**Methodology**

This section outlines the step-by-step process used in this research, covering both data cleaning and the application of Natural Language Processing (NLP) libraries and evaluation metrics.

1. Data Cleaning Process:

The dataset preparation began by consolidating responses from each chatbot into a single Excel sheet. This consolidation aimed to make coding easier. Subsequently, we carefully removed rows with missing responses from any chatbot. This was crucial to ensure that every question in the dataset had responses from all three chatbots, ensuring the dataset's completeness and reliability for analysis.

2. Natural Language Processing (NLP) Libraries and Evaluation Metrics:

We utilized the Python Natural Language Toolkit (NLTK) for analyzing chatbot responses. Specifically, we calculated BLEU and METEOR scores using NLTK, providing a solid foundation for evaluating response quality.

For ROGUE scores, we chose the rogue-score library for its smoother performance compared to the implementation in NLTK. The ROGUE score, including components like 'rouge1', 'rouge2', and 'rougeL', played a vital role in evaluating semantic similarity. In our analysis, we took the average F-measure across these components as the representative ROGUE score, offering a nuanced evaluation of chatbot responses.

3. Thorough Comparison of Responses:

To comprehensively assess responses, we systematically compared responses from each chatbot to the others:

* ChatGPT vs Google Bard
* ChatGPT vs Microsoft Bing
* Microsoft Bing vs Google Bard

The overall similarity score for each chatbot was determined by averaging ROGUE, METEOR, and BLEU scores for the respective pair. This method allowed for a holistic evaluation of chatbots, considering linguistic quality, semantic similarity, and overall performance of each metric used.