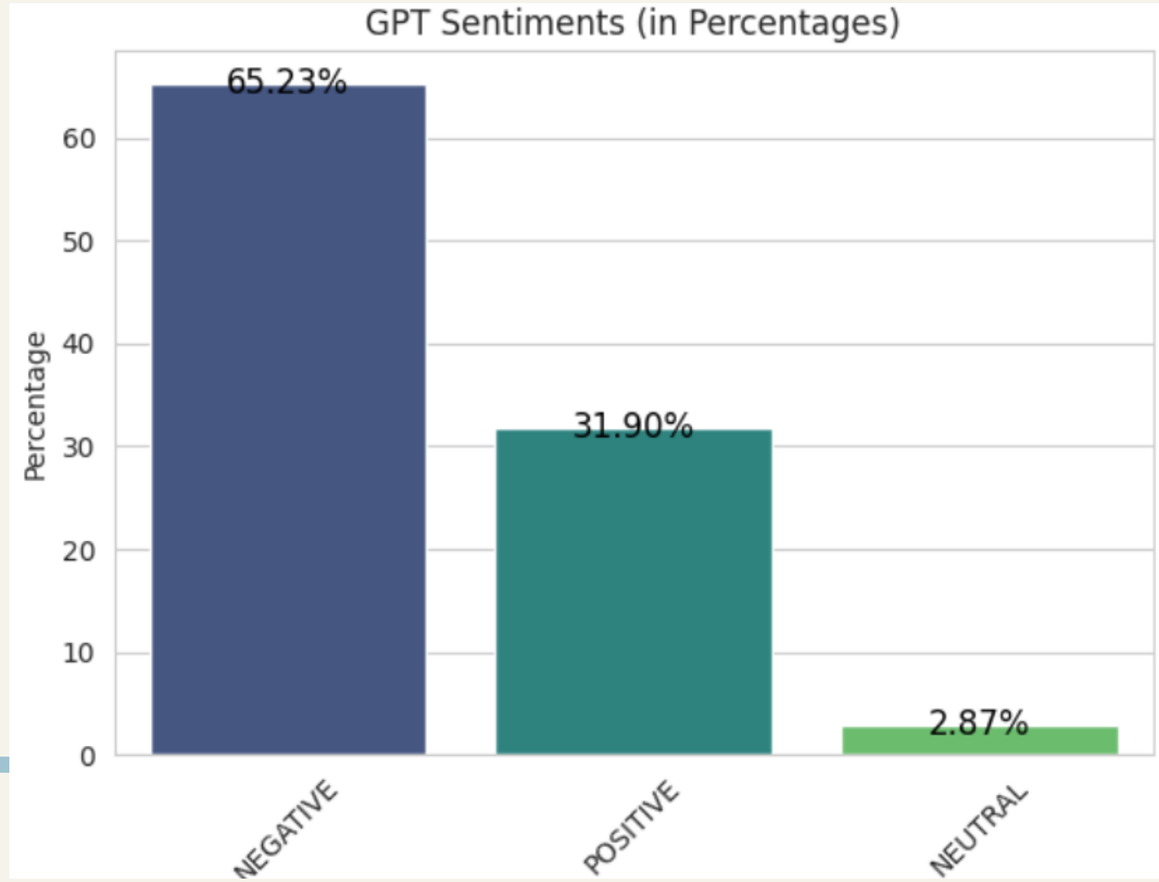
- A primeira avaliação realizada foi quanto ao sentimento das perguntas. Os sentimentos foram classificados em 3 classes: POSITIVO, NEGATIVO, NEUTRO.
- Utilizado um modelo de código aberto aberto disponível no Hugging face.

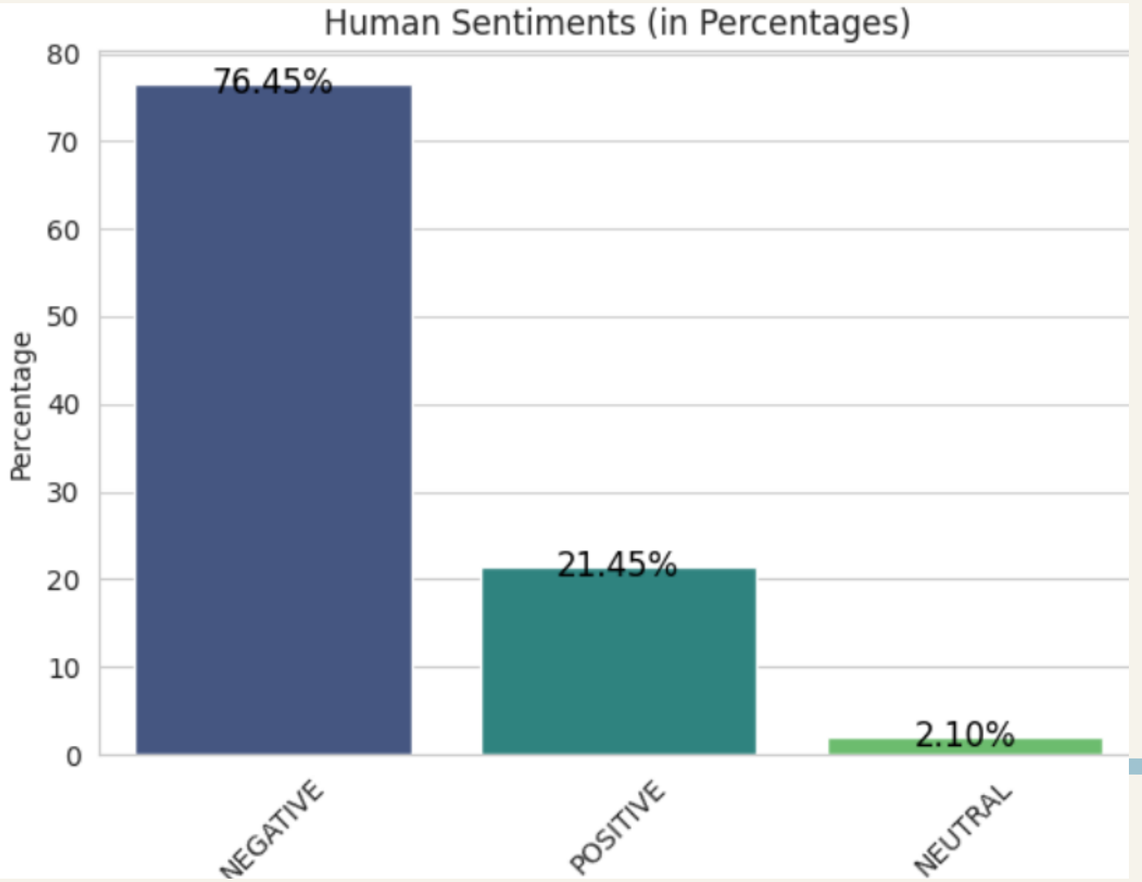




2 - ANÁLISE DE SENTIMENTO

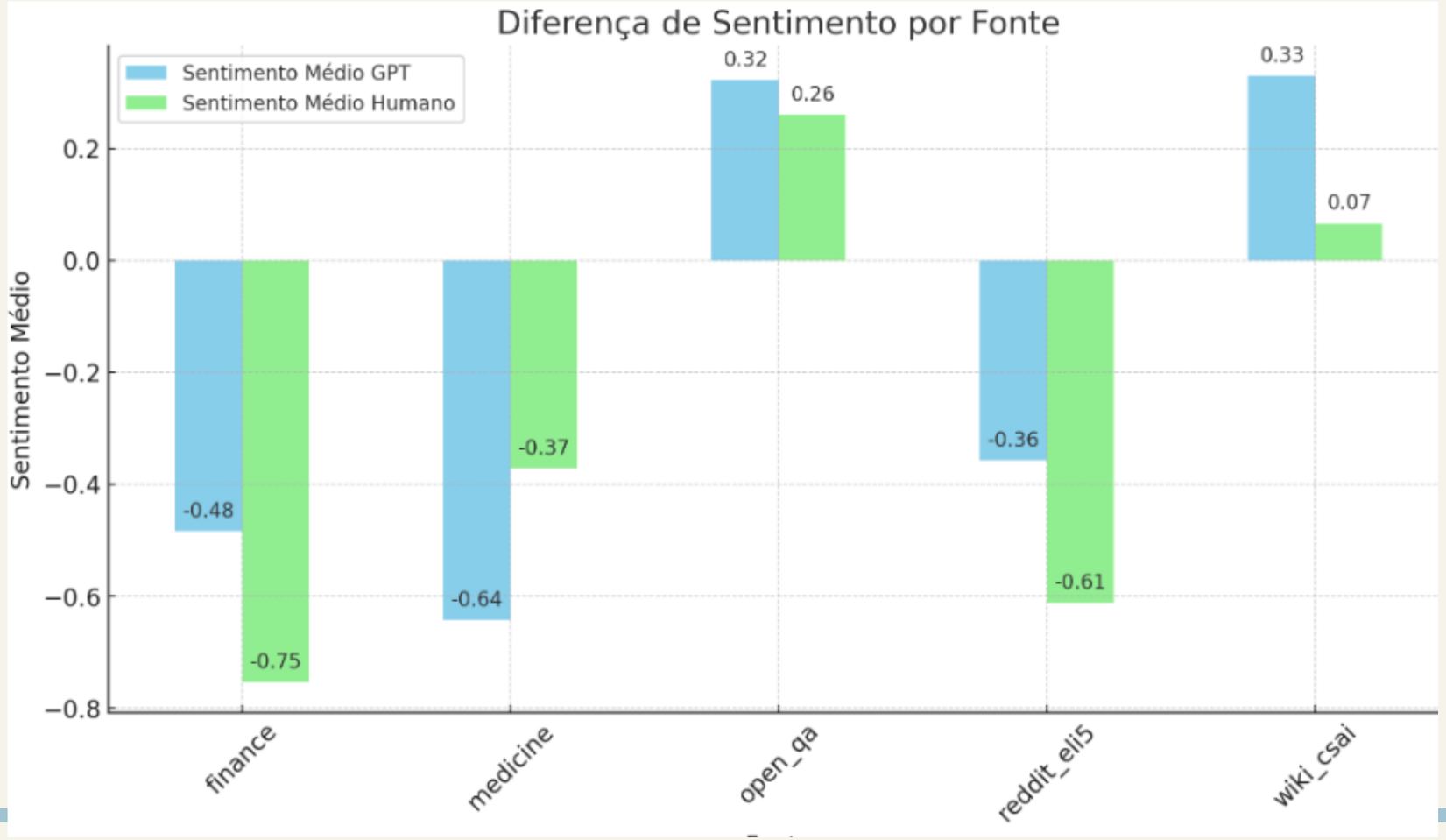
- **Humanos tem mais respostas com sentimentos negativos do que GPT.**
- **Número similar de respostas neutras, em consequência, número de respostas positivas do gpt é maior.**





