Ben Heinze, Braxton McCormack, Michael Hagin

Research Question
Has the diction in academic research papers within the computer science community changed after January 2023 at a faster rate than usual as a result of the potential assistance of large language models (LLMs)?

Project Proposal Outline
Overview of the Problem: Investigate whether the integration of LLMs in academic writing has altered the language and style of academic papers quicker than the previous rate of change..

Stakeholders:
Academic Researchers: Insights into the impact of LLMs on scholarly writing could influence how they approach their own writing and publication strategies.
Journal Editors and Reviewers: Understanding these changes could assist in developing guidelines or detection tools for LLM-generated content.
Academic Institutions: Insights could help in shaping policies regarding the use of AI in academic work.

Dataset Summary:
Data Sources: Papers published in selected academic journals before and after January 2023. Selected journals will have a range of impact factors and papers will be published no earlier than 2018.
Attributes: Textual data from abstracts and main content, publication date.
Preprocessing Needs: Tokenization, stopword removal, lemmatization.

Data Mining Techniques:
Natural Language Processing (NLP): Techniques to analyze text features, including tokenization, stopword removal, lemmatization, and part-of-speech tagging.
Statistical Analysis: Compare frequencies of linguistic features found in academic papers predating LLMs against frequencies of linguistic features postdating LLMs over time.

Evaluation Strategy:
Quantitative Metrics: Use Term Frequency (TF) and Term Frequency by Inverse Document Frequencies (TF-IDF) to measure significant differences in diction between periods.

Future Work Acknowledgment:
- Increase the number of research papers analyzed.
- Comparing/contrasting larger ranges of differently ranked academic journals, as well as analyzing them collectively.

- Attempting to classify LLM-generated words to the LLM that generated them. (It's important to keep in mind that just because a word is more frequent with AI, that does not conclude the word was generated with it. Exposure to AI-assisted research papers may influence the verbatim readers use for their own papers.)

- Gemeni, Claude, and GPT were released/popularized at different times. Instead of a simple before/after, it would be nice to work with a larger sample size and examine more regular intervals.