Datasets to train language models can span terabytes and petabytes (Xue et al., 2020; Brown et al., 2020). With the scaling of training data and the demand for larger datasets, manually auditing and ensuring data quality becomes less feasible. This project aims to develop a data cleaning technique to clean LLM datasets with regard to sentence competition and determine the scale of performance improvement. WikiText, a popular open-source dataset, was collected and analyzed to address a risk assessment variable, with a data cleaning technique developed and iterated. The data cleaning technique was then applied to the WikiText dataset, and the cleaned data was used to train the GPT2 architecture. The resultant model was benchmarked and evaluated to establish the degree of significance of the performance increase. The results emphasize that this risk assessment variable is a prominent issue within various popular datasets. The developed data cleaning technique was effective, as the resultant model demonstrated increased performance. The data cleaning technique developed and iterated in this project can be used to create generalized heuristic functions and filtering methods for large-scale LLM datasets. With developing methods of mitigating data scarcity, filtering techniques could be largely significant for ensuring quality in accordance with scaling.

Keywords: Large Language Models (LLMs), Data Cleaning Techniques, Data Quality, Bias, Accuracy, Training Data, Datasets, Artificial Intelligence, Natural Language Processing