2 - ANÁLISE DE SENTIMENTO

- O modelo detectou mais positividade do GPT no wiki_csai e mais negatividade do gpt no medicine
- O modelo detectou mais negatividade dos humanos no finance e reddit_eli5



2 - ANÁLISE DE SENTIMENTO

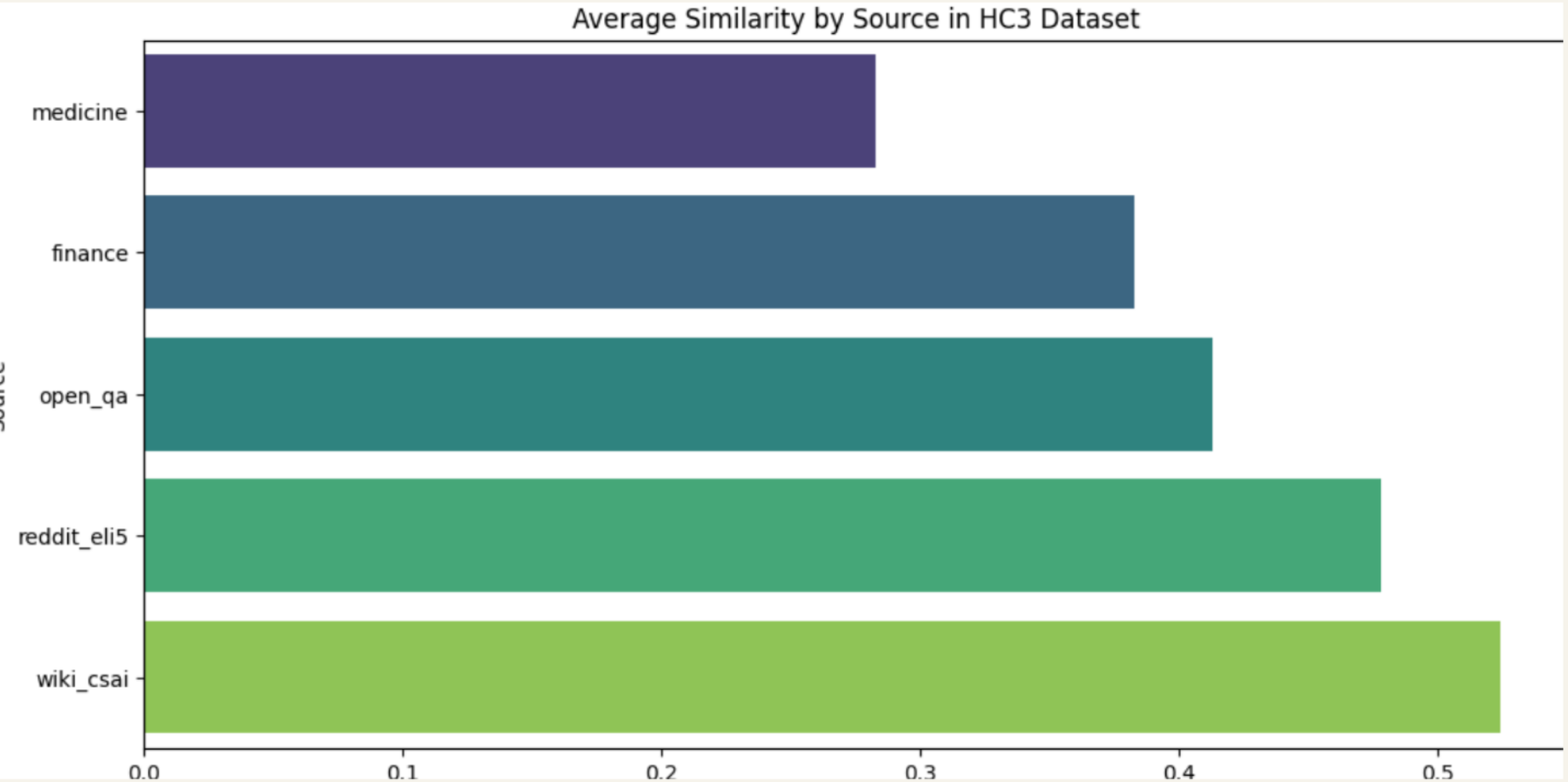
- **Visualização entre tamanho da pergunta vs comprimento da resposta com coloração por sentimento**

2 - ANÁLISE DE SENTIMENTO

- **Visualização entre tamanho da pergunta vs comprimento da resposta com coloração por sentimento**

