**BIL471/571 NLP Project - Sprint Update**

**Project Title:** TurkishMedQA-LLM: Specialized Medical Question-Answering Model

**1. Sprint Goal & Summary of Progress**

* **Sprint Goal(s):**
  + Expand and diversify the Turkish medical QA dataset
  + Evaluate doctor responses using an LLM for informativeness and factual accuracy
  + Select high-quality informative samples for future fine-tuning
  + Prepare fine-tuning dataset for domain-adapted LLM
* **Key Accomplishments:**
  + Collected additional Q&A pairs from diverse medical topics to enrich dataset scope
  + Used a LLM to assess informativeness and verification quality of doctor responses
  + Identified and selected high-quality, informative examples as candidates for training
  + Designed preprocessing scripts to format selected samples into fine-tuning-ready format

**2. Detailed Updates**

* **Literature Survey (/docs/Literature\_Survey/):**
  + Reviewed IEEE Paper: “The Large Language Model as a Judge: Evaluating LLMs with LLMs” (IEEE, 2024)
  + Gained insights into using LLMs for scoring relevance, factual accuracy, and helpfulness of generated answers
  + Adapted this protocol for evaluating doctor-written responses in our dataset
  + Informed scoring rubric and thresholding method for selecting training candidates
* **Data Collection & Preprocessing (/data/):**
  + Expanded dataset with newly scraped Q&A pairs across multiple specialities
  + Labeled examples based on informativeness using LLM scoring criteria
* **Code & Development (/src/):**
  + Developed automation script to run LLM-as-a-Judge evaluations in batches
  + Implemented scoring threshold logic to classify Q&A pairs as “suitable” or “needs revision”
  + Scripted data conversion for formatting informative examples into supervised fine-tuning JSONL
* **ML Model / Experiments:**
  + No new LLM training this sprint
  + Focused on dataset curation and evaluation pipeline
  + Created a shortlist of high-quality Q&A pairs to be used in the next sprint for fine-tuning
  + Began early stage prompt experimentation for response verification consistency

**3. Challenges & Blockers**

* LLM-based evaluation is compute-intensive, making large-batch scoring time-consuming
* Some doctor responses are highly context-dependent, making informativeness difficult to score automatically
* Subjective edge cases require manual review despite LLM assessment

**4. Task Breakdown & Contributions**

| Task Description | Team Member(s) Responsible | Status (Completed / In Progress / Blocked) |
| --- | --- | --- |
| *Expand and diversify scraped dataset* | *Bekir Bilgean Tekin*  *Mehmet Yasin Tosun*  *Semih Uçan* | *Completed* |
| *Implement LLM-as-a-Judge pipeline* | *Semih Uçan* | *Completed* |
| *Evaluate doctor responses for informativeness* | *Bekir Bilgehan Tekin*  *Mehmet Yasin Tosun* | *Completed* |
| *Select top-scoring Q&A samples for fine-tuning* | *Mehmet Yasin Tosun*  *Semih Uçan* | *Completed* |
| *Format selected data into supervised training format* | *Bekir Bilgehan Tekin* | *Completed* |

**5. Goals for Next Sprint**

(What are your specific, measurable goals for the next two weeks before our next meeting?)

* Fine-tune Mistral or Hippocrates model on selected Turkish medical QA dataset
* Conduct initial fine-tuning run and evaluate zero-shot vs fine-tuned performance
* Begin integrating Retrieval-Augmented Generation (RAG) into the architecture
* Draft Evaluation and Methods sections of final report

PROJECT MEETING:

Increase and diversify data,

LLM as a Judge for informative and verification of responses

Select most informative ones for training LLM

LLM evaluation of doctors

LLM finetuning/training dataset preparations.

https://ieeexplore.ieee.org/document/10852500

Medgemma etc.