Large Language Models (LLMs) for Learning Spreadsheet Tabular Data

Large language models (LLMs) have shifted the paradigm of modern artificial intelligence (AI) to an unprecedented scale. Although developed on text data, LLMs have shown versatile capability in learning and sharing knowledge for non-text data. Spreadsheets in the form of tabular data are popular examples of non-text data widely used in business, commerce, engineering, banking, and countless other applications. Tabular data sets, structured in rows and columns, may contain text and numerical data, whereas modern AI can process only numerical data representations. LLM may overcome learning such heterogeneous data representations. This research proposes a novel method to learn and process spreadsheet data of mixed data types by algorithmically finetuning an LLM. Our research uses an off-the-shelf LLM with 82 million parameters previously trained on large text databases. We finetune this LLM by infusing new knowledge from a spreadsheet data set we target to learn and analyze. We use ten spreadsheet data sets from diverse application domains to benchmark our approach compared to other baseline methods. Benchmarking involves predicting an outcome variable (e.g., loan eligibility) from customer data and information structured in the columns of spreadsheets. The finetuned version of the LLM is significantly more accurate in predicting patterns in unseen spreadsheet data than using the off-the-shelf LLM without such finetuning. Our proposed approach also results in superior accuracy compared to using an LLM application programming interface (API), such as GPT 3.5, launched on the cloud by OpenAI, with prompt engineering. Our comparative analysis highlights the trade-offs in computational efficiency, domain-specific learning, and generalization across tasks. When the number of columns in the spreadsheet is limited to ten, the accuracy of our proposed fine-tuning approach outperforms state-of-the-art baseline deep and machine learning methods. This research underscores the importance of tailoring LLMs to application-specific requirements and presents a novel framework for leveraging LLMs for learning tabular data, paving the way for transformative advancements in domains reliant on structured data analysis.