Evaluation of Large Language Models as a Data Validation Tool

Scan Me

Author: Sauhard Dubey

Objective

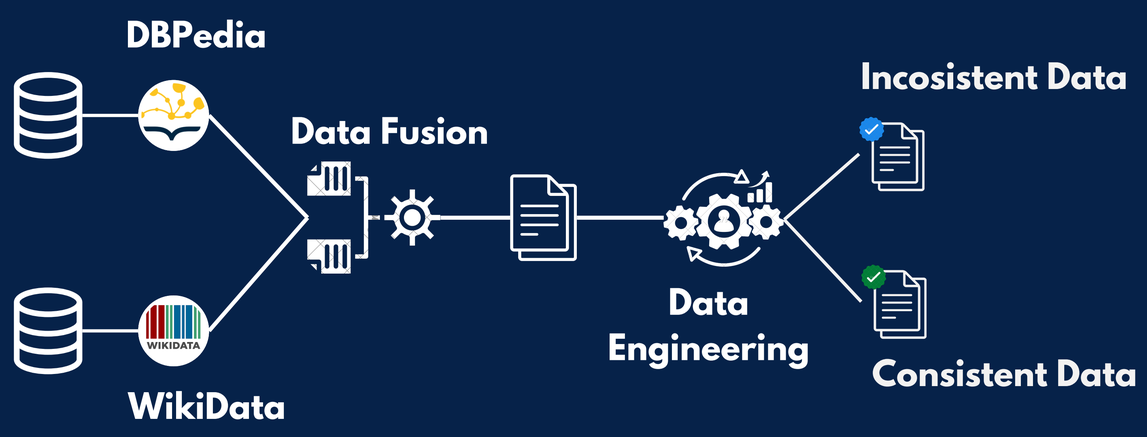
THE PROJECT AIMS TO INVESTIGATE THE RELAIBAILTY OF LARGE LANGUAGE MODELS(LLMS) FOR IDENTIFYING AND RESOLVING INCOSISTENCISES IN LARGE PUBLIC DATASETS LIKE WIKIDATA AND DBPEDIA.

Limitations

DBpedia and WikiData Sparql Query Server cannot process more than 40000 elements

Limited Input tokens for Gemini in the free version

Approach

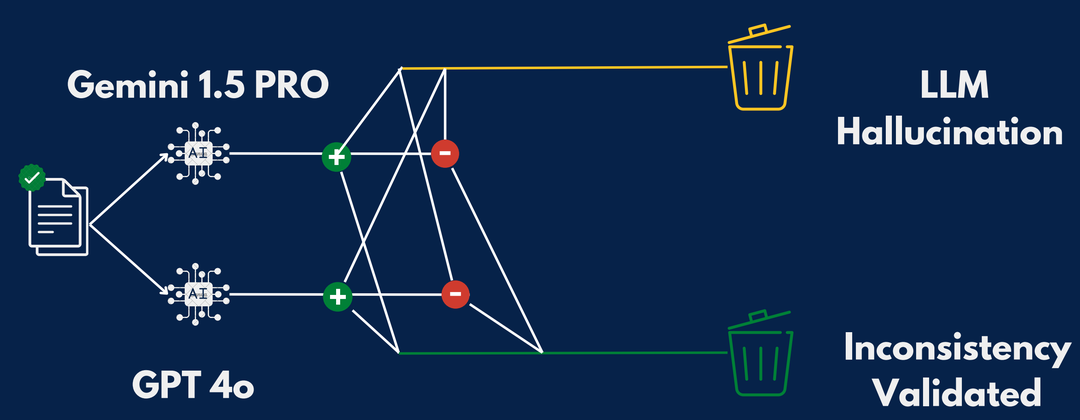


Phase 1

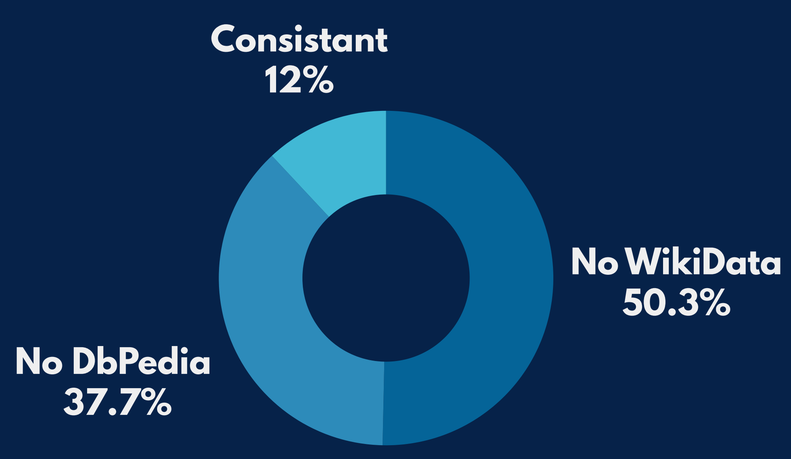
- Data Extraction from SPARQL Servers
- Data Fusion with the help of the redirected Wikipedia Links
- Creating separate datasets for consistent and inconsistent data

Phase 2

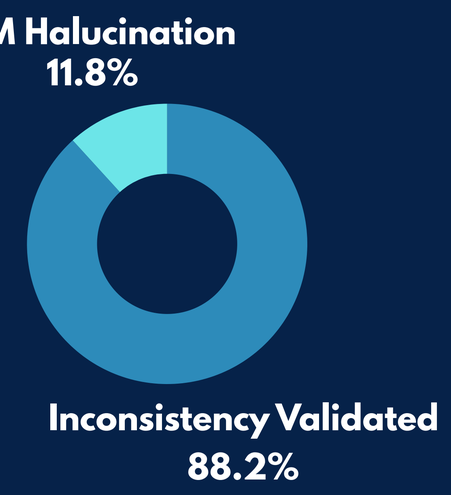
- Consistent Data is passed to GPT and Gemini to test the accuracy of both LLMs in Validation Tasks
- If the accuracy of consistent data is high, a common answer from both LLMs can be used to remove inconsistencies from the Knowledge bases



Findings and Results



- The amount of inconsistencies in the two knowledge bases is extremely high.
- GPT-4o: Validates consistent data with an accuracy of **99%** while Gemini 1.5 Pro: Provides an accuracy of **94%** for the same data.
- For inconsistent data, the LLMs exhibit hallucination rates **below 12%**, making them effective tools for validation purposes.



Technical University of Munich , TUM School of Computation, Imformation and Technology

Bachelor Practical Course in Data Engineering 2024

sauhard.dubey@tum.de