The Optimization of Large Language Model (LLM) Performance with Data Preparation Techniques

Grant Proposal

Armaan Priyadarshan

Mass Academy at WPI

85 Prescott St, Worcester MA

Executive Summary (Eng)

The abstract would contain an overall summary of what you (as the author) would like to convey. It would include some of the knowledge gaps that would eventually lead to researchable questions you have identified in the field.

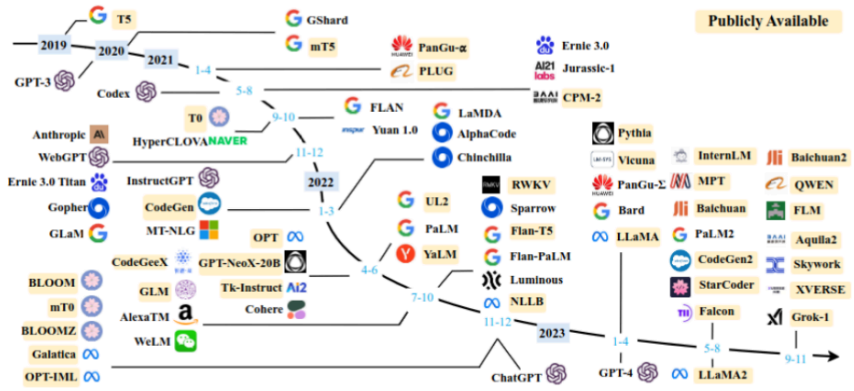
Keywords: emotion understanding, interest, social development, prosocial behavior, infants

The Optimization of Large Language Model (LLM) Performance with Data Preparation Techniques

Large Language Models

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. These models, exemplified by architectures such as GPT-4, BARD, LLaMA, and many more, are built upon complex neural networks composed of billions of parameters to perceive patterns in language. Modern models have the capacity to understand and manipulate various aspects of language, from syntactic and semantic structure to context-specific meanings. These models use unsupervised learning, a form of machine learning that involves learning from unlabeled data, to assimilate large amounts of language examples and acquire the ability to predict the next word or sequence of words in a sentence, honing their understanding of grammar, semantics, and context. However, the rise of LLMs also comes with concerns about bias and inaccuracy. Generating erroneous text, termed “hallucinations,” is common in LLMs, where models confidently generate false information (Zhang et al., 2023). Various forms of bias,

Figure 1



A timeline of existing large language models (having a size larger than 10B) in recent years (Zhao et al., 2023)

including racial bias, gender bias, and more, are also present to a degree in LLMs, with significant correlation to the training dataset (Sun et al., 2023). As a result, the training examples present in the datasets that LLMs use to train are a significant aspect to consider when evaluating and addressing the ethical implications of these models.

The Demand for High-quality LLM Training Data

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 of parameters to trillions of parameters (Brown et al., 2020). However, 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). As LLMs continue to grow and improve, training data must be scaled accordingly, meaning that datasets will also need to grow significantly. Datasets to train language models already span terabytes and petabytes (Xue et al., 2020; Brown et al., 2020). With the scaling of training data and the demand for larger datasets that comes with larger models, manually scanning and ensuring data quality becomes less feasible.

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)

Quality Issues in Existing Datasets

Quality issues exist in available datasets and are a significant concern. In a 2022 study, Julia Kreutzer and other researchers audited the quality of 205 language-specific corpora released with five major public datasets (CCAligned, ParaCrawl, WikiMatrix, OSCAR, mC4). The researchers found that lower-resource corpora have systematic issues: At least 15 corpora have no usable text, and a significant fraction contains less than 50% of sentences with acceptable quality. The study also mentions that many corpora are mislabeled or use nonstandard and ambiguous language codes, which could lead to erroneous interpretations. It demonstrates that these issues are noticeable even to non-experts and can be identified using automated analyses (Kreutzer et al., 2022). The information justifies concerns with data quality and shows how it could lead to incorrect results and a negative impact on the performance of the trained model.

