TurkishMedQA-LLM: Specialized Medical Question-Answering Model

Bekir 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—**We present the development of a Turkish medical question-answering (QA) system based on large language models (LLMs). To overcome the lack of domain-specific resources in Turkish, we collected and filtered over 47,000 doctor-patient Q&A pairs from online medical platforms using a multi-agent LLM-based evaluation pipeline. A Turkish-aligned LLM was then fine-tuned using instruction-style supervision and parameter-efficient techniques. While standard metrics such as BLEU and ROUGE showed limited utility, a custom multiple-choice benchmark revealed significant performance gains after fine-tuning. Additional qualitative evaluations demonstrated the model's ability to generate context-aware and stylistically consistent responses. Our findings highlight the feasibility of building reliable medical QA systems for low-resource languages through data curation, targeted adaptation, and prompt design.**

Index Terms—**Turkish NLP, Medical Question Answering, Large Language Models, Instruction Fine-Tuning, In-Context Learning, Low-Resource Languages, Clinical NLP**

I. Introduction

The rise of large language models (LLMs) has transformed the landscape of natural language processing (NLP), enabling machines to engage with human language at an unprecedented level of fluency, coherence, and contextual understanding. In healthcare, this breakthrough has catalyzed advancements in clinical documentation, diagnostic support, medical education, and patient communication. Among the most promising applications is medical question answering (QA), where LLMs act as intelligent agents capable of synthesizing complex clinical knowledge and responding to user queries in natural language.

Recent progress in medical large language models has led to strong results on standardized QA benchmarks, often reaching or even exceeding expert-level performance. However, these advancements have heavily relied on large, high-quality biomedical datasets, which are mostly available in English. This creates a disparity in access for non-English-speaking communities and limits the global applicability of medical AI tools.

Turkish, spoken by over 80 million people, remains significantly underrepresented in clinical NLP research. Its rich morphology, agglutinative grammar, and data scarcity pose unique challenges for model training and evaluation. Moreover, existing Turkish medical datasets are often small in scale, limited in domain coverage, or unstructured. This linguistic and infrastructural gap hampers the development of trustworthy, general-purpose Turkish medical QA systems.

To address this problem, we propose a comprehensive framework for developing a domain-adapted Turkish medical QA system built on instruction-tuned LLMs. Our approach includes: (1) constructing a large-scale, high-quality dataset of Turkish doctor-patient Q&A pairs; (2) filtering the data using a multi-agent LLM-based pipeline to ensure clinical and linguistic adequacy; and (3) fine-tuning a Turkish-aligned instruction-following model using parameter-efficient methods.

Our system supports multiple medical specialties such as cardiology, neurology, dermatology, endocrinology, and otolaryngology, and is designed to serve both clinicians and the general public. We evaluate our approach using both standard surface-level metrics (BLEU, ROUGE) and a custom multiple-choice benchmark for factual consistency. Furthermore, we explore chat-style in-context evaluation techniques to simulate personalized, context-aware doctor–patient interactions.

This work contributes a reproducible and scalable methodology for building medical QA systems in low-resource languages and lays the foundation for broader clinical NLP applications in Turkish.

II. Related Work

Recent progress in medical large language models (MedLLMs) has enabled significant advances in clinical question answering. This section reviews major contributions in the field, including instruction-tuned and domain-pretrained models, large-scale medical foundation models, and recent surveys highlighting trends and challenges in trustworthy QA systems. While most existing efforts have focused on English-language corpora, their methodologies inform the design of language-specific systems like ours. We group the literature into individual model families and surveys, focusing on training strategies, benchmark performance, and applicability to real-world medical tasks.

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 Preprocessing

A. Data Source and Scope

To construct a domain-specific dataset for Turkish medical question answering, we performed structured web scraping on the publicly accessible platform `doktorsitesi.com`, one of the most widely used medical Q&A forums in Türkiye. This site contains real-world doctor-patient interactions, covering a wide range of medical specialties. We targeted 25 distinct specialties including cardiology, dermatology, psychiatry, gynecology, infectious diseases, and more. For each specialty, we aimed to collect approximately 2,500 high-quality question-answer (QA) pairs, resulting in an initial raw corpus of 62,500 QA pairs.

