GenAI for human-centric table data generation and understanding

Project Description:

FANAR (https://fanar.qa/en) is Qatar's home-grown foundational model, developed by QCRI with the objective of preserving Muslim cultural values. We are developing multi-modal components for FANAR to facilitate knowledge acquisition from various information sources such as images, videos, and text. One resource yet to be fully utilized is human-centric table (HCT) data, commonly found in statistical reports, such as those available on the National Planning Council website (https://www.npc.qa/).

For example, a real estate promoter might seek statistical data on housing and wages from national reports to guide their next project. They would analyze this data to identify trends and make informed decisions regarding the optimal location and type of real estate to develop. Similarly, a manager preparing a report for company leadership might use internal statistical data to provide insights into company performance and strategic direction. By examining sales data, customer feedback, and operational metrics from internal HCT reports, the manager can assist leadership in making informed decisions about future investments, resource allocation, and strategic initiatives.

However, such tables pose significant challenges for current foundational models. To address this, we have developed a benchmark consisting of hundreds of triplets of HCTs, questions, and answers in natural language, as well as a synthetic table and questions generator, for evaluating existing large language models (LLMs). Our goal is to leverage this data to enhance existing LLMs, particularly FANAR, to better handle HCTs.

The recent agentic approach involves developing specialized agents for specific tasks. In this proposed project, we aim to explore the feasibility of developing agents specialized in understanding HCTs. These agents would enable accurate responses to questions related to HCTs (RAG) or summarization of HCT data into text based on information from multiple tables, whether accessed in the prompt, during pre-training, or through fine-tuning.

We propose to investigate the following directions:

- Finetuning FANAR to improve its accuracy in answering natural language questions about HCTs. This includes subtasks such as selecting the appropriate HCT from a set of HCTs containing the answer and providing an accurate response from the selected table. We will compare two approaches: fine-tuning directly based on HCT-QA triplets, and transforming HCTs into relational tables and natural language questions into SQL queries to derive the answer.
- Another area of interest involves generating HCTs from prompt instructions given relational data tables. This approach aims to facilitate data reporting.

By pursuing these directions, we seek to enhance FANAR's capabilities in managing and interpreting human-centric table data effectively.

Project Type: Research and Engineering

Internship Batch:

- **Batch 1:** May 11 to July 10, suitable for Education City students, i.e., CMUQ, TAMUQ and HBKU students

Duties/Activities:

Analyze the problem, select the relevant data and models, implement with Python (Pytorch, Hugging-Face) and test various options to help decide the best.

Required Skills:

Python, LLMs prompting and fine-tuning, Hugging-Face

Preferred Intern Academic Level:

Last year undergraduate, graduate

Learning Opportunities:

LLMs prompting and fine-tuning, FANAR, Hugging-Face, Google-cloud

Expected Team Size: *it is preferable to have team projects*

Two students per project.

Mentors

Name: Michael AUPETIT, Mohamed ELTABAKH

email: maupetit@hbku.edu.qa, meltabakh@hbku.edu.qa