Previous Research on Data Filtering Techniques

The need for automated quality control and filtering tools becomes vital to maintain the effectiveness and standards of LLMs. Previous research on data cleaning techniques shows that common datasets can have significant issues, such as biases and inaccuracies, which can detrimentally affect the performance and reliability of LLMs. In a study conducted by Katherine Lee and others (2021) on deduplication, a data preparation technique that involves identifying and removing duplicate data from a dataset, popular datasets such as C4 could consist of up to 7.18% of duplicate training examples. Cleaning this data could have a significant positive impact on the model performance regarding perplexity, or how confidently a model predicts words, and data overlap, which is

Table 2

	% train tokens with dup in train	% train tokens with dup in valid	% valid with dup in train
C4	7.18%	0.75 %	1.38 %
RealNews	19.4 %	2.61 %	3.37 %
LM1B	0.76%	0.016%	0.019%
Wiki40B	2.76%	0.52 %	0.67 %

The percentages of duplicate examples in the training and validation sets in various datasets and the percentage of overlap (Lee et al., 2022)

when duplicate examples are present in both the training and validation sets causing inflated metrics. The benefits of deduplication ranged from reducing memorized data to allowing for more efficient model training and sizes (Lee et al., 2021). This study underscores how datasets can be plagued by various data quality issues, primarily duplicate examples, and underscores the importance of automated data cleaning methods in enhancing the overall reliability and performance of LLMs.

Research Objective

This project aims to investigate and quantify the impact of various data cleaning techniques on LLMs through local training and benchmarking with online datasets and smaller language models. While past research has reported metrics such as perplexity, this project will also consider BIG-bench metrics to determine more general model performance improvements.

Section II: Specific Aims

This proposal's objective is to outline and describe the various facets of this project, including background, aims, methodology, and preliminary results. It should describe the need for the dataset optimization of LLMs, how it is proposed to be achieved, and the significance of the potential outcomes.

Our long-term goal is to identify methods of refining data quality that can impact the performance and continual improvement of LLMs. This project aims to approach a standardized system, algorithm, or technique for ensuring large-scale data quality over massive text corpora used in LLM training. The central hypothesis of this proposal is that applying data cleaning and filtering techniques to LLM datasets will result in a positive outcome on model performance. The rationale is that previous research shows that datasets can have shortcomings in quality,

and the application of smaller-scale data cleaning techniques has been shown to have a positive impact on the outcome of the model, so a general approach to data optimization with the application of various filters or techniques should be possible (Lee et al., 2021; Longpre et al., 2023). The work we propose here will contribute to the improvement of LLMs in terms of their viability as an assistive tool.

Specific Aim 1: Quantify the shortcomings of various publicly available datasets (C4, the Pile, WikiText, etc.) with regard to a variable addressed by the intended data cleaning technique(s)

Specific Aim 2: Apply the data cleaning technique(s) to the dataset and observe and generalize the difference it can make

Specific Aim 3: Train a pre-trained model architecture (XL) on the datasets, both those that are unmodified and those that are modified, and evaluate whether there are significant results and the extent of those results

The expected outcome of this work is a general approach to autonomously ensuring the quality of the dataset of an LLM. This outcome will also encompass improving the performance of the LLM training off of the cleaned dataset.

Section III: Project Goals and Methodology

Relevance/Significance

This project addresses a crucial gap in artificial intelligence by focusing on automating the data cleaning process for Large Language Models. As the demand for high-quality training data grows with the increasing sophistication and scale of LLMs, existing datasets exhibit

notable quality issues (Kreutzer et al., 2022). By leveraging data cleaning techniques, the research aims to provide a standardized approach for ensuring large-scale data quality in LLM training. The specific aims, ranging from quantifying dataset shortcomings to evaluating model performance, contribute to establishing a systematic method for refining data quality. The anticipated outcome is a general approach to autonomously ensuring dataset quality, offering potential improvements in LLM performance, and promoting responsible and effective use of these language models in diverse applications.

Innovation

This project introduces innovation by pioneering an automated and systematic approach to address the pressing challenge of data quality in LLM training. The innovation lies in the development and application of advanced data cleaning techniques tailored specifically for large-scale language datasets. By going beyond existing research and proposing a comprehensive methodology, the project innovatively combines a quantitative analysis of dataset shortcomings with the application of data cleaning techniques and a thorough evaluation of model performance.