B. Ethical Considerations

Data collection strictly adhered to ethical scraping practices. We respected site-specific `robots.txt` policies, maintained polite request intervals, and ensured that no personally identifiable information (PII) was stored. Additionally, each QA pair was anonymized and only publicly visible doctor names and credentials were retained as metadata.

C. LLM-based Data Filtering and Scoring Pipeline

Given the noisy and unstructured nature of real-world medical Q&A data collected from online platforms, it was crucial to implement a robust quality control pipeline to ensure that only clinically coherent and linguistically fluent examples would be used for training.

To this end, we designed a multi-stage evaluation pipeline powered entirely by Gemini 2.5, a state-of-the-art multimodal LLM with strong instruction-following and medical reasoning capabilities. The pipeline consists of three distinct AI agents, each fulfilling a specific role in the filtering process:

1) **Agent 1: Relevance Filter**
   This agent was tasked with eliminating completely irrelevant or nonsensical answers. Given a Q&A pair, it determined whether the answer addressed the question meaningfully. If deemed irrelevant or too generic, the pair was excluded from further evaluation.
2) **Agent 2: Scoring Evaluator**
   This agent assigned a numerical score between 0 and 10 to each QA pair based on a set of evaluation criteria: grammatical fluency, medical accuracy, contextual relevance, and completeness. Only entries that received a score of ≥ 7 were considered sufficiently high quality.
3) **Agent 3: Fine-Tuning Appropriateness Checker**
   Even among high-scoring pairs, some may not be suitable for supervised training due to ambiguity, length, or overly broad content. This agent answered a binary yes/no question: "Is this QA pair suitable for instruction fine-tuning?" Only pairs that received a "Yes" were retained.

The outputs of Agent 2 and Agent 3 were then aggregated. A QA pair was accepted into the final dataset only if:

- It was not flagged by Agent 1,
- Received a score of 7 or higher from Agent 2,
- And was marked as suitable by Agent 3.

This pipeline was implemented and visualized using a no-code workflow orchestration tool, as shown in Figure 1. Each agent invocation was executed via an API call to Gemini 2.5, with results aggregated and post-processed using custom logic.

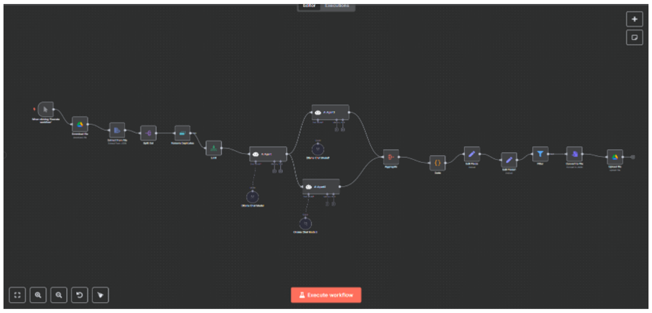

Fig. 1. Three-stage Gemini 2.5-based QA filtering pipeline. The leftmost branch uses Agent 1 to filter out irrelevant answers. In the middle section, the flow splits into two parallel agents: Agent 2 assigns quality scores (0–10), while Agent 3 evaluates fine-tuning suitability. Their results are aggregated before the final filtering and formatting stages.

Using this system, we reduced the noisy raw dataset to approximately 47,000 high-confidence QA pairs, suitable for fine-tuning domain-specific LLMs. This not only improved the factual and linguistic quality of training data but also aligned the examples more closely with the stylistic and structural expectations of supervised instruction tuning.

D. Data Schema and Metadata

Each QA pair is stored using a custom JSON schema with the following fields:

- `question`: The patient's query (in Turkish).
- `answer`: The corresponding doctor's response.
- `specialty`: The medical category (e.g., Neurology, Urology).
- `doctor_info`: A structured sub-object containing the doctor's name, title (e.g., Prof., Uzm. Dr.), clinic name, and area of expertise.

This metadata allowed for fine-grained filtering, stratified evaluation, and category-based analysis in subsequent stages.

