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, 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 self-supervised 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, including racial bias, gender bias, and more, are also present 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 to trillions of parameters (Brown et al., 2020). However, despite this growth in parameter count, models 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 tokens 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.

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 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.