# The Optimization of Large Language Model Performance with Data Cleaning Techniques

Armaan Priyadarshan
Advisor: Dr. Kevin Crowthers, Ph. D.

Ar

## BACKGROUND

### LARGE LANGUAGE MODELS

LLMs are a form of artificial intelligence that train on vast amounts of data to perform various natural language tasks, such as understanding and generating text.

### RECENT GROWTH

A crucial factor in the development of LLMs is the large-scale text corpora used to train. LLMs continue to grow in size, with models being introduced that are comprised of hundreds of billions to trillions of parameters (Xue et al., 2020; Brown et al., 2020).

### NEED

As such, the need for automated quality control and filtering tools becomes increasingly pressing to maintain the effectiveness and standards of LLMs.

## Research Need

With the growing size of Large Language Model datasets, there is a need for methods of autonomously ensuring high data quality.

## Project Goal

This project aims to investigate and quantify the impact of data cleaning techniques on datasets and the performance of resultant trained language models.

## PROJECT METHODOLOGY

### DATASET ANALYSIS

Collect commonly used LLM datasets and analyze and pinpoint variables for risk assessment

### QUANTIFICATION

Isolate and determine the scale of the issues associated with the variables

### DATA CLEANING

Develop and iterate data cleaning techniques for especially prominent variables

### EVALUATION

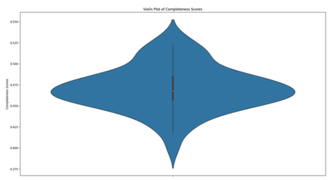
Apply the data-cleaning techniques to create clean datasets, and train and benchmark a pre-trained architecture

## Results

## Previous Research

- Despite models growing in size, they still often remain undertrained due to the focus on scaling model size without consideration of proportional training data size (Hoffmann et al., 2022; Dey et al., 2023).
- Research on other data cleaning techniques, such as deduplication, a data preparation technique that involves identifying and removing duplicate data from a dataset, has shown that popular existing datasets have quality issues present and cleaning their data could have a significant positive impact on the resultant model's performance (Lee et al., 2021).

## Figure 1




A violin plot of completeness scores of the first 100 training examples in the C4 dataset as evaluated by the BERT model and tokenizer for sequence classification.

## Figure 2

Decision Matrix

| Criteria | Weight | Readability | Lexical Diversity | Completeness Score |
|---|---|---|---|---|
| Correlation to lower-quality examples | 10 | 3 | 4 | 8 |
| Variability within datasets | 7 | 7 | 3 | 5 |
| Ease of implementation | 7 | 8 | 8 | 8 |
| Total | | 135 | 117 | 171 |

A decision matrix for determining which risk assessment variable to target in the data cleaning algorithm.

## Analysis

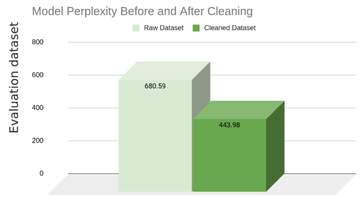
- Similarly to Figure 3, subsets of common datasets, such as (C4, WikiText, the Pile, etc.), will be put through Python programs and analyzed and scored with respect to different variables associated with quality.
- These results will be visualized using libraries such as Matplotlib and Seaborn.
- The training of resultant models will be analyzed using metrics relevant to the respective variables (eg. the diversity coefficient or perplexity).
- The models will also be evaluated using performance on downstream tasks.

## Table 1

| Model | Size (# Parameters) | Training Tokens |
|---|---|---|
| LaMDA (Thoppilan et al., 2022) | 137 Billion | 168 Billion |
| GPT-3 (Brown et al., 2020) | 175 Billion | 300 Billion |
| Jurassic (Lieber et al., 2021) | 178 Billion | 300 Billion |
| *Gopher* (Rae et al., 2021) | 280 Billion | 300 Billion |
| MT-NLG 530B (Smith et al., 2022) | 530 Billion | 270 Billion |
| *Chinchilla* | 70 Billion | 1.4 Trillion |

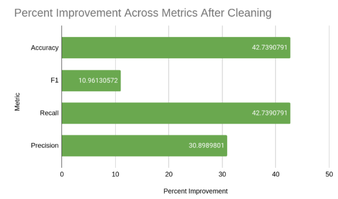
The lack of training tokens among larger LLMs relative to Chinchilla, a compute-optimal model (Hoffmann et al., 2022).

## Figure 3




The difference in the reported perplexity of the model trained on raw data versus the model trained on cleaned data.

## Figure 3




The difference in the reported perplexity of the model trained on raw data versus the model trained on cleaned data.

## Future Works

- While some preliminary research has been conducted on examples in C4 with a basic classification algorithm, more data and more variables must be analyzed.
- If the results are significant, multilingual datasets could potentially be analyzed and cleaned as well.
- A heuristic filter must be developed to clean the datasets.
- A pre-trained model architecture, likely the XL models, must be trained and evaluated on the datasets using Python wrappers for frameworks such as TensorFlow or PyTorch.