E. Preprocessing

The raw HTML content obtained through scraping was processed through a multi-step preprocessing pipeline implemented in Python. The key steps were as follows:

- **Noise Removal:** All HTML tags, escape characters, null bytes, and control characters were removed from the scraped documents using `BeautifulSoup` and regular expressions.
- **Question–Answer Extraction:** Q&A pairs were parsed from embedded JSON-LD blocks (`application/ld+json`) on each question page. We extracted the patient question, doctor response, and topic metadata from the `mainEntity` and `acceptedAnswer` fields.
- **Doctor Metadata Retrieval:** Each answer contained a hyperlink to the doctor's profile page. A secondary scraping pass parsed each unique profile and extracted structured metadata including name, title, specialty, clinic name, location, biography, and average review score. Titles such as "Uzm. Dr." and "Prof. Dr." were normalized using a pattern-matching function.
- **De-duplication:** QA entries were hashed using question–answer string pairs to detect and remove duplicates. Additionally, overly generic or trivial questions (e.g., ”Yardım eder misiniz?”) were filtered out using rule-based heuristics.
- **Validation and Storage:** Cleaned QA pairs were stored in UTF-8 encoded JSON files using a two-part schema: one file for Q&A content, and another for corresponding doctor metadata. Each doctor was assigned a unique UUID to facilitate relational mapping across entries.

F. Dataset Statistics

The final dataset contains approximately 47,000 QA pairs across 25 medical specialties. It is split into 70% training, 15% validation, and 15% test sets using stratified sampling to preserve category balance.

Table I summarizes the key properties of the finalized dataset.

TABLE I
SUMMARY STATISTICS OF THE TURKISH MEDICAL QA DATASET

	Train	**Val**	**Test**
# QA Pairs	32,900	7,050	7,050
Avg. Question Length (tokens)	45.4	46.5	45.3
Avg. Answer Length (tokens)	64.1	64.0	63.8
# Specialties Covered	25	25	25

IV. METHODOLOGY

A. Model Selection and Motivation

As an initial step, we selected the `epfl-llm/meditron-7b` model for fine-tuning, due to its strong performance in the medical QA domain. According to the literature, Meditron-7B is a large-scale language model that has undergone continued pretraining on biomedical corpora, instruction tuning, and reinforcement learning with human feedback (RLHF). It is fully open-source and achieves competitive results across multiple medical benchmarks such as MedQA, MedMCQA, and PubMedQA.

Given its transparency and proven effectiveness in clinical NLP tasks, we deemed it a suitable starting point for adapting to Turkish medical question answering.

However, despite its domain expertise, the Meditron-7B model was originally trained on English-language data. In our experiments, we observed that it struggled to generate fluent and coherent responses in Turkish. This was particularly problematic for tasks requiring patient-oriented explanations, where language quality is critical. As a result, we transitioned to a Turkish-aligned model.

B. Adopting a Turkish-Aligned Medical LLM

To address the limitations of Meditron in handling Turkish input and output, we adopted the `malhajar/Mistral-7B-v0.2-meditron-turkish` model. This model combines the general instruction-following capabilities of Mistral-7B with domain-adaptive continued pretraining on Turkish medical corpora. It thus represents a more suitable base for instruction fine-tuning on our high-quality Turkish QA dataset.

C. Prompt Formatting

In both training and inference, we employed a consistent prompt format to guide the model's behavior. Each sample was structured as:

Soru: `<question text>`
Yanıt: `<answer text>`

This format was chosen for its clarity and alignment with instruction-tuned models. The explicit "Soru" and "Yanıt" tokens act as delimiters and help the model distinguish between the input and the expected output.

D. Instruction Fine-Tuning Process

We applied instruction fine-tuning to further adapt the selected model to Turkish medical question answering. Our fine-tuning dataset consisted of approximately 47,000 high-quality QA pairs, curated from patient–doctor interactions and filtered using an LLM-based quality scoring mechanism. These pairs were divided into training (70%), validation (15%), and test (15%) sets, ensuring balanced representation across 25 medical specialties.

To enable efficient training without requiring full model updates, we used the Low-Rank Adaptation (LoRA) method. LoRA injects a small number of trainable parameters into specific layers, allowing for resource-efficient fine-tuning. We targeted key projection layers such as `q_proj`, `k_proj`, `v_proj`, and others.