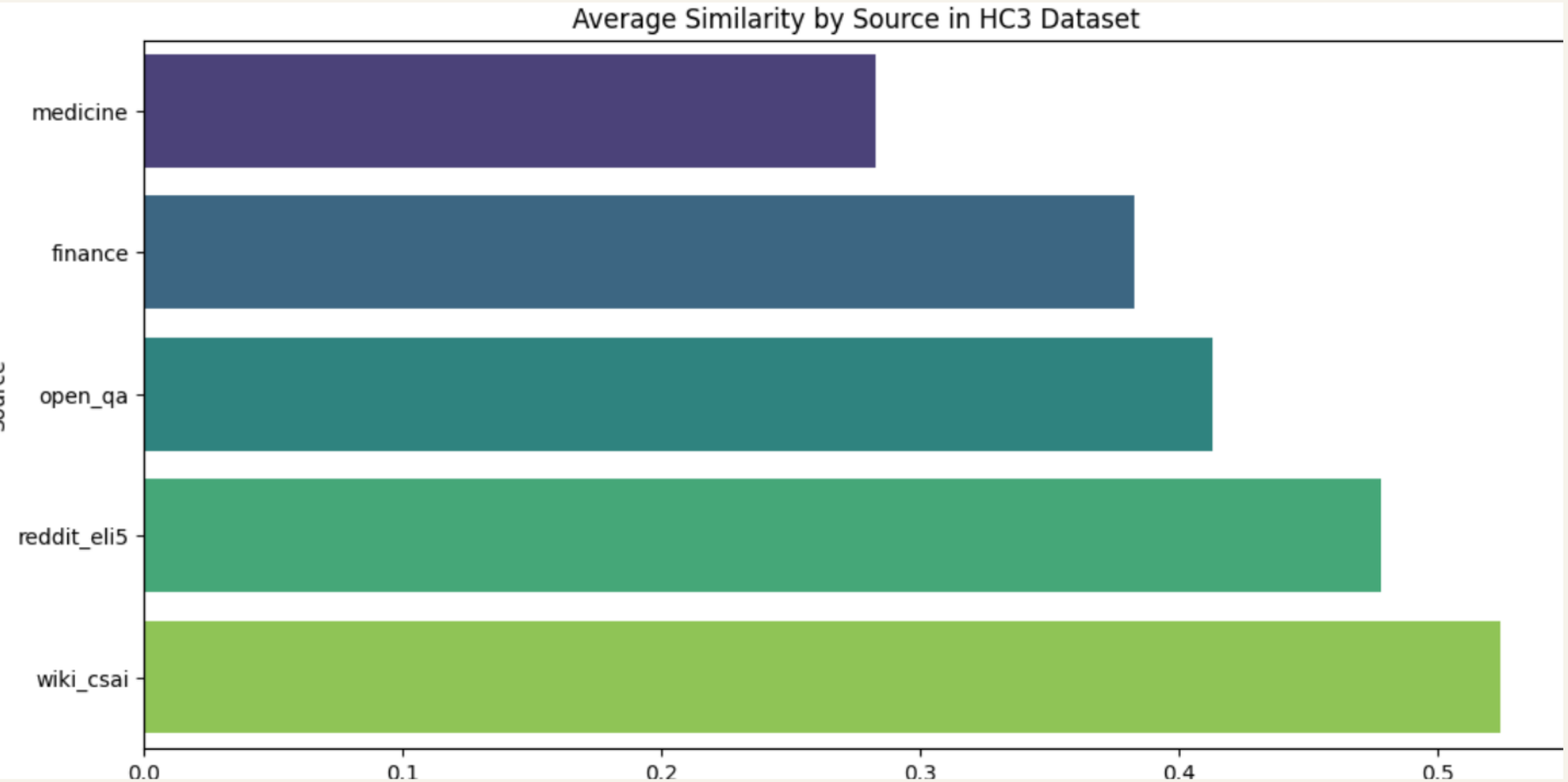
3 - SIMILARIDADE

- Comparação entre a similaridade entre as respostas de gpt e humanos por fonte, valor entre 0 e 1, quanto mais perto de 1, mais similar.
- Modelo disponível pelo Hugging Face.



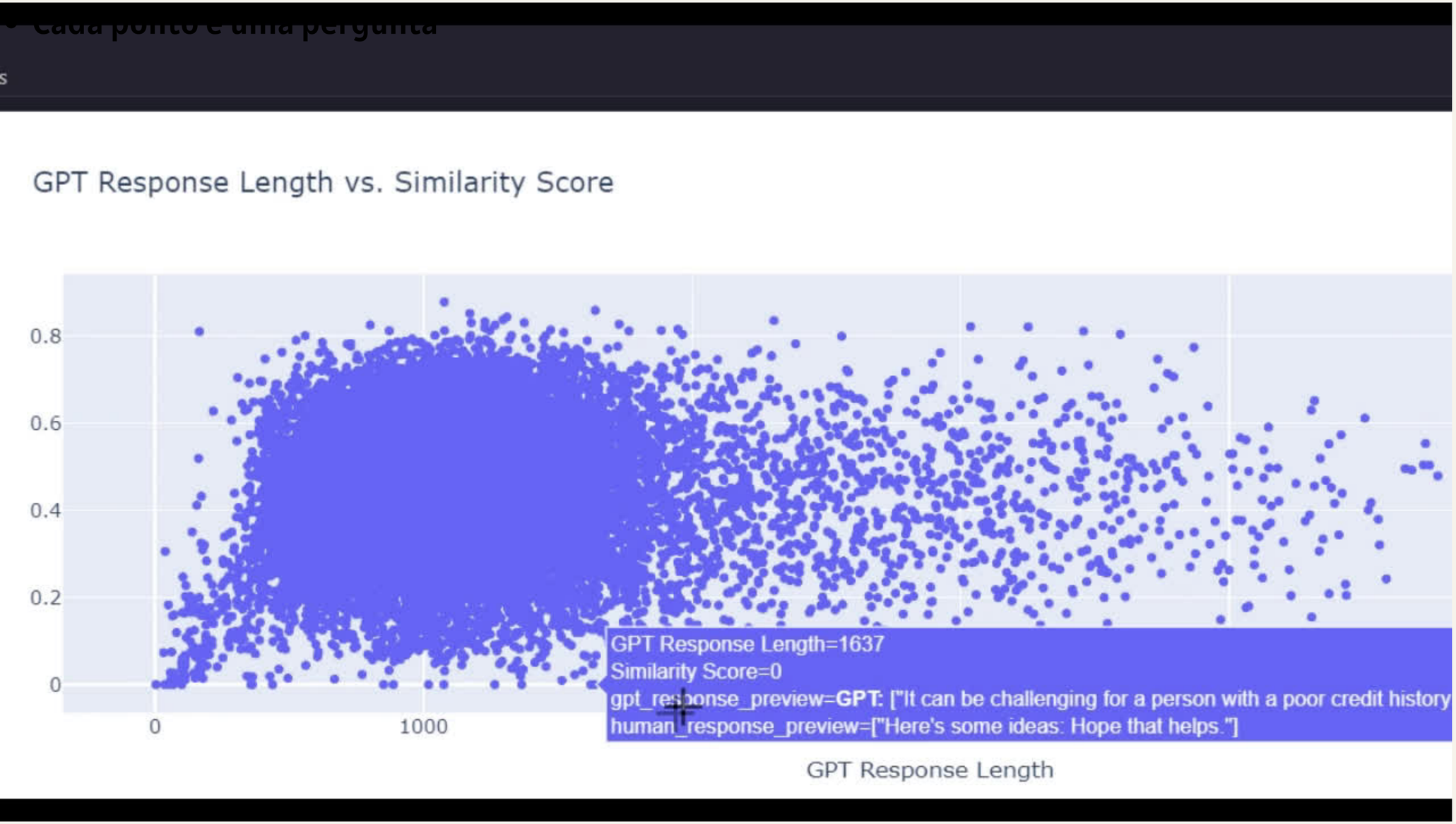
3 -SIMILARIDADE: CONCLUSÕES INICIAIS

- **Medicina é o tópico em que o modelo detectou maior diferença entre humanos e gpt. Similaridade muito baixa, cerca de 30%.**



3 –SIMILARIDADE: VISUALIZAÇÃO COMPRIMENTO DA RESPOSTA DO GPT VS SIMILARIDADE

- **Ferramenta iterativa para ver o tamanho da resposta do GPT por similaridade com resposta humana.**
- **Cada ponto é uma pergunta**



3 -SIMILARIDADE: VISUALIZAÇÃO COMPRIMENTO DA PERGUNTA VS SIMILARIDADE

- Ferramenta iterativa para ver o tamanho da pergunta e a similaridade entre as respostas
- Tendência de diminuição da similaridade conforme o tamanho da pergunta.

3 – SIMILARIDADE: VISUALIZAÇÃO COMPRIMENTO DA PERGUNTA VS SIMILARIDADE

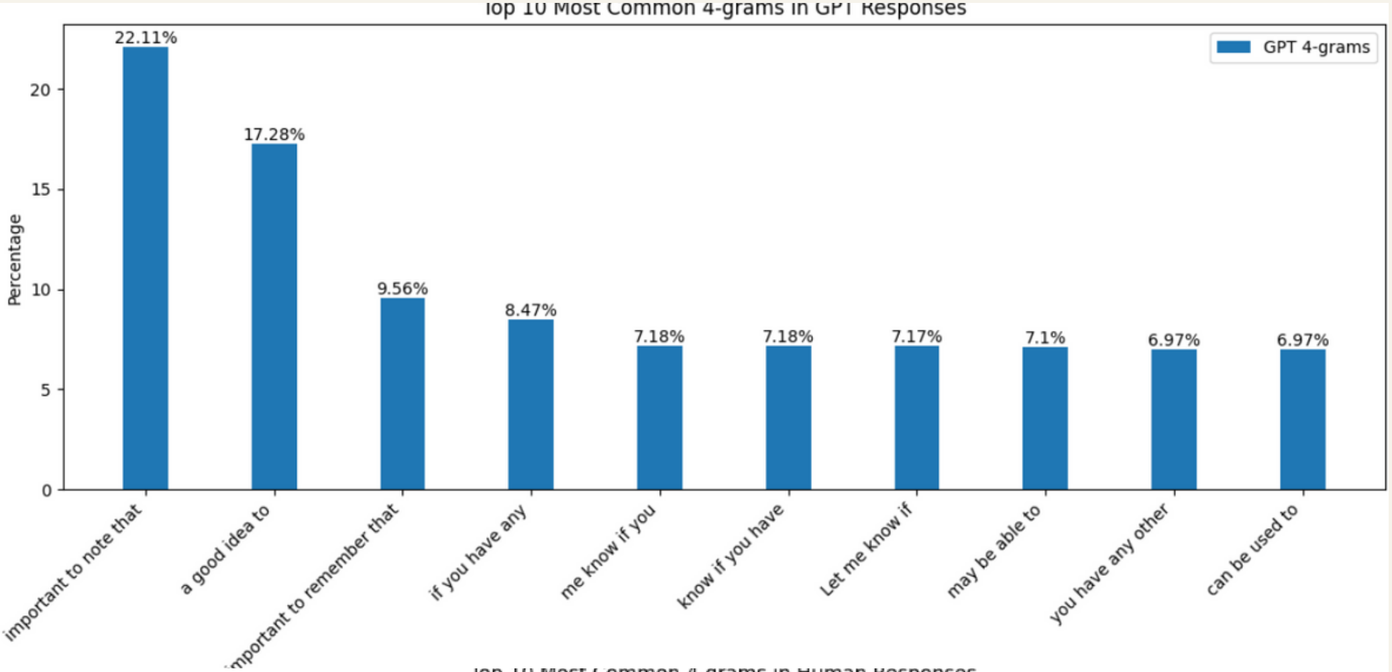
- Ferramenta iterativa para ver o tamanho da pergunta e a similaridade entre as respostas
- Tendência de diminuição da similaridade conforme o tamanho da pergunta.

