**Progress Report**

**Project Title: Turkish Medical Q&A Model Evaluation for Otolaryngology (ENT)**

**Date:** **June 7, 2025**

**1. Data Collection**

We collected a dataset of **1,200 Turkish question-answer pairs** specifically related to the field of **Otolaryngology (Ear, Nose, and Throat / KBB)**.  
The dataset consists of real patient questions and expert doctor answers, providing a relevant benchmark for medical language model evaluation.

**2. Experimental Setup**

From the full dataset, **10 representative Q&A pairs** were selected for controlled experiments.  
The selected samples were used to evaluate the performance of the **Mistral-7B-Instruct-v0.2** language model in two different prompting scenarios:

* **Zero-shot:**  
  The model is provided only with the test question and is expected to generate a direct answer, without any sample Q&A examples.
* **Few-shot:**  
  The model is prompted with 3 example Q&A pairs before being presented with the test question. This approach aims to help the model better learn the expected answer style and content.

**3. Evaluation Metrics**

The generated answers were evaluated using widely accepted text similarity metrics:

* **ROUGE-1 F1:** Unigram (single word) overlap between generated and reference answers.
* **ROUGE-2 F1:** Bigram (two-word phrase) overlap.
* **ROUGE-L F1:** Longest common subsequence (captures fluency and structure).
* **BLEU:** Measures n-gram precision with a brevity penalty, commonly used for machine translation and text generation.

**4. Results**

**Zero-shot Results:**

* **ROUGE-1 F1:** 0.1471
* **ROUGE-2 F1:** 0.0179
* **ROUGE-L F1:** 0.0909
* **BLEU:** 0.0149

**Few-shot Results:**

* **ROUGE-1 F1:** 0.2224
* **ROUGE-2 F1:** 0.1311
* **ROUGE-L F1:** 0.1897
* **BLEU:** 0.0867