Abstract

The rapid advancement and increasing accessibility of large language models (LLMs) have sparked significant interest in their potential applications across various fields, including auditing. This study aims to evaluate the performance of publicly accessible LLMs, particularly open-source models, in executing audit tasks and queries to determine if they have reached a level of competency comparable to human benchmark. By employing a human benchmarking approach, the research compares the outputs of LLMs to the exam scores of undergraduate accounting students from Universitas Gadjah Mada. The LLMs are tested using the same set of audit questions and additional audit tasks, with each test repeated multiple times to assess consistency and reliability. Furthermore, the study explores the enhancement of LLM performance through techniques such as Self Reflection and Agentic Workflow. The findings aim to provide insights into the feasibility of integrating open-source LLMs into auditing processes, specifically addressing concerns related to model performance. This research contributes to the body of knowledge by offering a direct comparison between LLMs and human auditors, evaluating methods to improve AI performance in auditing, and assisting organizations in making informed decisions about adopting LLM technologies for audit applications.

Keywords: artificial intelligence, large language models, auditing, open-source models, audit tasks, self reflection, agentic workflow, audit applications.