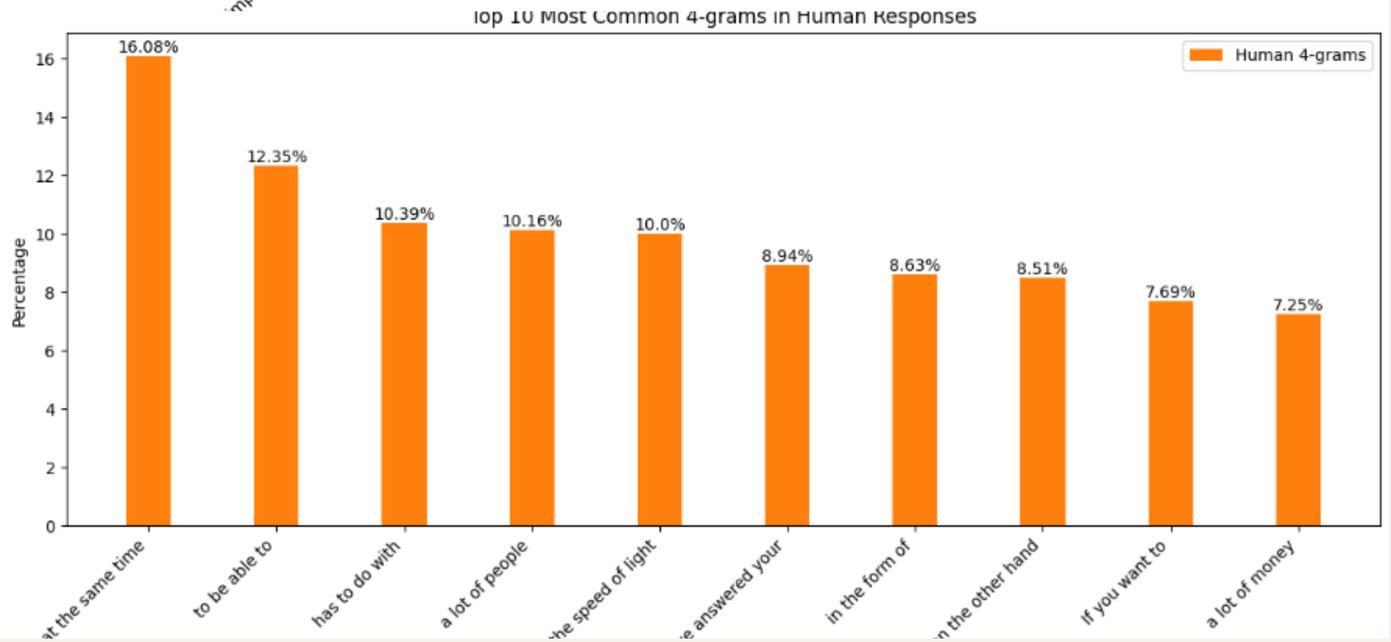
```
df['question_length'] = df['question'].str.len()

# Criar gráfico de dispersão
fig = px.scatter(df, x='question_length', y='similarity', color='source',
                 labels={'question_length': 'Comprimento da Pergunta', 'similarity': 'Similaridade'},
                 title='Similaridade vs Comprimento da Pergunta, Colorido por Fonte')

fig.show()
```