Methodology

Initially, publicly available datasets commonly used in LLM training will be gathered. These datasets will be analyzed to determine variables that pose risks and could contribute to specific data-cleaning techniques. These variables could be used to identify the quality of training examples in the dataset and filter low-quality ones. Once a variable has been isolated, its specific concerns will be numerically and statistically quantified. Upon determining the scale

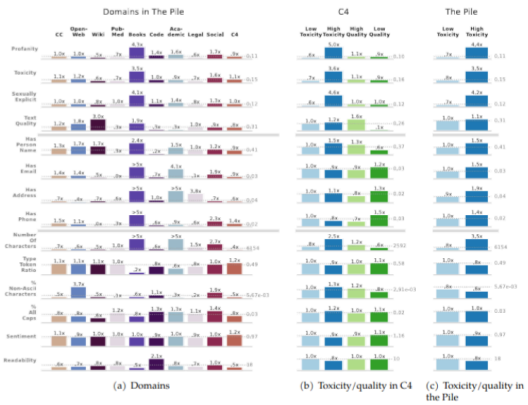
of the issue, a corresponding data cleaning technique will be developed and iterated to address the concerns posed by the specific variable. The data cleaning techniques will be applied to create new versions of the gathered datasets. Once the original and clean datasets have been collected and organized, they will be tested by using a pre-trained LLM architecture to train on both the original and cleaned datasets. The trained models will be benchmarked separately, and statistical analysis will be conducted to determine statistical significance. The effectiveness of the data cleaning techniques will be determined subsequently using BIG-bench metrics. Relative performance impact will be established by comparing metrics from the models trained on the raw data and cleaned data, respectively.

Specific Aim #1: Quantify the shortcomings of various publicly available datasets (C4, the Pile, WikiText, etc.) with regard to a variable (e.g. sentence completeness, lexical diversity, etc.) addressed by the intended data cleaning technique(s)

Justification and Feasibility. Quantifying the shortcomings of the available datasets is crucial to establishing the context for the scale of the problem being addressed. For the planned data cleaning techniques, the solution being attempted will be further established through this part of the methodology. Previous studies involving data cleaning or error analysis in datasets have included tabular or graphical information to quantify the issue.

Summary of Preliminary Data. Studies, such as that conducted by Shayne Longpre and

Figure 3



Feature differences across slices of the pretraining datasets (Longpre et al., 2023)

other researchers, included data about the features of the dataset relevant to the required need (Longpre et al., 2023). In Figure 3, the toxicity/quality in C4 and the toxicity/quality in the Pile dataset are laid out before their impact on the model is analyzed.

Expected Outcomes. The overall outcome of this aim is to characterize the scale of the issue within current datasets and provide a basis of relative comparison to establish the efficacy of the techniques applied.

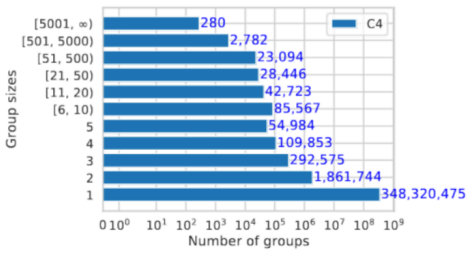
Potential Pitfalls and Alternative Strategies. We expect existing popular datasets to potentially have minimal issues regarding certain variables. In this case, a different data cleaning technique and variable within the datasets must be analyzed to establish a problem to solve. Once a data cleaning is developed, it can be applied to all LLM datasets as long as they are formatted as text collections.

Specific Aim #2: Apply the data cleaning technique(s) to the dataset and observe and generalize the difference it can make.

Justification and Feasibility. This aspect of the project is necessary to evaluate how effectively the data cleaning techniques can filter training samples and create filtered datasets to test by training models on them. Previous data cleaning studies also applied cleaning techniques and benchmarked the differences they made on the datasets.

Summary of Preliminary Data. The study on deduplication by Katherine Lee and other researchers

Figure 4



The distribution of near-duplicate cluster sizes from running NearDup, a data filtering technique, on C4, a popular dataset (Lee et al., 2021)

ran a deduplication method on C4, a popular dataset (Lee et al., 2021). Figure 4 shows how NearDup, one of the deduplication techniques used in the paper, clustered duplicate examples, showing how NearDup works and its impact on the dataset.

