Progress Report: Development of a Turkish Medical QA System Using Large Language Models

Bilgehan Tekin[1] – 211401008
Artificial Intelligence Engineering
TOBB University of Economics and Technology
Ankara, Türkiye
bekirbilgehantekin@etu.edu.tr

Mehmet Yasin Tosun[2] – 211401003
Artificial Intelligence Engineering
TOBB University of Economics and Technology
Ankara, Türkiye
mehmetyasin.tosun@etu.edu.tr

Semih Uçan[3] – 211401012
Artificial Intelligence Engineering
TOBB University of Economics and Technology
Ankara, Türkiye
s.ucan@etu.edu.tr

***Abstract*—This project aims to develop a specialized Turkish medical question-answering (QA) system by leveraging large language models (LLMs). A domain-specific dataset was constructed by ethically scraping publicly available medical QA content from Turkish medical websites, particularly from doktorsitesi.com. The curated data was structured using a custom JSON schema and preprocessed for use in downstream tasks. Initial experiments involved zero-shot and few-shot evaluations using the Mistral-7B-Instruct-v0.2 model, benchmarked with standard NLP metrics such as BLEU and ROUGE. Preliminary results demonstrate that few-shot prompting significantly improves response quality. The ultimate goal is to fine-tune a medical LLM on the Turkish dataset and assess its effectiveness in domain-specific QA and basic symptom-checking, comparing its performance to general-purpose open- and closed-source models.**

I. Introduction

The advent of large language models (LLMs) has introduced a paradigm shift in natural language processing (NLP), enabling machines to understand and generate human-like text with remarkable fluency and contextual awareness. When applied to healthcare, these models have the potential to revolutionize a wide range of applications, including clinical documentation, diagnostic support, medical education, telemedicine, and patient communication. Their ability to synthesize extensive medical knowledge and respond to natural language queries makes them highly suitable for question-answering (QA) systems aimed at both clinicians and patients.

Recent advances in models such as GPT-4, Med-PaLM 2, and MEDITRON-70B have demonstrated exceptional performance on medical QA benchmarks. These systems can answer complex clinical questions with high factual accuracy, sometimes even surpassing expert-level performance in standardized evaluations. However, these capabilities are largely enabled by access to large-scale, high-quality biomedical corpora—resources that are predominantly available in English. As a result, the benefits of these technologies are disproportionately skewed toward English-speaking populations and healthcare systems.

This linguistic asymmetry presents a significant barrier to equitable access in AI-powered healthcare. Languages such as Turkish—spoken by over 80 million people—remain underrepresented in the medical AI landscape. Turkish's morphologically rich and syntactically complex structure adds further challenges, and there is currently a lack of robust, domain-specific Turkish datasets for clinical NLP applications.

Moreover, many existing medical QA systems are limited in scope, often focusing on narrow specialties. This hinders the development of general-purpose and trustworthy tools capable of addressing a wide array of patient questions across multiple medical domains. To address this gap, we propose the creation of a large-scale, multilingual, and multi-specialty Turkish medical QA system. Rather than focusing on a single domain (e.g., internal medicine or otolaryngology), our system is designed to support diverse specialties including cardiology, dermatology, endocrinology, neurology, pediatrics, psychiatry, gynecology, and infectious diseases.

In the initial phase of this project, we constructed a pilot dataset of question-answer pairs based on real doctor-patient interactions collected from publicly available Turkish medical platforms. Data collection followed ethical web scraping practices, adhering to site-specific access policies (e.g., `robots.txt`) and ensuring anonymization of personal information. We developed a structured JSON schema to represent each QA entry along with metadata on medical specialty, provider, and clinical context.

To evaluate the applicability of current instruction-tuned LLMs in Turkish medical QA, we conducted initial experiments using the Mistral-7B-Instruct-v0.2 model. Selected for its strong instruction-following and multilingual capabilities, Mistral was tested under both zero-shot and few-shot prompting strategies. While zero-shot responses demonstrated basic comprehension, few-shot prompting yielded substantially more accurate and contextually relevant answers.