Training was conducted for 8 epochs on a single NVIDIA H100 GPU (80GB). The tokenizer sequence length was set to 512 tokens, and causal language modeling was used as the objective. Fine-tuning was implemented using the Huggingface `Trainer` API and PEFT library. Batch size and gradient accumulation steps were tuned to accommodate GPU memory limitations.

The final fine-tuned model achieved a substantial improvement in both fluency and task relevance for Turkish QA tasks, compared to both its untuned base version and the original Meditron-7B model.

V. EXPERIMENTS AND RESULTS

A. Response Generation

Before any quantitative evaluation metrics were applied, model responses were first generated using various decoding configurations to construct a fair and representative evaluation set. Temperature values of 0.4, 0.7, and 1.0 were tested during answer generation for both the base and fine-tuned models. These configurations were manually explored across multiple test samples to assess the influence of temperature on response fluency, creativity, and factual accuracy.

It was observed that low-temperature outputs (e.g., 0.4) tended to be overly deterministic—resulting in rigid, brief, and occasionally repetitive responses with limited linguistic variation. On the other hand, high-temperature generations (e.g., 1.0) produced more diverse outputs, but often introduced hallucinated facts or strayed from medically appropriate phrasing. A medium temperature setting of 0.7 was found to offer the most effective balance: outputs were fluent, contextually appropriate, and aligned with plausible clinical reasoning.

Accordingly, a temperature value of 0.7 was adopted for all subsequent evaluations to ensure consistency and maintain the overall quality of generated answers.

B. BLEU and ROUGE for Baseline Evaluation

In the initial stage of our evaluation, we employed two widely used lexical overlap metrics—**BLEU** and **ROUGE**—to benchmark the quality of model-generated answers against real doctor-written responses. These metrics are standard in machine translation and summarization tasks but often fall short when applied to generative tasks involving semantic reasoning, such as medical QA.

BLEU [8] (Bilingual Evaluation Understudy) computes the geometric mean of n-gram precisions between the generated and reference texts, penalized by a brevity factor:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{1}$$

Here, p_n is the modified n-gram precision, w_n is the weight for each n-gram (typically uniform), and BP is the brevity penalty:

$$\text{BP} = \begin{cases} 1, & \text{if } c > r \\ e^{(1-r/c)}, & \text{if } c \leq r \end{cases} \tag{2}$$

where c and r are the lengths of the candidate and reference answers, respectively.

ROUGE [9] (Recall-Oriented Understudy for Gisting Evaluation) emphasizes recall by computing the overlap between reference and generated texts. For instance, ROUGE-N is defined as:

$$\text{ROUGE-N} = \frac{\sum_{\text{ref}\in\text{Ref}} \sum_{\text{gram}_n \in \text{ref}} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{\text{ref}\in\text{Ref}} \sum_{\text{gram}_n \in \text{ref}} \text{Count}(\text{gram}_n)} \tag{3}$$

These surface-level metrics provide useful initial signals but fall short of capturing semantic correctness—especially in domains like medicine, where the same clinical information can be expressed in multiple valid forms.

For example, the reference answer "*Bu tablo derin ven trombozuna işaret ediyor olabilir*" and the generated response "*Bacak damarınızda pıhtı oluşmuş olabilir*" convey essentially the same medical insight. However, due to differences in wording, such responses may receive low BLEU or ROUGE scores, despite being factually equivalent.

Our empirical results using BLEU and ROUGE on a 10-sample ENT subset are summarized in Table II. While few-shot prompting improved performance compared to the zero-shot setting, the overall scores were modest and did not align well with human judgment of answer quality. Consequently, we opted for a task-specific multiple-choice benchmark as the primary evaluation strategy, discussed in the following subsection.

TABLE II
BLEU AND ROUGE SCORES ON 10-SAMPLE ENT SUBSET

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

C. Transition to Multiple-Choice QA Benchmark

Given the limitations of generative similarity metrics in capturing clinical correctness, we adopted an alternative evaluation setup using multiple-choice question answering. Inspired by prior work such as Meditron [3], which follows a constrained inference format for factual assessment (cf. Meditron paper, p.7), we designed a benchmark where the model is presented with a medical question and four options (A–D). It must select the most appropriate answer by returning only a single letter.

