Section IV: Discussion

The primary objective of this project was to develop a data cleaning technique to improve the quality of Large Language Model (LLM) datasets, specifically focusing on mitigating the identified risk assessment variable, sentence completion, within the WikiText dataset. Through analysis and iterative development, the proposed data cleaning technique effectively addressed these variables, resulting in a significant improvement in the performance of the trained GPT-2 model.

The analysis of risk assessment variables such as sentence completion, readability, and lexical diversity provided valuable insights into the quality of the WikiText dataset. By leveraging tools like BERT sequence classification for sentence completion, Flesch-Kincaid Grade Score for readability, and Moving Average Type-Token-Ratio for lexical diversity, we were able to identify areas of improvement within the dataset.

Despite the success of the project, several limitations and confounding variables should be acknowledged. Firstly, the effectiveness of the data cleaning technique may vary depending on the specific characteristics of different datasets, and further validation on diverse datasets would strengthen the generalizability of the findings. Additionally, the reliance on pre-existing metrics for risk assessment variables introduces the potential for biases inherent in those metrics, which could influence the interpretation of results.

This project contributes to the field of Natural Language Processing (NLP) by providing a practical and effective approach to improving the quality of LLM datasets through data cleaning techniques. By addressing common risk assessment variables such as sentence completion, readability, and lexical diversity, this research enhances the reliability and utility of LLMs for various NLP applications.

Future Research

Moving forward, further research could explore the applicability of the developed data cleaning technique to other LLM datasets and investigate additional risk assessment variables for comprehensive quality assurance. Moreover, integrating machine learning algorithms for automated data cleaning processes could streamline and enhance the efficiency of quality control efforts in LLM dataset preparation.

Section V: Conclusion

This project successfully achieved its objectives by developing a robust data cleaning technique to enhance the quality of LLM datasets, particularly focusing on mitigating risk assessment variables within the WikiText dataset. Through a comprehensive analysis of sentence completion, readability, and lexical diversity, coupled with the application of rigorous statistical tests, the effectiveness of the proposed technique was confirmed. The iterative refinement of the data cleaning process, supported by computational resources and mentor guidance, resulted in a significant improvement in the performance of the GPT-2 model. By addressing key challenges and limitations, this research contributes to advancing the field of NLP by providing a practical solution to ensure the reliability and utility of LLMs for various applications. The developed data cleaning technique has the potential to impact the scalability and quality assurance efforts in the domain of large-scale LLM datasets, paving the way for more robust and efficient NLP solutions in the future.