This report outlines the foundational phase of our system development, establishing a framework for expanding

into a high-coverage, clinically useful QA platform. Future work will focus on increasing the dataset to over 5,000 QA pairs across more than 10 specialties, integrating retrieval-augmented generation (RAG) for enhanced factual grounding, and incorporating expert review to ensure clinical safety and reliability. Our ultimate goal is to deliver a robust, accessible, and domain-aware medical QA system for Turkish-speaking users..

II. Related Work

A. Med-PaLM 2

Med-PaLM 2 [1] is an advanced medical language model built upon Google's PaLM 2 architecture. It employs two innovative strategies: Ensemble Refinement (ER), which aggregates multiple candidate responses to improve consistency, and Chain of Retrieval (CoR), which grounds the response in relevant medical literature. It achieved 86.5% accuracy on the MedQA (USMLE) benchmark, surpassing the average performance of human physicians on multiple clinical dimensions. The model was also evaluated for factuality, bias, and harmfulness, positioning it as a benchmark in trustworthy medical QA.

B. MeLLaMA

MeLLaMA [2] extends the LLaMA-2 model by instruction-tuning it with 129 billion tokens of biomedical and clinical data. It demonstrates strong multitask performance across question answering, named entity recognition (NER), summarization, and clinical diagnosis. Notably, it outperforms GPT-4 on 5 out of 8 standard medical benchmarks, showcasing the effectiveness of extensive domain-specific fine-tuning. The study emphasizes the role of instruction variety and task diversity in achieving robust performance across medical NLP applications.

C. MEDITRON-70B

MEDITRON [3] is an open-source LLM with 70 billion parameters, fine-tuned on a comprehensive biomedical corpus including PubMed articles, clinical practice guidelines, and scientific abstracts. Built on the LLaMA-2 base model, MEDITRON excels on medical QA benchmarks such as MedMCQA, PubMedQA, and MedQA, where it performs on par with or better than GPT-3.5. Its strong few-shot capabilities and open accessibility make it a promising candidate for healthcare applications requiring transparency and reproducibility.

D. Hippocrates (Hippo-7B)

The Hippocrates framework [4] introduces the Hippo-7B and Hippo-Mistral models, which leverage continued pretraining (CPT) on 298M tokens of high-quality medical data and instruction tuning with 292K curated prompts. Additionally, Direct Preference Optimization (DPO) using GPT-4-based annotations aligns model outputs with human expectations. The models demonstrate robust performance across four major QA datasets (MedQA, PubMedQA, MedMCQA, and USMLE-style questions), especially in few-shot settings, despite their smaller size compared to GPT-4.

E. Survey: Med-LLMs

A comprehensive survey by Tong et al. [5] offers a taxonomy of foundation models in healthcare, covering architecture types, pretraining data sources, adaptation strategies, and downstream tasks. The survey highlights emerging techniques like Retrieval-Augmented Generation (RAG) and Parameter-Efficient Fine-Tuning (PEFT) as crucial for domain adaptation. It also emphasizes the lack of benchmark datasets in low-resource languages and calls for more diverse evaluation protocols.

F. Survey: LLMs in Healthcare

Another recent survey [6] systematically reviews challenges in deploying LLMs in real-world healthcare settings, including concerns of hallucination, bias, and lack of explainability. It stresses the need for transparent model evaluation, user trust, and regulatory compliance. The study proposes integrating symbolic reasoning, structured knowledge bases, and human-in-the-loop feedback for safer QA deployment.

G. Trustworthy Medical QA

Zhang et al. [7] propose a framework for defining and evaluating trustworthiness in medical QA systems. Trust is conceptualized as a multidimensional construct involving factual accuracy, consistency, robustness, and transparency. The survey reviews commonly used evaluation methods, such as BLEU, ROUGE, Exact Match (EM), and human preference studies, and suggests the importance of hybrid metrics tailored for clinical safety.

III. Dataset Collection and Processing

A. Data Source and Ethics

We compiled a domain-specific Turkish medical QA dataset focused on Otolaryngology (ENT), using publicly accessible Q&A entries from Turkish medical websites, particularly doktorsitesi.com. All data collection was conducted ethically in compliance with the websites' `robots.txt` files and terms of service. We followed a polite scraping policy with appropriate request intervals and proper user-agent identification to avoid burdening the servers.

