MBA17

Leveraging Large Language Models to Democratize Access to Costly Financial Datasets for Academic Research

Julian Junyan Wang[1], Victor Xiaoqi Wang[2]

[1]University of Oxford, Oxford, Oxfordshire, United Kingdom. [2]California State University Long Beach, Long Beach, CA, USA

Abstract

Unequal access to costly datasets essential for empirical research has long hindered researchers from disadvantaged institutions, limiting their ability to contribute to their fields and advance their careers. Recent breakthroughs in Large Language Models (LLMs) have the potential to democratize data access by automating data collection from unstructured sources. We develop and evaluate a novel methodology using GPT-4o-mini within a Retrieval-Augmented Generation (RAG) framework to collect data from corporate disclosures. Our approach achieves human-level accuracy in collecting CEO pay ratios from approximately 10,000 proxy statements and Critical Audit Matters (CAMs) from more than 12,000 10-K filings, with LLM processing times of 9 and 40 minutes respectively, each at a cost under $10. This stands in stark contrast to the hundreds of hours needed for manual collection or the thousands of dollars required for commercial database subscriptions. To foster a more inclusive research community by empowering researchers with limited resources to explore new avenues of inquiry, we share our methodology and the resulting datasets.

Conference Track

MIS and Business Analytics