ABSTRACT

Large Language Models (LLMs) have become increasingly prevalent across industry and academia. However, despite their widespread adoption, the internal reasoning processes of these models remain largely opaque. This project investigates the alignment between the language output of LLMs and the underlying thought processes that generate those outputs. The objective is to assess whether the responses produced by LLMs exhibit coherence not only in fluent language but also in reasoning.

To evaluate this alignment, the study uses Knowledge Graphs (KGs), structured factual data sources, as an external benchmark for reasoning. Natural language questions are converted into SPARQL queries using two approaches: one directly through an LLM, and another via a dedicated Knowledge Graph Question Answering (KGQA) system. By comparing SPARQL queries, their results, and other aspects, the project assesses the consistency between the natural language response of the LLM and the structured query logic.

This dual-query methodology enables a heuristic evaluation of both language-thought alignment and response reliability. The findings are intended to contribute to a deeper understanding of how LLMs reliably reason, offering insight into their potential and limitations in tasks that require factual grounding and logical consistency.