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

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

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

* **Sprint Goal(s):**
  + Construct an initial Turkish medical QA dataset from public sources
  + Conduct baseline experiments using an instruction-tuned LLM
  + Evaluate few-shot vs zero-shot prompting effectiveness
* **Key Accomplishments:**
  + Collected and cleaned 1,200 Q&A pairs in Turkish from medical platforms (ENT domain)
  + Designed a structured JSON schema for both Q&A pairs and doctor metadata
  + Evaluated Mistral-7B-Instruct-v0.2 with zero-shot and few-shot prompting strategies
  + Measured BLEU and ROUGE metrics for model output comparisons

**2. Detailed Updates**

* **Literature Survey (/docs/Literature\_Survey/):**
  + Reviewed key medical LLMs: Med-PaLM 2, MeLLaMA, MEDITRON, Hippocrates
  + Highlighted advanced training techniques such as CPT, DPO, RAG, and PEFT
  + Informed our fine-tuning and evaluation design (e.g., hybrid metrics, expert feedback loop)
* **Data Collection & Preprocessing (/data/):**
  + Scraped ENT-domain Q&A pairs from doktorsitesi.com following ethical scraping guidelines. We have also started scraping Q&A pairs from other areas from doktorsitesi.com
  + Cleaned and normalized the text, removed HTML artifacts
* **Code & Development (/src/):**
  + Wrote data cleaning and web-scraping scripts
  + Implemented prompt generation logic for zero-shot and few-shot testing
  + Integrated ROUGE and BLEU metric calculators for automated evaluation
* **ML Model / Experiments:**
  + Model: Mistral-7B-Instruct-v0.2
  + Setup: Prompted using 10 Q&A samples (zero-shot and few-shot modes)

| **Setting** | **ROUGE-1** | **ROUGE-2** | **ROUGE-L** | **BLEU** |
| --- | --- | --- | --- | --- |
| Zero-shot | 0.1471 | 0.0179 | 0.0909 | 0.0149 |
| Few-shot | 0.2224 | 0.1311 | 0.1897 | 0.0867 |

* + Results:
  + Interpretation: Few-shot prompting significantly improved answer quality and fluency

**3. Challenges & Blockers**

* *Many doctor responses in the dataset are short and generic (e.g., “see a specialist”), limiting the richness of training signals*
* *Hardware limitations during web scraping caused prolonged data collection times due to slow processing and memory constraints*
* *Limited hardware resources restrict fine-tuning large models*

**4. Task Breakdown & Contributions**

| Task Description | Team Member(s) Responsible | Status (Completed / In Progress / Blocked) |
| --- | --- | --- |
| *Web-scraping of Turkish medical data* | *Mehmet Yasin Tosun Semih Uçan* | *In Progress* |
| *Convert scraped data to structured JSON format* | *Bekir Bilgehan Tekin* | *Completed* |
| *Literature review on MedLLMs* | *Mehmet Yasin Tosun*  *Semih Uçan* | *Completed* |
| *Run zero-shot and few-shot experiments* | *Bekir Bilgehan Tekin*  *Semih Uçan* | *Completed* |
| *Performance evaluation using BLEU/ROUGE metrics* | *Bekir Bilgehan Tekin*  *Mehmet Yasin Tosun* | *Completed* |

**5. Goals for Next Sprint**

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

* Expand dataset to cover at least 10,000 QA pairs across 10+ specialties
* Fine-tune open-source medical LLMs on Turkish data
* Draft the “Methods” section of the 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