✓ 0.1s Python

lldtte: not installed

4 - EXPRESSÕES MAIS USADAS

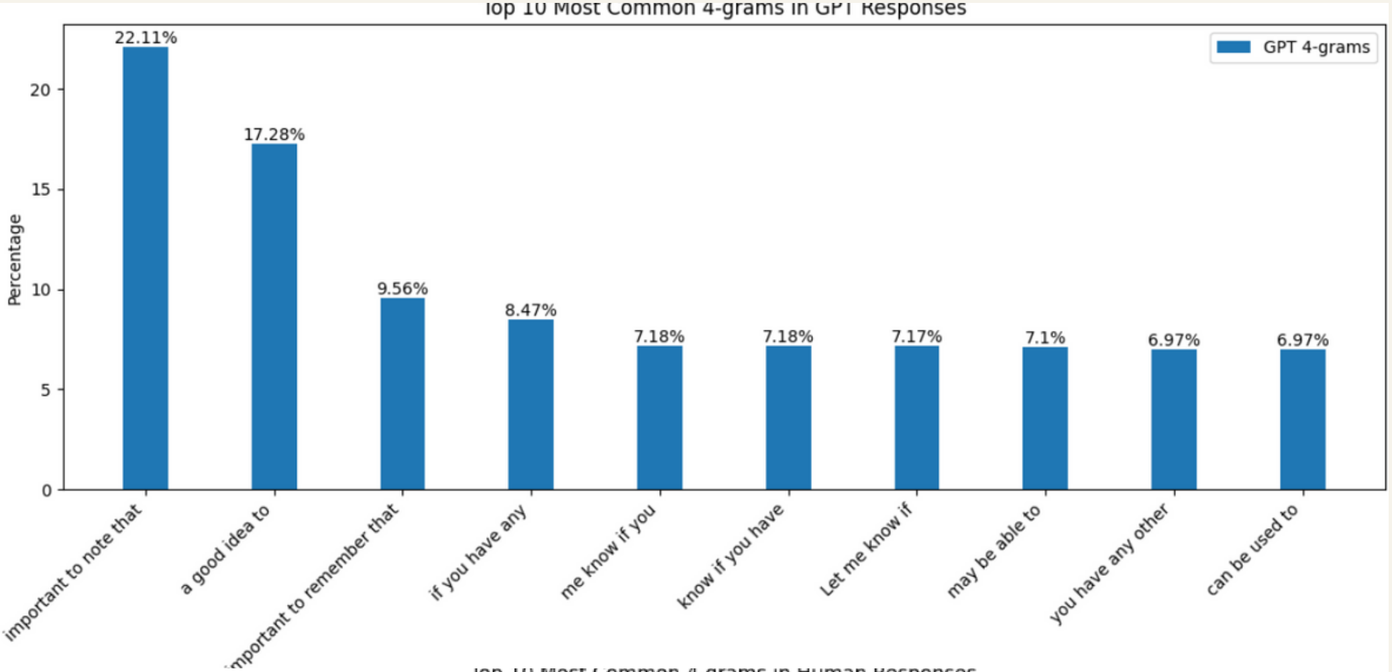


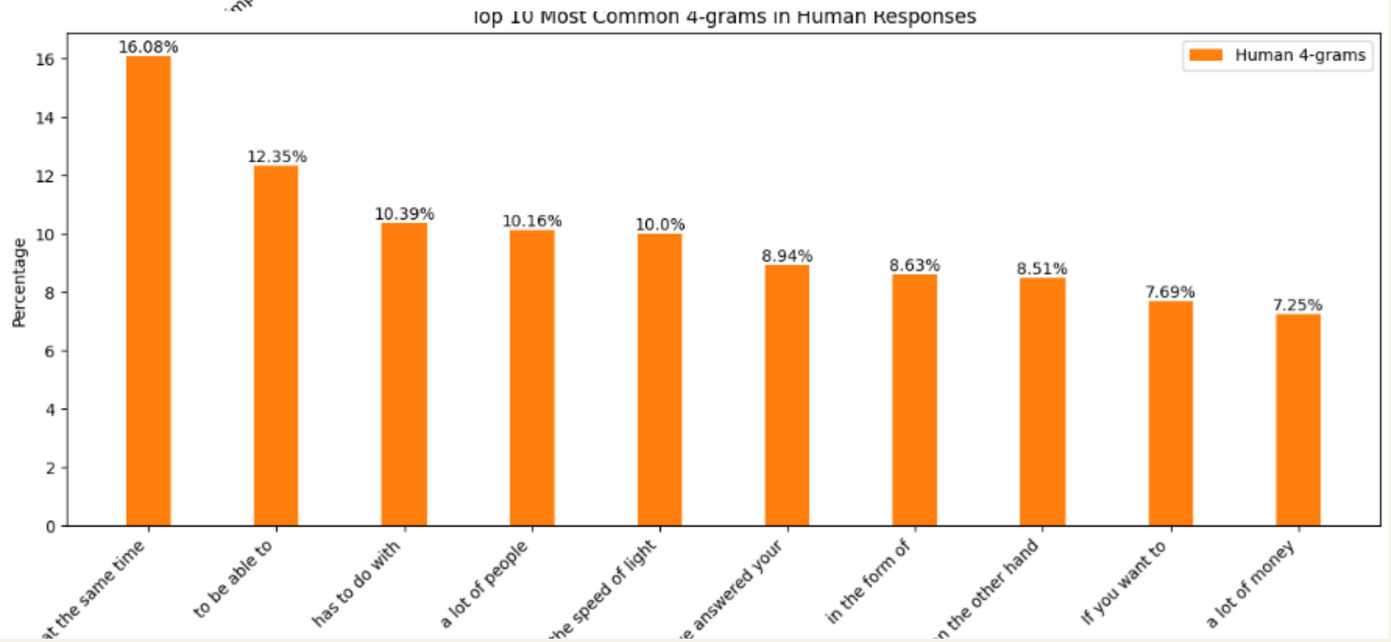


- **Avaliar a frequência de utilização de expressões entre as respostas**

- **Utilizado modelo oferecido pelo Natural Language Toolkit: punkt**

4 - EXPRESSÕES USADAS: CONCLUSÕES INICIAIS



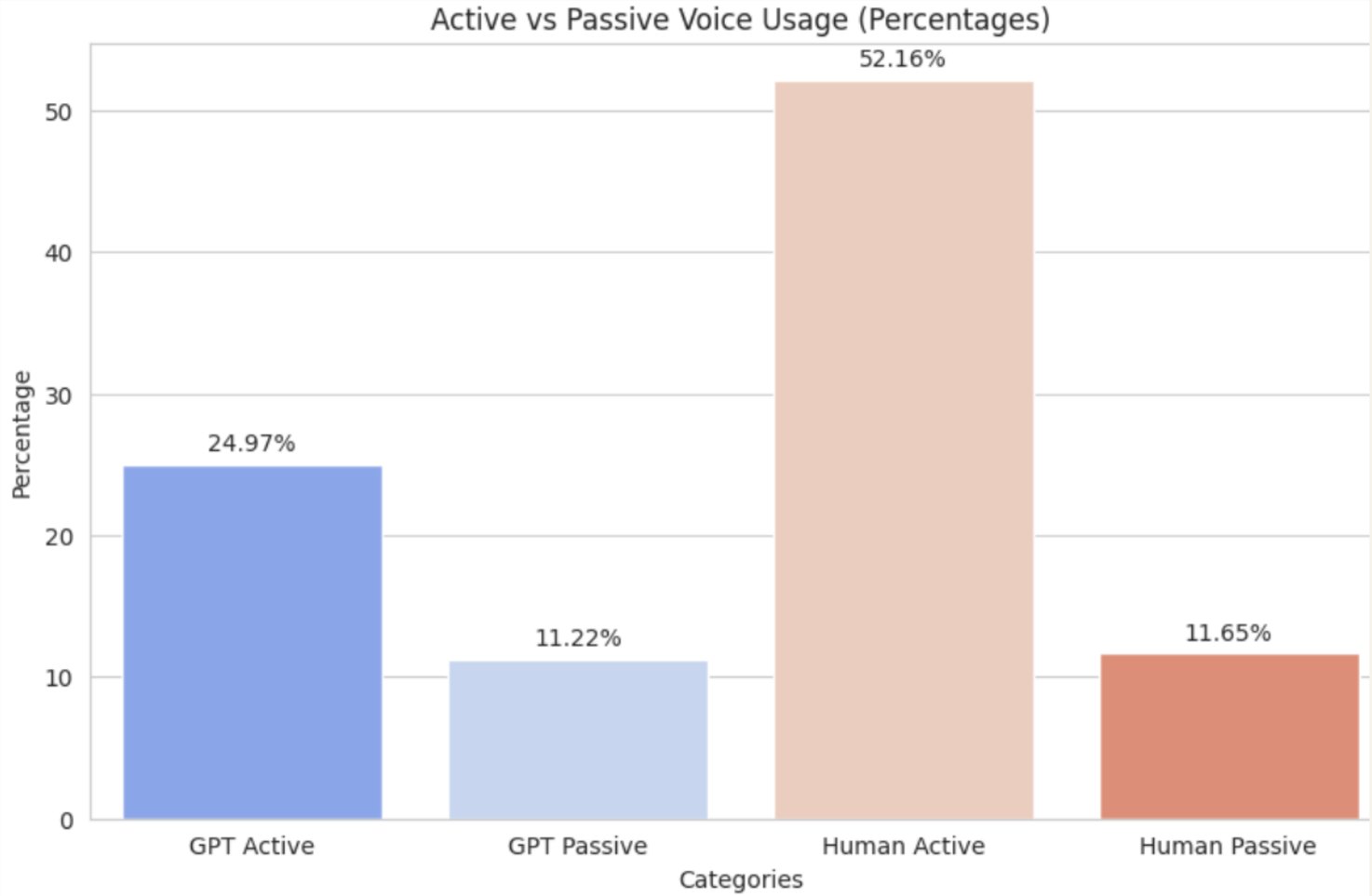


- **GPT utiliza em mais de I/5 dos casos " É importante notar que ". Em um dataset de 23688 respostas, isso é considerável.**

- **Humanos distribuem melhor suas expressões.**

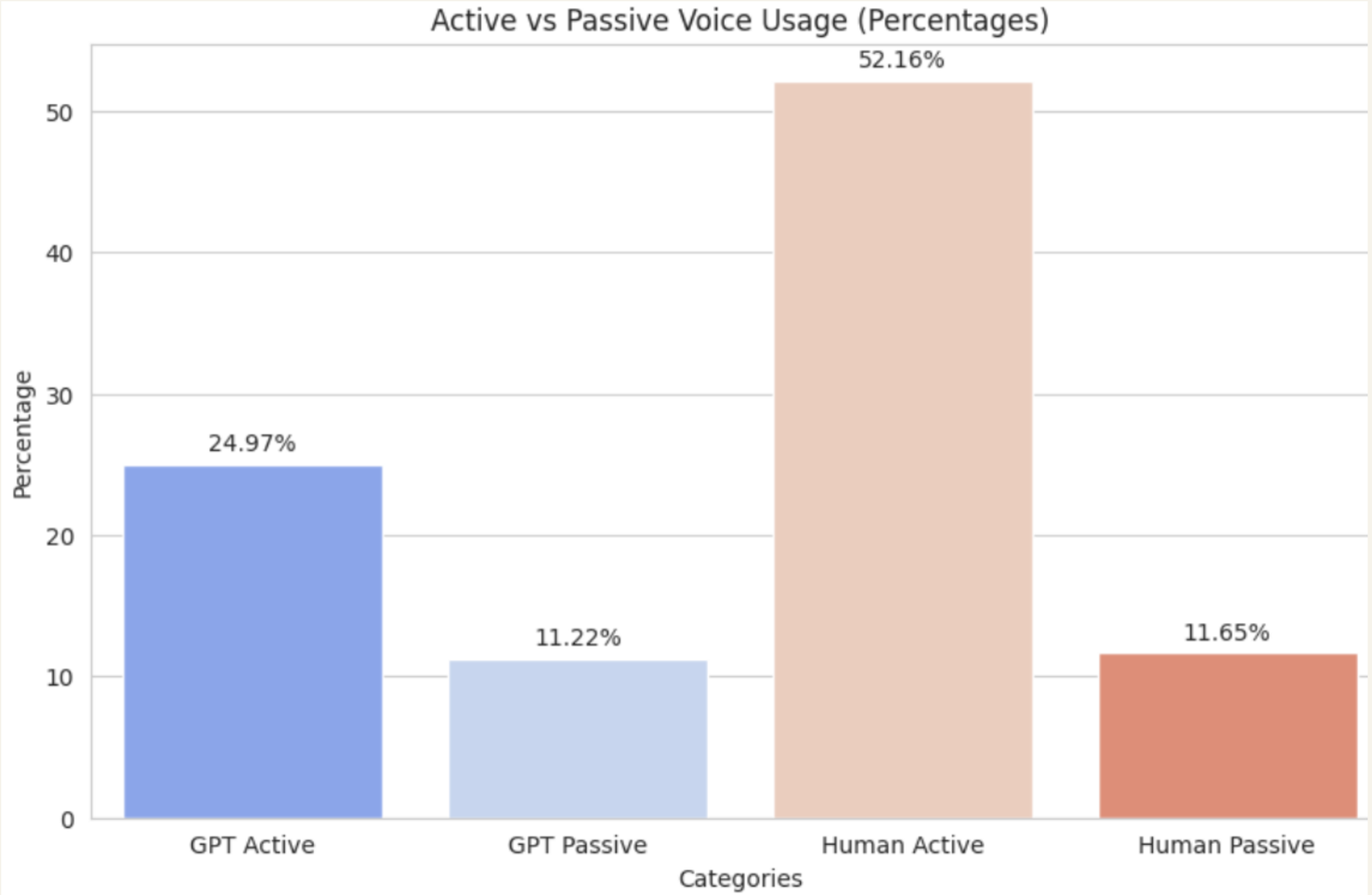
5 - VOZ PASSIVA E ATIVA

- Avaliar as respostas quanto a voz utilizada. Modelo fornecido no hugging face, baseado em BERT.
- Voz ativa: “O gato perseguiu o rato” - Objeto (gato) realiza a ação(persegue)
- Voz passiva: “O rato foi perseguido pelo gato.” - o objeto (o rato) recebe a ação (foi perseguido)



