Evaluations Document

Overview

This document provides a comprehensive evaluation of the performance of two prominent large language models, DialogGPT and LlaMA-2, within the context of character-specific chatbots. The evaluation focuses on key metrics, including perplexity and BLEU scores. Furthermore, fluency and relevancy are visually represented through pie charts derived from manual assessments conducted by a diverse population of 500 individuals.

Methodology

1. Perplexity Evaluation:

- Perplexity serves as a measure of how well the language model predicts a sample. Lower perplexity values indicate better performance.
- Results are presented in a table format, comparing values obtained from both DialogGPT and LlaMA-2.

2. BLEU Scores Assessment:

- BLEU scores quantify the similarity between the model-generated responses and human-generated references.
- The evaluation includes a table comparing BLEU scores for both models, offering insights into their linguistic quality.

3. Fluency Visualization:

- Fluency, a qualitative measure, is visually represented through pie charts. Participants rated the naturalness and coherence of responses.
- Detailed charts provide an intuitive understanding of how DialogGPT and LlaMA-2 compare in terms of fluency.

4. Relevancy Visualization:

- Relevancy, another qualitative aspect, is also depicted using pie charts. Participants assessed the appropriateness of responses.
- The visual representation aids in comparing the relevancy of outputs from both models.

Results

1. Perplexity and BLEU Scores:

- Table 1: Perplexity and BLEU scores for DialogGPT and LlaMA-2.

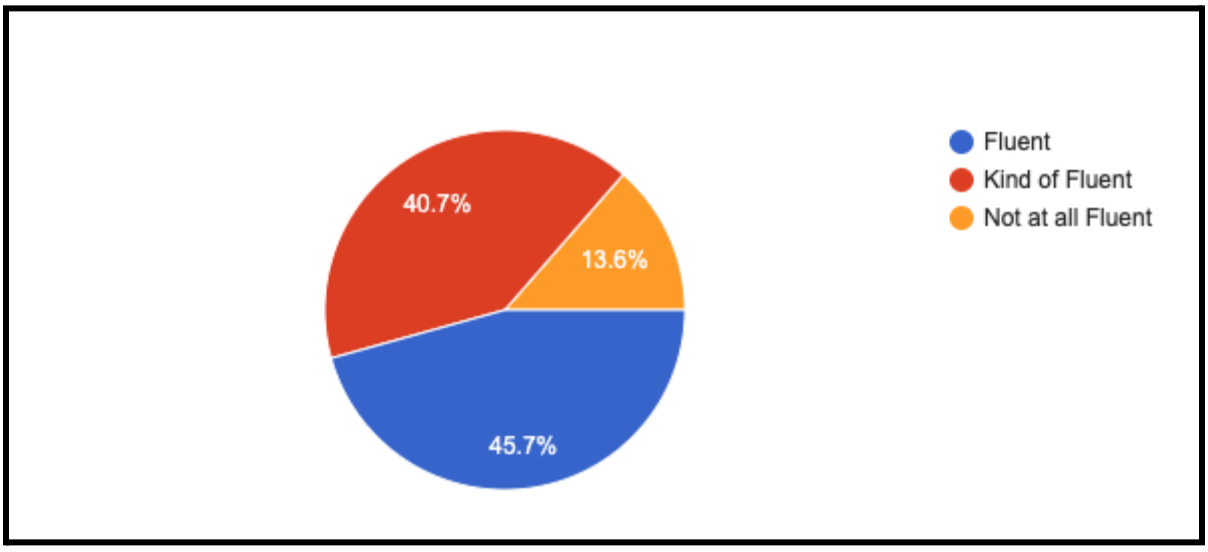
DialogGPT	Rick and Morty	Big Bang Theory
Perplexity	3.0322	3.0448
BLEU score	0.5027	0.5020

Llama - 2	Rick and Morty	Big Bang Theory
Perplexity	19.323	17.612
BLEU score	0.8978	0.5223

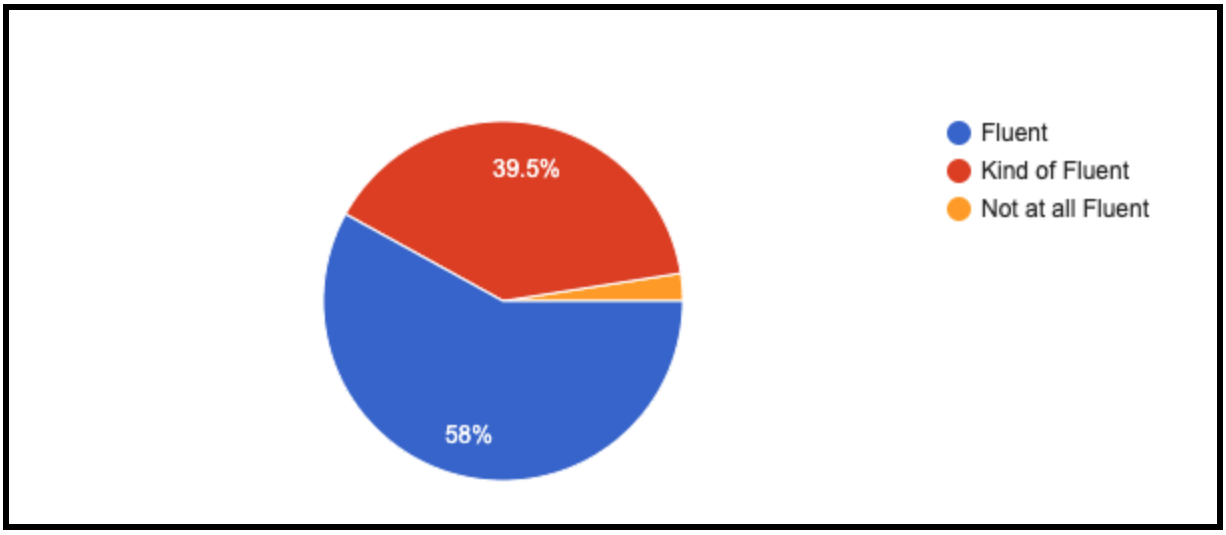
2. Fluency Visualization:

- Chart 1: Fluency comparison between DialogGPT and LlaMA-2.