Expected Outcomes. The overall outcome of this aim is to optimize the datasets with respect to the cleaning or filtering techniques applied to them. Once the datasets are optimized, they can be used to train the language models and benchmark.

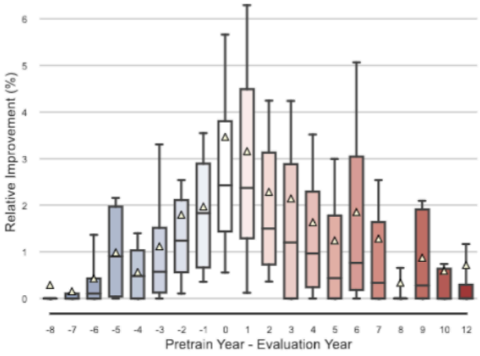
Potential Pitfalls and Alternative Strategies. The fulfillment of this aim relies heavily on the previous one. If the variable in the dataset being analyzed does not pose a significant issue in the datasets used, then the cleaning techniques will have to be varied. Additionally, due to the size of the datasets and potentially the complexity of the filters, the process might be expensive computationally and time-wise. An alternative strategy to work around this is using devices with more computational resources for testing.

Specific Aim #3: Train a pre-trained model architecture (XL) on the datasets, both those that are unmodified and those that are modified, and evaluate whether there are significant results and the extent of those results

Justification and Feasibility. A crucial aspect of this project is gauging the impact of data cleaning techniques on overall model performance. An LLM must be trained on the cleaned datasets and benchmarked to verify whether cleaning the datasets improves LLM performance. Previous studies have similarly benchmarked LLMs using BIG-bench metrics to validate experimental results.

Summary of Preliminary Data. In studies such as that conducted by Shayne Longpre and other researchers (2023), which investigated the impact of various data curation choices, language models are trained on different experimental groups of datasets and evaluated to highlight the differences. Figure 5 shows the relative improvement of models as they are trained on more temporally aligned data.

Figure 5



The relative performance increases in XL models as data age alignment improves (Longpre et al., 2023)

Expected Outcomes. The overall expected outcome of this aim is to determine a performance improvement in a model trained on cleaned data versus original data and assess the significance and degree of said improvement.

Potential Pitfalls and Alternative Strategies. Locally training the models might prove to be computationally infeasible. If this is the case, cloud resources or more powerful compute clusters will be needed for training.

Section IV: Resources/Equipment

Since this project is mainly based on software, there likely will be few physical resources or equipment necessary. As for software, the main necessary ones will be Python, relevant Python libraries, VSCode, HuggingFace, Git, and potentially Google Colab or data-oriented programming languages such as Julia or R. The only major concern is whether local computational resources will be sufficient for data cleaning and model training. If they are not, access to computing clusters at WPI will also be necessary.

Section V: Ethical Considerations

Since this project will not use biological organisms or chemicals, there should not be any major ethical or safety concerns. The only ethical consideration is the sourcing of the data, as it should not contain anything that violates the privacy of other individuals or groups. However, since mainly publicly available and popularly used datasets will be utilized for this project, data concerns will also likely not pose much of an issue.

Section VI: Timeline

Phase 1: Reading Literature and Narrowing the Scope of the Project (August through October)

- Read and annotate journal articles and patents relevant to LLM data quality, curation, and cleaning
- Research the knowledge gaps and determine a specific field to delve into further

Phase 2: Reaching out to Mentors and Companies for Guidance (September through November)

- Make professional contact with mentors relevant to preliminary research and pursue guidance

Phase 3: Proposal Writing Process (November through December)

- Write any necessary technical documentation describing and proposing the premise of the project

Phase 4: Performing the Methodology and Gathering Data (December through January)