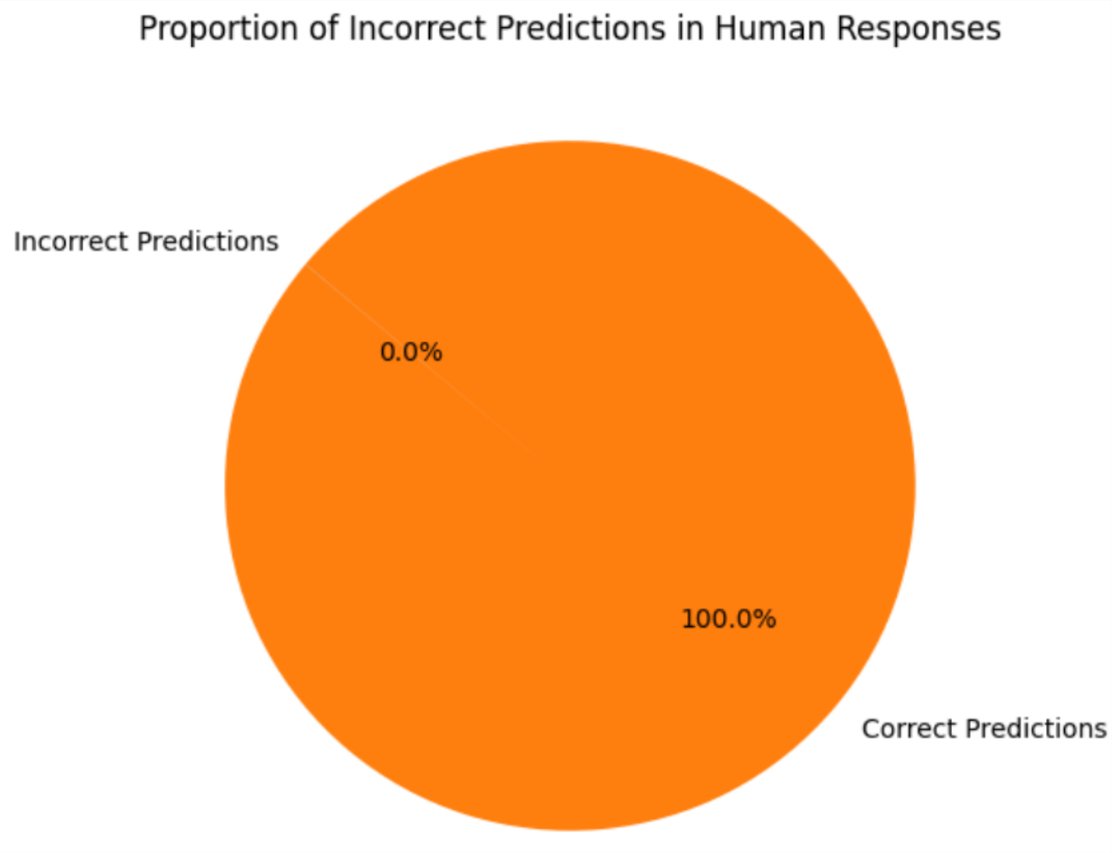
5 – VOZ PASSIVA E ATIVA: CONCLUSÕES INICIAIS

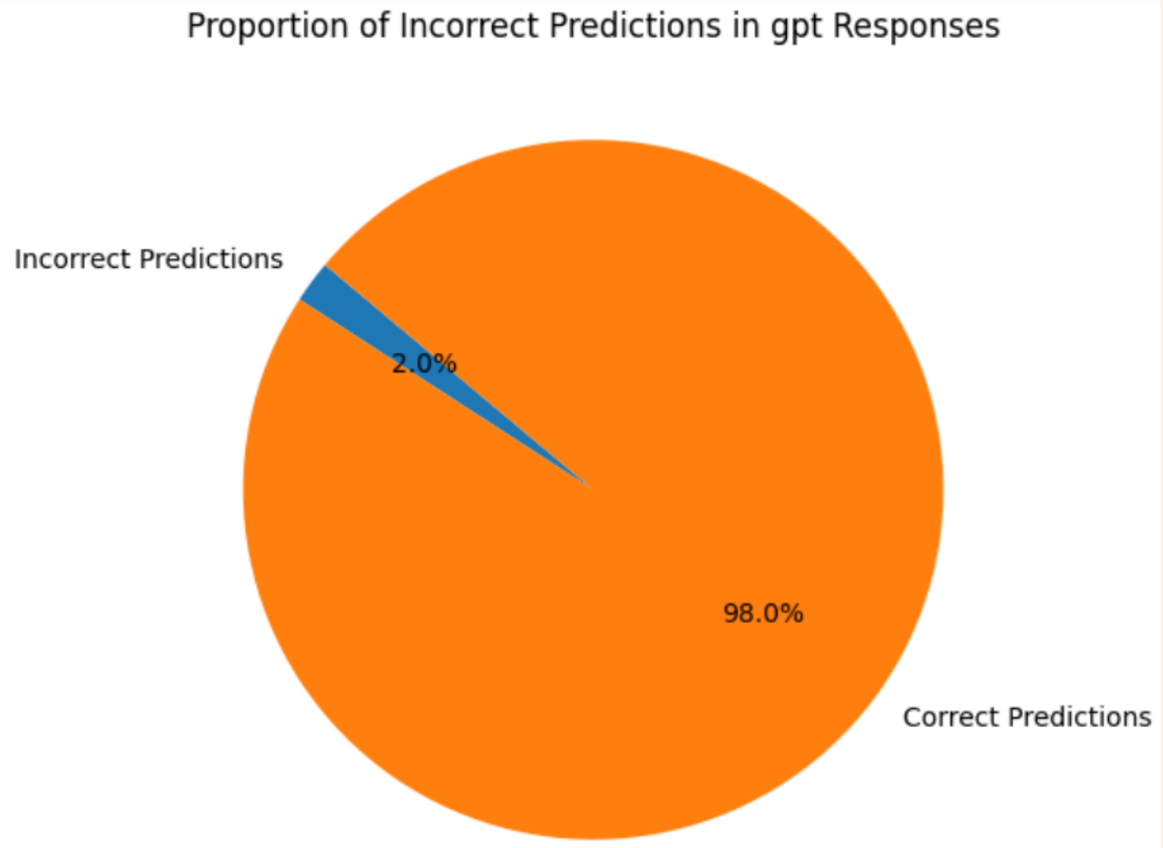
- **Humanos utilizam muito mais da voz ativa que o GPT**
- **A voz passiva em ambos os casos é pouco utilizada**
- **GPT tem mais variancias no discurso, podendo utilizar outras vozes como reciproca, causativa e impessoal**



6 - DETECÇÃO DE TEXTO ESCRITO POR IA

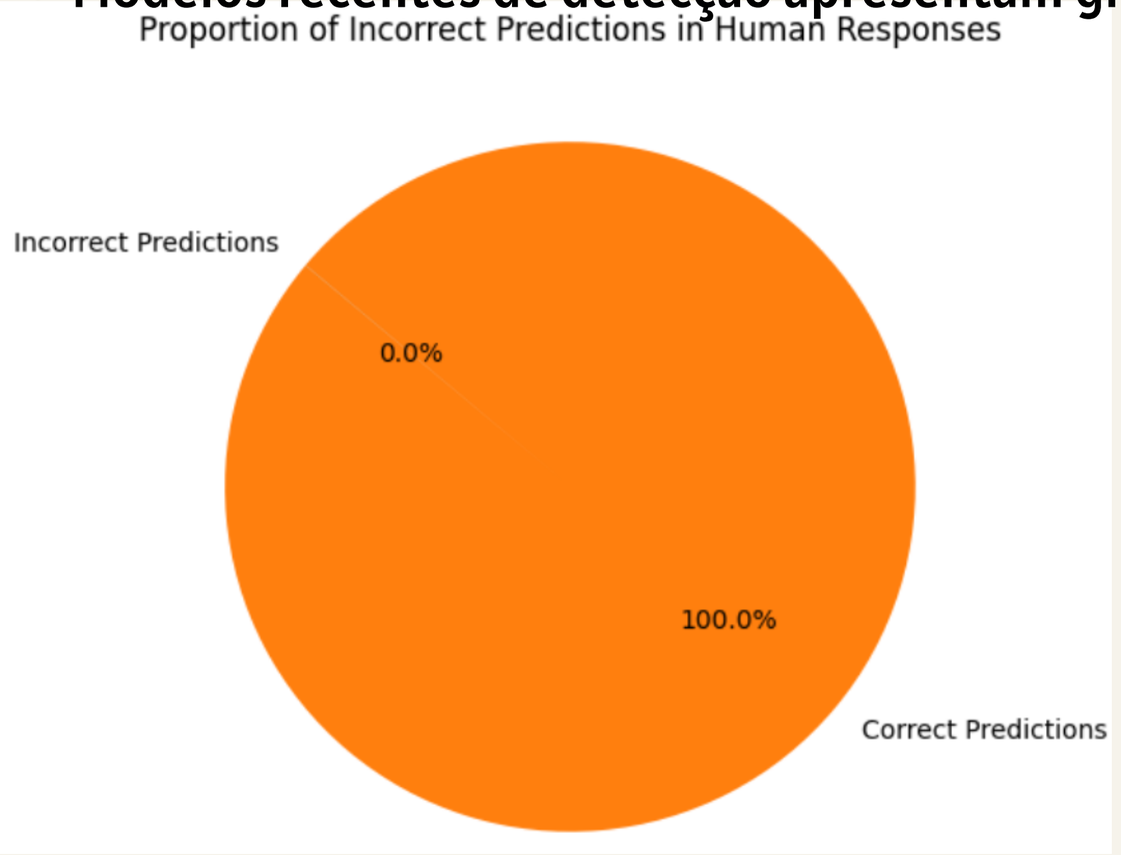
- **Utilizar um modelo recente de detecção de IA baseado em RoBERTa disponível no hugging face para avaliar taxa de erros e falsos positivos/negativos.**