  - Fluency of Dialog GPT Model on both Dataset



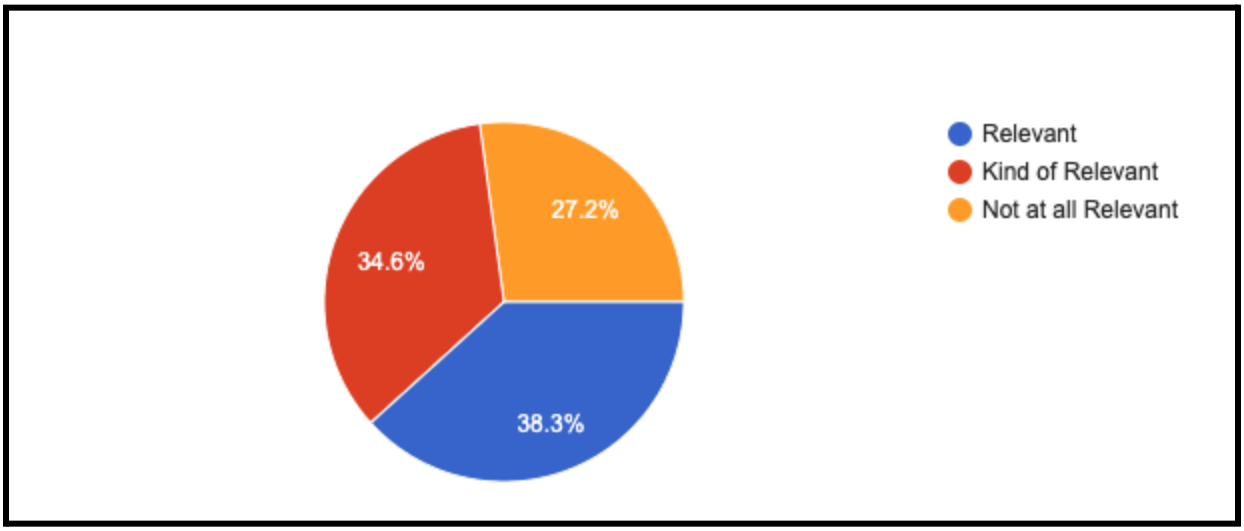
- Fluency of Llama - 2 Model on both Dataset



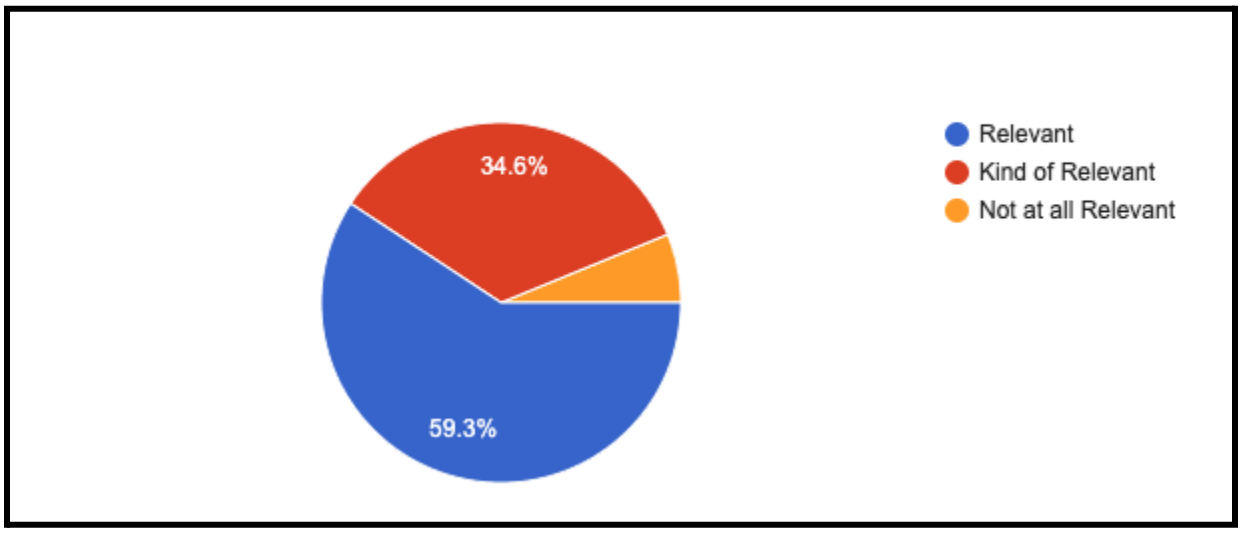
3. Relevancy Visualization:

- Chart 2: Relevancy comparison between DialogGPT and LlaMA-2.

- Relevancy for DialoGPT Model on both Dataset



- Relevancy of Llama-2 model for both Dataset



Conclusion

This evaluation provides a nuanced understanding of the strengths and weaknesses of DialogGPT and LlaMA-2 in the context of character-specific chatbots. The combination of quantitative metrics and qualitative assessments offers a holistic view, aiding stakeholders in informed decision-making for model selection and deployment.