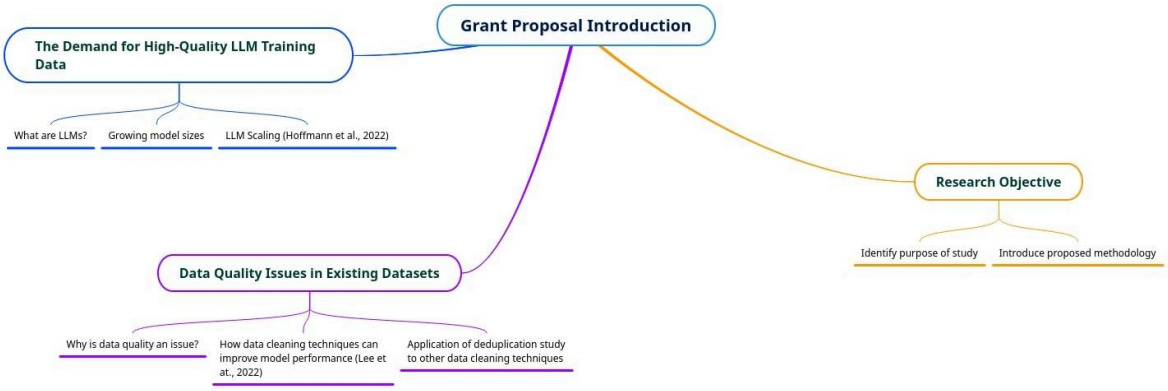
- Follow the procedure and steps outlined in the project methodology and gather data to be analyzed further with the necessary software.

Phase 5: Data Analysis

- Discuss and draw conclusions from the data gathered in Phase 4
- Generalize the conclusions and apply them to the broader field

Section VII: Appendix

Appendix 1: Grant Proposal Mindmap



Section VIII: References

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., . . . Amodei, D. (2020). Language models are few-shot learners (arXiv:2005.14165). arXiv. https://arxiv.org/abs/2005.14165

Dey, N., Gosal, G., Zhiming, Chen, Khachane, H., Marshall, W., Pathria, R., Tom, M., & Hestness, J. (2023). *Cerebras-GPT: Open compute-optimal language models trained on the Cerebras wafer-scale cluster* (arXiv:2304.03208). arXiv. https://arxiv.org/abs/2304.03208

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. de L., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G. van den, Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., … Sifre, L. (2022). *Training compute-optimal large language models* (arXiv:2203.15556). arXiv. https://arxiv.org/abs/2203.15556

Kreutzer, J., Caswell, I., Wang, L., Wahab, A., van Esch, D., Ulzii-Orshikh, N., Tapo, A., Subramani, N., Sokolov, A., Sikasote, C., Setyawan, M., Sarin, S., Samb, S., Sagot, B., Rivera, C., Rios, A., Papadimitriou, I., Osei, S., Suarez, P. O., … Adeyemi, M. (2021). *Quality at a glance: An audit of web-crawled multilingual datasets*. (arXiv:2103.12028). arXiv. https://arxiv.org/abs/2103.12028

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., & Carlini, N. (2021). *Deduplicating training data makes language models better* (arXiv:2107.06499). arXiv. https://arxiv.org/abs/2107.06499

Longpre, S., Yauney, G., Reif, E., Lee, K., Roberts, A., Zoph, B., Zhou, D., Wei, J., Robinson, K., Mimno,

D., & Ippolito, D. (2023). *A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity*. (arXiv:2305.13169). arXiv. https://arxiv.org/abs/2305.13169

Sun, H., Pei, J., Choi, M., & Jurgens, D. (2023). *Aligning with whom? Large language models have gender and racial biases in subjective NLP tasks* (arXiv:2311.09730). arXiv. https://arxiv.org/abs/2311.09730

Xue, L., Constant, N., Roberts, A., Kale, M., Siddhant, A., Barua, A., & Raffel, C. (2020). *MT5: A massively multilingual pre-trained text-to-text transformer* (arXiv:2010.11934). arXiv. https://arxiv.org/abs/2010.11934

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2023). *Siren's song in the AI ocean: A survey on hallucination in large language models* (arXiv:2309.01219). arXiv. https://arxiv.org/abs/2309.01219

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2023). *A survey of large language models*. (arXiv:2303.18223). arXiv. https://arxiv.org/abs/2303.18223