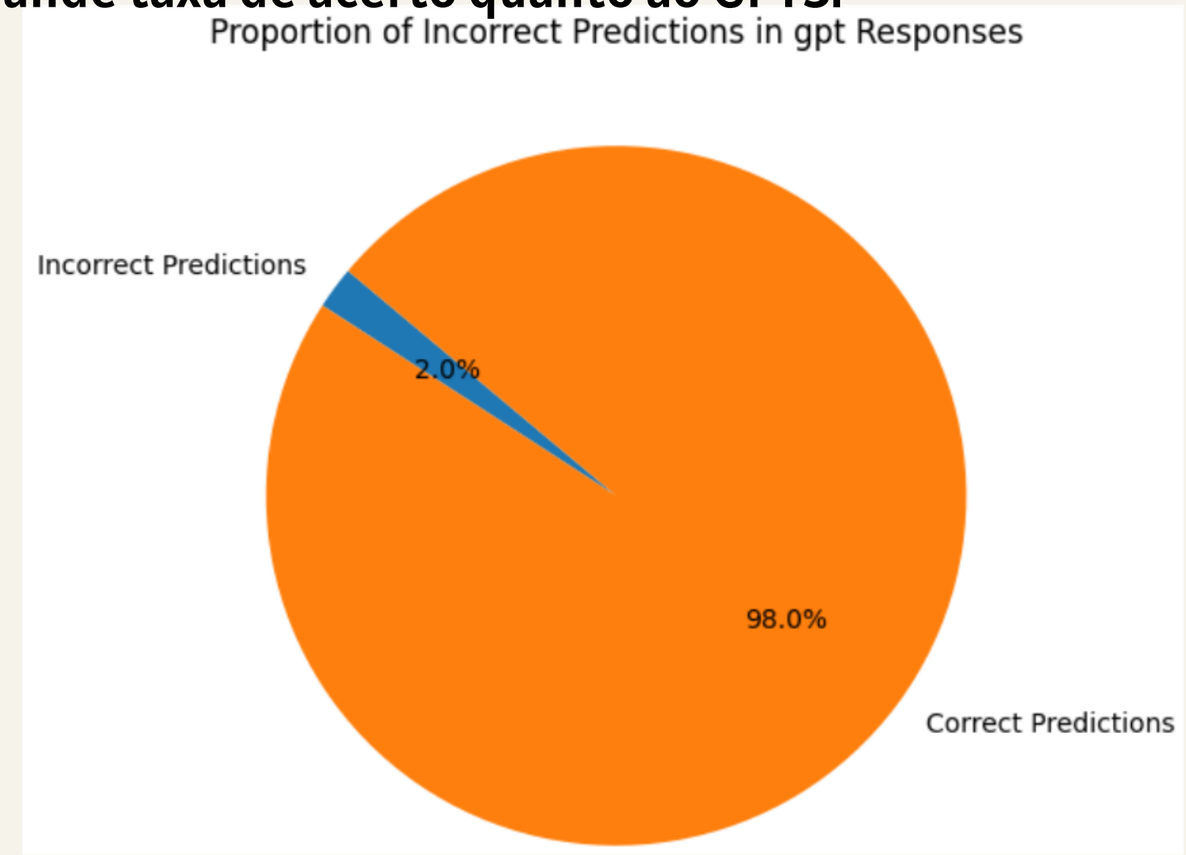




6 - DETECÇÃO DE TEXTO ESCRITO POR IA: CONCLUSÕES INICIAIS

- **O modelo acertou que em 100% dos casos o testo escrito por humanos foi realmente escrito por humanos.Em 2% dos casos de respostas do GPT detectou o texto sendo escrito por humanos quando na verdade foi escrito por IA.**
- **Modelos recentes de detecção apresentam grande taxa de acerto quanto ao GPT3.**



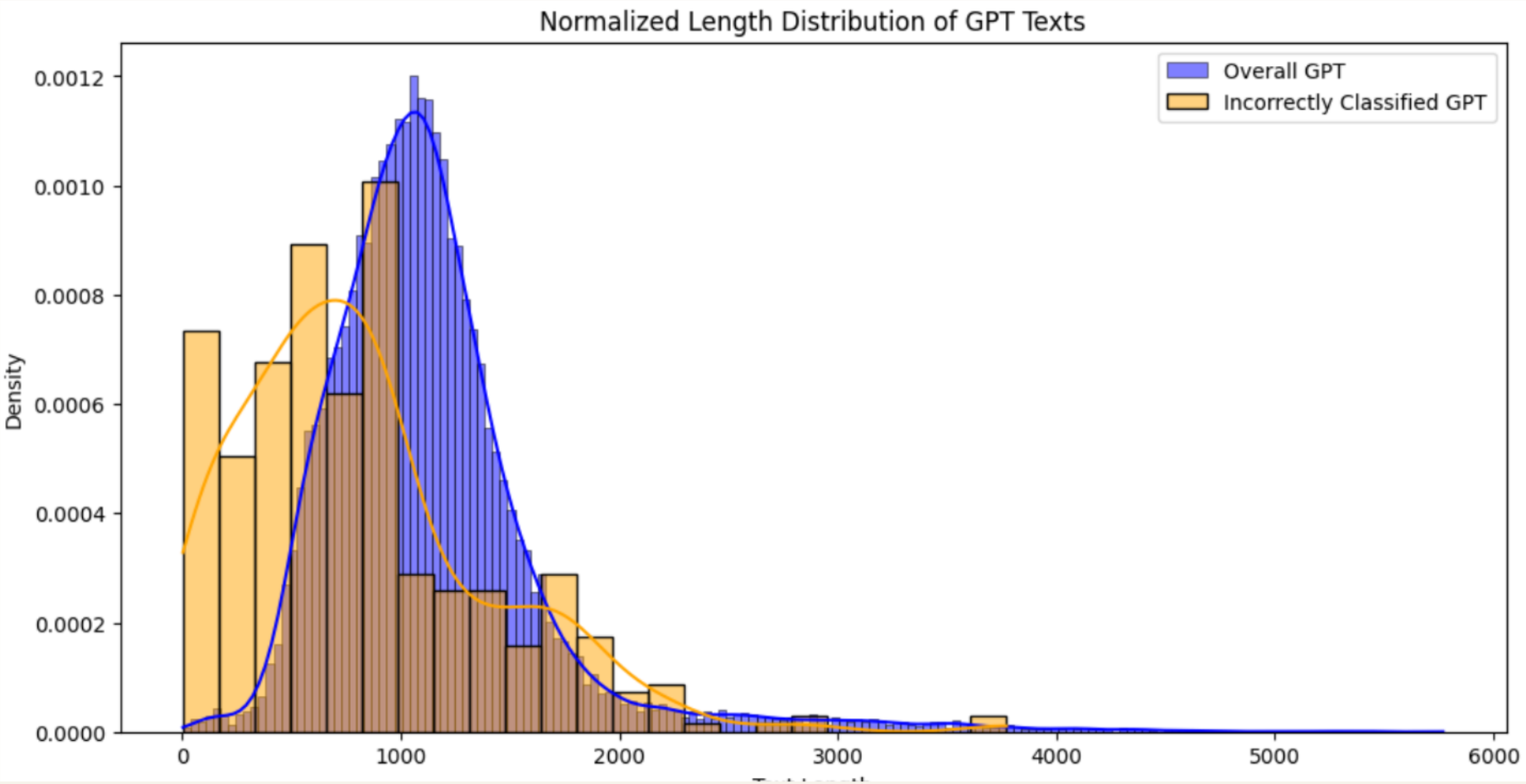


6- DETECÇÃO DE TEXTO ESCRITO POR IA:VISUALIZAÇÃO POR SIMILARIDADE

- **Visualização das respostas em que a diferença de probabilidade de serem feitas por IA foi pequena**