B. Dataset Description

Size: 1,200 high-quality Q&A pairs involving real patient queries and doctor-written answers.
Language: Turkish.
Domain: ENT (Ear, Nose, and Throat).
Topics Covered: Sinusitis, hearing loss, throat infections, nasal polyps, tinnitus, deviated septum, and post-nasal drip.
Use: The dataset serves as both training and evaluation material for LLM-based QA systems.
A smaller subset of 10 diverse and representative samples was used for initial experiments.

C. Data Schema (JSON Format)

Each Q&A entry follows a structured JSON format:

Q&A Entry:

```
{
  topic: "Kulak Burun Boğaz",
  title: "Geniz akıntısı ve boğazda yanma",
  question: "Yaklaşık 1 aydır geniz akıntısı
  var ve boğazım yanıyor...",
  answer: "Muhtemelen kronik rinosinüzit
  kaynaklıdır. Kulak burun boğaz uzmanına
  başvurmalısınız.",
  doctorID: "ENT001"
}
```

Doctor Metadata:

```
{
  doctorID: "ENT001",
  name: "Dr. Ayşe Demir",
  title: "Uzman Dr.",
  specialty: "Rinoloji, Kulak Hastalıkları",
  clinicName: "İstanbul KBB Merkezi",
  clinicAddress: { street: "Barbaros Cad.",
  city: "İstanbul" },
  about: "Hacettepe Tıp Fakültesi mezunu,
  12 yıl deneyim.",
  rating: 4.8
}
```

D. Cleaning and Storage

The raw scraped text was cleaned to remove HTML tags, special characters, and noise. Turkish characters were preserved during normalization. The processed Q&A pairs were validated for completeness and correctness before being stored in structured JSON format. The dataset is stored using a NoSQL-compatible schema to facilitate fast querying and future extension to additional specialties. Diversity across ENT subtopics was prioritized to improve generalization during model evaluation.

IV. Initial Experiments

A. Model Selection

For the initial phase of our model evaluation, we selected the open-source **Mistral-7B-Instruct-v0.2** model. This model was chosen due to its strong multilingual generalization ability, which is particularly valuable for Turkish language with relatively limited medical NLP resources. Mistral-7B-Instruct is optimized for instruction-following tasks and benefits from dense pretraining on diverse corpora, followed by supervised fine-tuning on human-annotated instruction datasets. These characteristics make it a competitive baseline for zero-shot and few-shot scenarios without requiring extensive task-specific fine-tuning. Additionally, Mistral supports efficient inference on commodity hardware with limited VRAM, which aligns with the constraints of our research infrastructure.

B. Experimental Setup

To evaluate the model's medical QA performance, we extracted a focused subset of 10 question-answer pairs from the Ear, Nose, and Throat (ENT) domain in our custom Turkish dataset. Each Q&A pair contains a real-world medical question posed by a patient and a corresponding expert response. Two distinct prompting settings were used:

- **Zero-shot Prompting:** The model is given only the raw user question with no additional context or examples. This setting assesses the model's ability to generalize medical reasoning from its pretraining alone.
- **Few-shot Prompting:** Prior to the test question, the prompt includes three manually selected Q&A pairs relevant to the ENT domain. These examples are formatted in a consistent style and serve as in-context learning signals. Few-shot prompting is particularly useful in scenarios where model fine-tuning is infeasible, as it allows leveraging small annotated datasets to condition the model dynamically.

In both settings, prompts were carefully constructed to minimize ambiguity and ensure consistency in the evaluation procedure. The few-shot examples were curated to include variations in medical subtopics (e.g., tonsillitis, ear infections, nasal congestion) to increase generalizability.

C. Evaluation Metrics

We employed both surface-level and semantic metrics to evaluate the model's responses. Specifically:

- **ROUGE-1, ROUGE-2, and ROUGE-L:** These metrics capture unigram and bigram overlap, as well as the longest common subsequence, between the generated and reference answers.
- **BLEU:** A precision-based metric traditionally used in machine translation, BLEU evaluates how many n-grams in the model output match the reference response.