The prompt used during evaluation was standardized as follows:

Aşağıda bir çoktan seçmeli tıbbi soru verilmiştir.
Soruyu dikkatle okuyun ve en doğru cevabı seçin.
Sadece A, B, C veya D harflerinden birini döndürün.
Soru: `<question>`
A) `Option A`
B) `Option B`
C) `Option C`
D) `Option D`
Yanıt: `A`

We constructed a 500-question benchmark covering 25 medical specialties. The model was evaluated in greedy decoding mode without sampling (`temperature=0.0`, `do_sample=False`). Accuracy was computed as the proportion of questions for which the model selected the correct option.

An example from the benchmark is shown below:

Soru (Kalp Damar Cerrahisi): Bacak damarımda
sert kord benzeri şişlik var, ne olabilir?
A) Tromboz
B) Kas spazmı
C) Varis
D) Gut
Yanıt: A

This approach allowed us to directly assess the factual reasoning ability of the model in a controlled setting and reduced the subjectivity associated with open-ended answer evaluation.

D. Accuracy Results

The multiple-choice evaluation yielded the following accuracy results:

TABLE III
MODEL ACCURACY ON MULTIPLE-CHOICE BENCHMARK

Model Variant	Accuracy (%)
Base Model (untuned)	56.32
Fine-tuned (ours)	**62.96**

The fine-tuned model demonstrated a notable gain of over 6 points in accuracy compared to the base model. This validates the effectiveness of our domain-specific instruction fine-tuning and supports the utility of multiple-choice QA as a robust evaluation paradigm in medical NLP for Turkish.

E. Chat-style In-Context Evaluation with Doctor Conditioning

In addition to quantitative metrics, we conducted a qualitative evaluation in a more realistic, chatbot-style setting. This scenario reflects the end-user experience more accurately, where patients pose free-form medical questions and expect fluent, accurate, and context-aware responses from the system.

To simulate this, we designed an **in-context learning (ICL)** setup where the model was presented with a new question along with multiple previously answered questions by the same doctor. The intuition behind this strategy is that by observing how a specific doctor has responded to past queries, the model can emulate that doctor's style, domain expertise, and level of detail. This structure allows the model to act as a personalized assistant conditioned on prior interaction history.

This approach draws on principles from three major research directions:

- **In-Context Learning (ICL)** [10]: where a large language model is shown a few example (question, answer) pairs as part of the input prompt to guide generation.
- **Retrieval-Augmented Generation (RAG)** [11]: where external knowledge or past examples are retrieved based on the current query and prepended to the input to improve factual correctness.
- **Persona-based Prompting** [12]: where the model is conditioned on the previous output style or behavioral traits of a specific speaker to generate personalized responses.

In our implementation, we used the Gemini 2.5 model as the inference engine due to its strong multilingual capabilities and instruction-following performance. For each test question, we retrieved up to three previously answered question–answer pairs by the same doctor from our dataset. These were selected using doctor ID and relevance filtering based on question similarity.

The final prompt presented to the model was structured as:

Soru: [Example Question 1]
Yanıt: [Example Answer 1]
Soru: [Example Question 2]
Yanıt: [Example Answer 2]

...
Soru: [Example Question N]
Yanıt: [Example Answer N]

Soru: [New patient question]
Yanıt:

This format enabled the model to "observe" how a specific doctor typically responds to medical queries, both in tone and in content, before attempting to answer a new question. An example prompt given to Gemini 2.5 is shown below:

Soru: Burun tıkanıklığım sabahları daha kötü oluyor, neden olabilir?
Yanıt: Sabahları artan burun tıkanıklığı genellikle alerjik rinit veya sinüzite bağlı olabilir. Özellikle toz, polen veya ev akarlarına karşı bir duyarlılığınız varsa bu durum sabah saatlerinde belirginleşebilir. Bir kulak burun boğaz uzmanına görünmenizi öneririm.

