Małgorzata Gierdewicz id: 148264

Introduction

When a clinician is delivering a message to the patient, they can sometimes overestimate the recipient's knowledge on said topic. The most important words about diagnosis can be delivered too harshly or without explaining key concepts. The presented agentic system is aimed to mitigate those problems, with an LLM system that will work as a judge for the clinicians answers. Based on that, knowledge graphs are built to aid the scoring provided by LLM's, improving the explanation of said models.

related works:
https://arxiv.org/abs/2505.23802
https://www.researchgate.net/publication/391707688_Communication_Styles_and_Reader_Preferences_of_LLM_and_Human_Experts_in_Explaining_Health_Information

Data

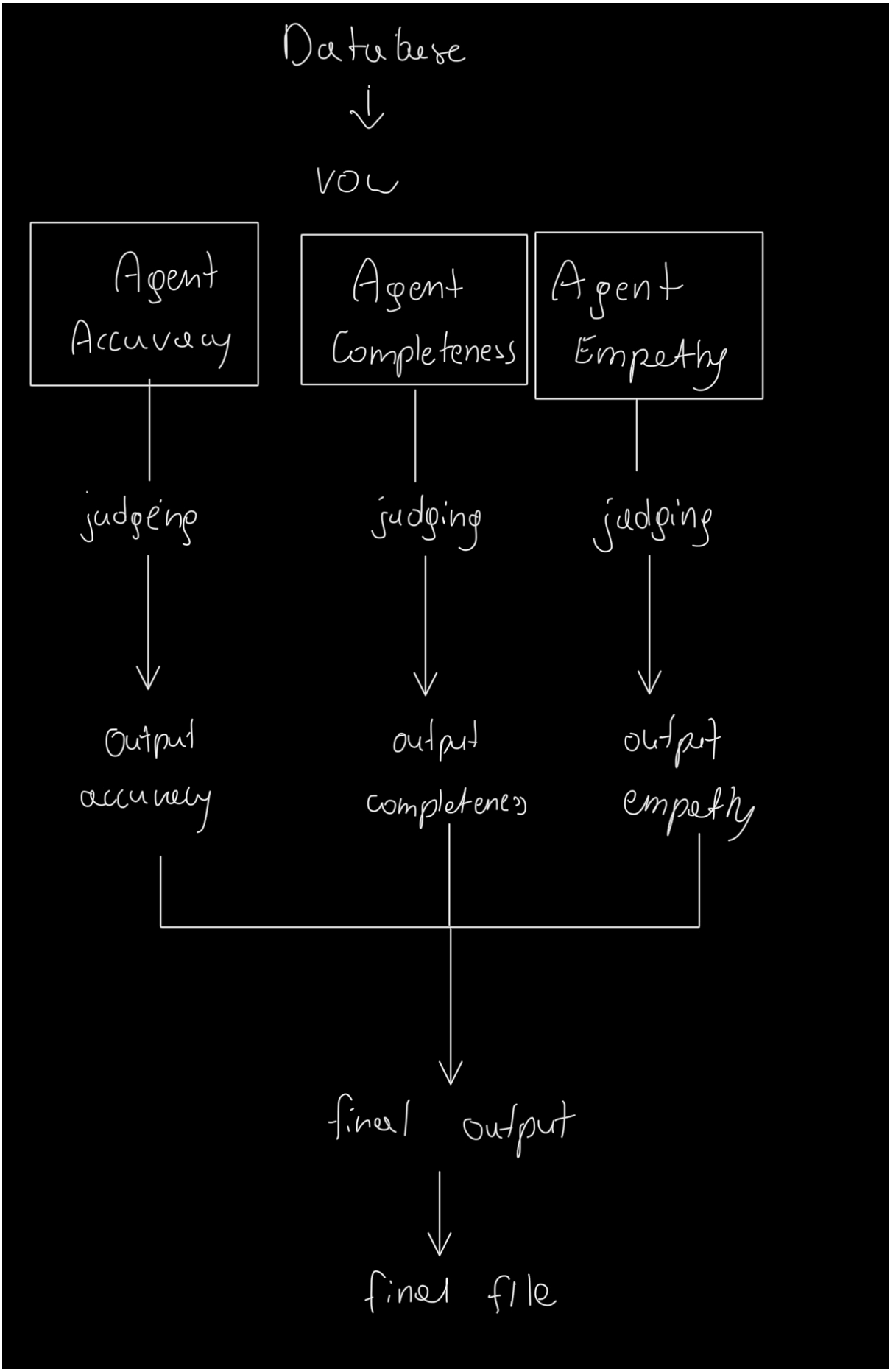
The data consists of 55 questions regarding health with responses made by real doctors. Data for this project was collected from the Roche x MiT Healthcare LLM-a-thon, which took place september 2025.

Method

The system depends on few important steps:

1. Agent Creation:
   To better evaluate the data, 3 concurrent agents were created, concerning 3 criteria on which doctor's response should be evaluated, the criteria being accuracy, completeness and empathy.

2. Judge LLM:
   A Gemini-2.5-flash model was chosen for the judging function, because of its popularity and availability. It is also an IFT model - finetuned for instruction prompts. Unlike simple binary classifiers, the Core Judge is prompted to extract a Knowledge Graph. It identifies key concepts (graph_nodes) and maps the logical relationships between them (graph_edges) that led to the verdict. This improves the overall explainability of the LLM judgment.
3. Agentic workflow:
   Each clinician answer is passed to the ai agents, who then call on the judge function to evaluate each answer. This approach lowers the runtime, however the system waits for each agent to process the text for it to move to another answer. The output from the agents are in the end merged in a dataframe row and saved to a csv file. Visualizations are then created based on this file.


Database
VOL
Agent Accuracy
Agent Completeness
Agent Empathy
judging
judging
judging
output accuracy
output completeness
output empathy
final output
final file

Evaluation

In LLM as a judge method, evaluation is performed by instructing the model by each agent to check the provided text for Accuracy, Completeness or Empathy, based on the criteria the agent initially judges on. Then the LLM judge is instructed via prompt to judge the answer, give it a score (0 if it doesn't meet the criteria and 1 if it meets them) and create a knowledge graph with all of the relations and items that were relevant to critique. After that the information from all 3 agents is merged into one pandas dataframe.

Conclusions

The system works as intended, meaning that the agents do judge based on criteria and the answers are scored. The system can be improved, by introducing RDFS, that will instruct the LLM to only use the properties defined in the schema. That way the LLM is less prone to hallucinating the answer, and the solution safeguards the LLM judgment.