Although these metrics primarily measure textual similarity, we recognize that in the medical QA domain, factual accuracy and clinical appropriateness are often more critical. Therefore, these automated metrics serve as a proxy for preliminary evaluation, pending future incorporation of human expert judgment and clinical validation.

V. Results

We conducted a series of initial evaluations using the Mistral-7B-Instruct-v0.2 model to assess its capability to generate accurate and contextually appropriate responses for Turkish ENT-related medical questions. We compared the model's performance in both zero-shot and few-shot prompting settings using a subset of 10 manually curated Q&A pairs from our dataset.

The evaluation was performed using standard NLP metrics including ROUGE-1, ROUGE-2, ROUGE-L, and BLEU. Few-shot prompting was implemented using three representative examples appended before the test question in each prompt.

TABLE I
ZERO-SHOT VS FEW-SHOT RESULTS

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**

The results, shown in Table I, indicate a notable performance improvement in the few-shot setting across all metrics.

These improvements are particularly prominent in ROUGE-2 and BLEU scores, which reflect better syntactic overlap and sequence fluency. This suggests that even without domain-specific fine-tuning, prompt engineering alone can substantially enhance LLM performance in Turkish medical QA tasks.

VI. CONCLUSION AND FUTURE WORK

In this progress report, we presented early-stage efforts toward building a reliable medical QA system for Turkish language queries, with an initial focus on the ENT domain. Using a curated dataset of 1,200 doctor-patient Q&A pairs, we evaluated the Mistral-7B-Instruct-v0.2 model in zero-shot and few-shot settings. The few-shot configuration consistently outperformed the zero-shot baseline, indicating the importance of contextual priming in underrepresented languages and domains.

Looking forward, our next steps include:

- **Expanding the dataset** to over 5,000 entries, covering at least 10 medical specialties such as cardiology, dermatology, and endocrinology.
- **Fine-tuning open-source medical LLMs** (e.g., MEDITRON, Hippocrates) using the collected Turkish dataset to improve factual grounding and coherence.
- **Benchmarking against commercial models** like GPT-4 and Med-PaLM 2 (via API access) to evaluate performance gaps.
- **Integrating advanced architectures** such as Retrieval-Augmented Generation (RAG) and Parameter-Efficient Fine-Tuning (PEFT) to further reduce hallucinations and enhance accuracy.
- **Collaborating with clinical professionals** to conduct human evaluations for real-world applicability, factual correctness, and patient safety.

This foundational work paves the way for a scalable, domain-adapted, and language-sensitive QA system capable of supporting medical communication and education in Turkish.

REFERENCES

[1] K. Singhal, S. Azizi, T. Tu, and et al., "Towards generalist biomedical ai," *arXiv preprint arXiv:2304.10512*, 2023.

[2] M. Kheirabadi, R. Tang, and et al., "Mellama: Multi-task instruction tuning for biomedical language models," *arXiv preprint arXiv:2311.16402*, 2023.

[3] Y. Lambert, T. Hildebrandt, and et al., "Meditron: Augmenting llama with medical domain knowledge," *arXiv preprint arXiv:2311.16079*, 2023.

[4] Y. Liu, J. Hu, and et al., "Hippocrates: Enhancing medical domain performance of large language models via continued pretraining, instruction tuning, and direct preference optimization," *arXiv preprint arXiv:2311.08398*, 2023.

[5] L. Tong, W. Du, R. Zhang, and et al., "Med-llms: A comprehensive survey of large language models in medicine," *arXiv preprint arXiv:2310.11459*, 2023.

[6] Y. Wang, Z. Wang, X. Wang, and et al., "Large language models in healthcare: A survey," *arXiv preprint arXiv:2305.15074*, 2023.

[7] Y. Zhang, C. Qian, B. Tang, and et al., "Trustworthy medical question answering: An evaluation-centric survey," *arXiv preprint arXiv:2402.11301*, 2024.