...
Soru: Son 2 gündür kulağımda uğultu var, gece uykumu bölüyor. Ne olabilir?
Yanıt:

We observed that Gemini 2.5 was able to maintain both linguistic consistency and clinical plausibility in its responses. The generated answers reflected the level of caution and specificity often observed in expert-written replies. Compared to zero-shot prompting, this method significantly improved relevance and tone alignment, even without any fine-tuning on the target model.

Although this evaluation is not quantitatively scored, it offered valuable insight into how LLMs can be adapted to personalized medical QA using simple prompt engineering. It also highlights the potential of combining retrieval and ICL techniques to simulate consistent, doctor-specific behavior in clinical dialogue systems.

VI. Conclusion and Future Work

In this study, we developed a Turkish-language medical question answering (QA) system by leveraging instruction-tuned large language models (LLMs) and a curated domain-specific dataset. Our work addresses the critical gap in non-English clinical NLP resources by constructing a large-scale, ethically sourced dataset of over 47,000 high-quality question–answer pairs across 25 medical specialties. We fine-tuned a Turkish-adapted version of Mistral-7B using parameter-efficient techniques (LoRA) and evaluated its performance under both open-ended and multiple-choice settings.

Initial experiments using BLEU and ROUGE metrics highlighted the limitations of surface-level text similarity in evaluating medical QA tasks. To better assess factual accuracy, we constructed a multiple-choice benchmark inspired by the Meditron framework and demonstrated a 6.6% absolute accuracy improvement over the base model. Furthermore, we explored a chat-style inference setup using Gemini 2.5, where prior doctor-patient interactions were used as context to simulate personalized, context-aware responses. This qualitative evaluation showed that in-context conditioning can significantly enhance the consistency and naturalness of generated answers without requiring additional fine-tuning.

Our results demonstrate that even in low-resource languages such as Turkish, careful dataset construction, domain adaptation, and prompt engineering can collectively yield reliable and clinically useful QA systems. However, challenges remain in assessing factual correctness, aligning model outputs with clinical safety standards, and handling ambiguous or multifaceted queries.

Future work will focus on the following directions:

- **Expert-based Evaluation:** Collaborating with licensed physicians to evaluate model responses for clinical safety, empathy, and factual correctness.

- **Dataset Expansion:** Increasing the dataset to over 100,000 QA pairs and including multilingual data to support broader accessibility.
- **Instruction Diversity:** Incorporating synthetic instruction-rewrite techniques to improve robustness across diverse patient expressions.
- **Retrieval-Augmented Architectures:** Integrating dense retrievers to enable document-grounded generation and reduce hallucinations.
- **Real-time Chatbot Deployment:** Building an interactive QA assistant interface for patients and clinicians, with feedback mechanisms for continuous improvement.

We believe that our work lays the foundation for a scalable, trustworthy, and linguistically inclusive medical AI system tailored to the Turkish-speaking population.

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.

[8] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[9] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004.

[10] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. i. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[11] P. Lewis, E. Perez, A. Piktus, V. Karpukhin, N. Goyal, I. Kulikov, A. Fan, V. Chaudhary, F. Petroni, W.-t. Yih *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Advances in Neural Information Processing Systems*, 2020.

[12] L. Zhou, S. Xie, W. Ma, and N. A. Smith, "Can language models learn from personal interaction histories? an empirical evaluation," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022.

Appendix

Appendix: Task Distribution

The project was collaboratively developed by the following team members:

- **B.B. Tekin** was responsible for converting scraped data into structured JSON format, designing and executing the instruction fine-tuning pipeline using LoRA, conducting zero-shot and few-shot experiments, and evaluating model performance using BLEU/ROUGE metrics. He also formatted the selected data into a supervised training format and led the overall integration of experimental results.
- **M.Y. Tosun** contributed to the initial web-scraping of Turkish medical Q&A content, performed literature review on medical LLMs, and co-led the evaluation of doctor responses for informativeness. He was also jointly responsible for the design and implementation of the chat-style in-context learning evaluation.
- **S. Uçan** worked on scraping and expanding the dataset, implemented the LLM-as-a-Judge scoring pipeline, and co-authored the literature review section. He also collaborated on few-shot prompting experiments and was a key contributor to the in-context evaluation framework using doctor-based prompt conditioning.

All members actively participated in weekly meetings, shared code reviews, and jointly authored the final report and presentation materials.