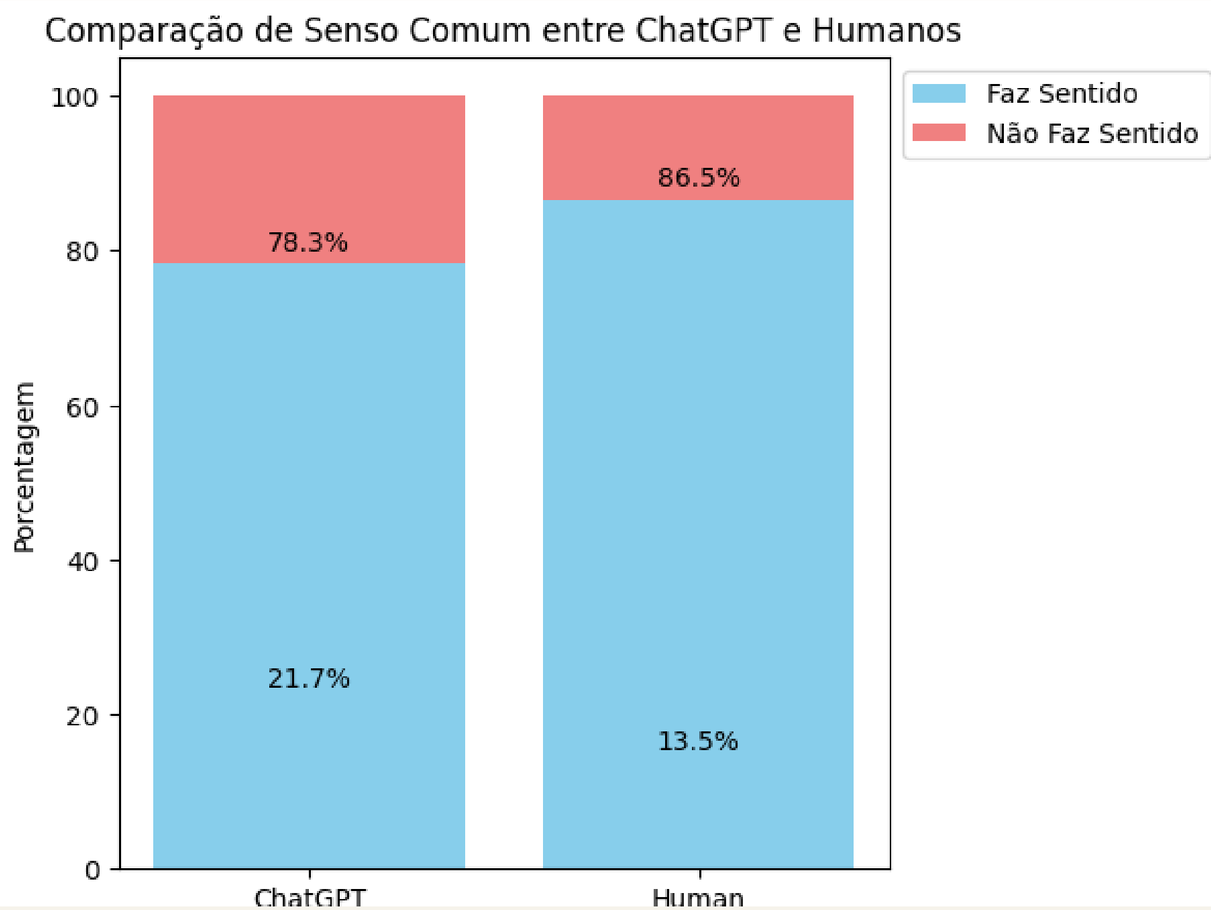
6 - DETECÇÃO DE TEXTO ESCRITO POR IA: AVALIAÇÃO DO ERRO

- **Nos 2% que o modelo detectou como escrito por humanos quando na verdade foi escrito pelo chatgpt percebemos que a distribuição do comprimento dos textos é menor que da média do conjunto todo de respostas do GPT.**



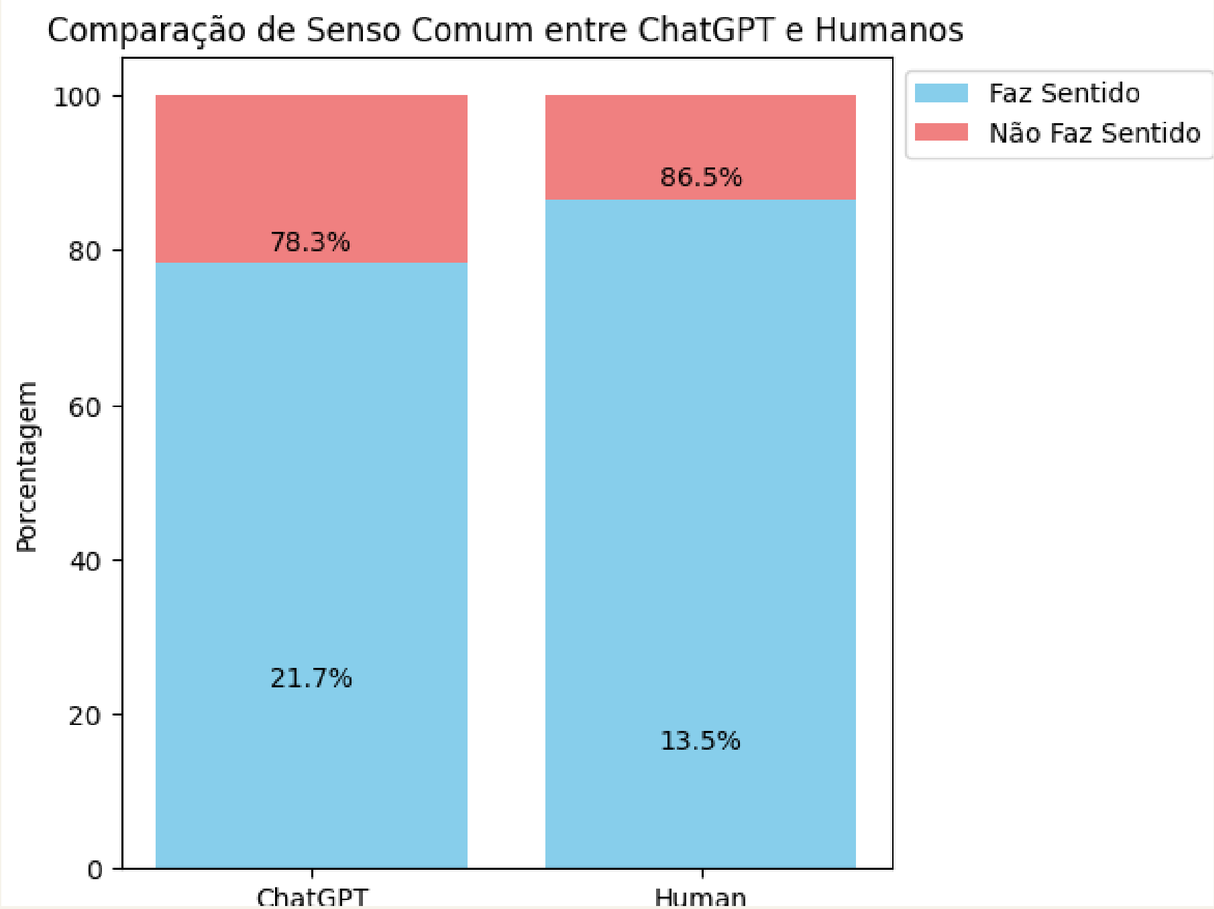
7 - RELAÇÃO DA RESPOSTA AO SENSO COMUM

- Avaliação em um modelo baseado em GPT2 disponível no hugging face
- Detecta se uma resposta faz sentido com relação ao senso comum
- Por exemplo: o gato subiu na árvore (faz sentido em relação ao senso comum, score = 1) / o elefante subiu na árvore (não faz sentido com relação ao senso comum, score = 0)



7 – RELAÇÃO DA RESPOSTA AO SENSO COMUM

- **GPT aparenta desviar mais do senso comum que os humanos.**



7 – RELAÇÃO DA RESPOSTA AO SENSO COMUM : CONCLUSÕES INICIAIS

- **Visualização de senso comum por comprimento da pergunta**

7 - RELAÇÃO DA RESPOSTA AO SENSO COMUM : CONCLUSÕES INICIAIS

- Visualização entre similaridade e senso comum, observando as diferenças entre as frases
- Observando as maiores diferenças entre probabilidade de senso comum para as mesmas respostas

CONCLUSÕES SOBRE O DATASET

- GPT tem tendência de diminuir o tamanho da resposta com o tamanho das perguntas
- GPT tem respostas com sentimentos mais positivos que os humanos
- Humanos distribuem melhor sua forma de se expressar que o GPT
- Humanos fazem mais sentido em relação ao senso comum que o GPT.

CONTRIBUIÇÕES

- Visão geral sobre as comparações entre respostas humanas e do GPT
- Gráficos iterativos que permitem buscar padrões em frases para estudos futuros

OBRIGADO :)

- Códigos e referências disponíveis em github.com/giovanicenta