**Literature Survey Report**

1. Recent advancements in LLMs applied to the medical domain:

* Med-PaLM2:
  + Built on the PaLM 2 model.
  + **Ensemble Refinement (ER):** Combines multiple responses to generate the best possible answer.
  + **Chain of Retrieval (CoR):** Gathers relevant sources before generating answers.
  + These innovative methods enhance the model’s inference and accuracy capabilities. It was tested with challenging datasets such as:
    - **Adversarial QA Dataset:** Includes misleading and uneven questions.
    - **Bedside Consultation Dataset:** Consists of real doctor questions from Stanford.
  + Med-PaLM 2 has shown performance rivaling or even surpassing GPT-4 in many scenarios. It was preferred over doctor responses in 8 out of 9 evaluation criteria.
  + ***Note:*** *Not fully open-source.*
* Hippocrates (Hippo-7B):
  + Hippocrates is a fully open-source framework that contributes significantly to the field of medical LLMs through its Hippo-7B models based on Mistral and LLaMA2. Unlike many other models, all training data, code, checkpoints, and evaluation protocols are openly available.
  + **Continued Pretraining (CPT):** The Hippo models were pretrained on 298 million tokens of specialized medical data, including clinical guidelines, PMC patient summaries, and PubMedQA.
  + **Instruction Tuning (SFT):** The models were fine-tuned with a general medical instruction set containing 292K examples and evaluated using samples from benchmarks such as MedQA.
  + **Preference Optimization (DPO):** Output optimization was conducted using datasets such as iCliniq-10K, with GPT-4 assistance, aligned with medical expert preferences.
  + The Hippo models demonstrated competitive or superior performance compared to 70B parameter models in benchmarks such as MedQA, MedMCQA, PubMedQA, and USMLE. The Hippo-Mistral model, in particular, stood out with an average accuracy of 62.1% in 5-shot learning scenarios.
* MEDITRON‑70B:
  + Based on the LLaMA 2–70B model.
  + **Continued Pretraining (CPT):** Trained on 48.1 billion tokens, including PubMed articles, clinical guidelines, and scientific abstracts.
  + Further fine-tuned using methods such as instruction tuning and RLHF for adaptation to medical QA tasks.
  + Achieved high performance across four major medical benchmarks: superior to GPT-3.5 and closely approaching GPT-4 (within 5%).
  + Fully open-source, including model, data, and training process
* Me‑LLaMA:
  + Built on the LLaMA 2 model.
  + **Continued Pretraining (CPT):** Domain adaptation with 129 billion tokens of medical data.
  + **Instruction Tuning (SFT):** Fine-tuned on 214K examples for QA, diagnosis, and summarization tasks.
  + Resistant to catastrophic forgetting: new knowledge is learned without forgetting previous information.
  + Outperformed GPT-4 in 5 out of 8 benchmarks and surpassed ChatGPT in 7.

2- Domain-specific fine-tuning techniques for LLMs, particularly for question-answering tasks:

* Continued Pretraining (CPT):
  + Retraining a pre-trained LLM with domain-specific (medical) data. Enhances understanding of medical terminology and concepts.
  + ***Examples:*** *Med-PaLM 2, MeLLaMA, MEDITRON*
* Instruction Fine-Tuning (SFT):
  + Training the LLM using task-specific examples in a “question → answer” format.
  + Improves direct task performance.
  + **Example:** 214K examples in MeLLaMA.
* PEFT – Parameter Efficient Fine-Tuning:
  + Instead of updating the entire model, small adapter layers are added.
  + Requires less GPU memory and computational power.
* Retrieval-Augmented Generation (RAG) + Fine-Tuning:
  + Generates answers by accessing external knowledge sources beyond the model’s memory.
  + Methods like CoR (Med-PaLM 2) and Re-RAG improve QA quality.

3- Existing Turkish NLP resources, medical QA datasets, and relevant benchmarks:

Datasets:

* MedTurkQuAD (2024):
  + 20.256 QA pairs.
  + Available via Hugging Face.
  + Covers general health, internal medicine, gynecology, psychiatry, etc
* TRMedQA (2023).
  + 8.000 QA pairs
  + Focused on basic health topics
  + Designed for short-form QA

4- Evaluation methodologies for medical QA systems:

* Exact Match (EM):
  + Checks whether the model’s response matches the ground truth exactly word-for-word.
* F1 Score:
  + Consider the overlap between the model’s output and the ground truth.
  + Captures partial correctness not detected by EM.
* BLEU (Bilingual Evaluation Understudy):
  + Measures n-gram overlap between model and reference answers.
  + Suitable for open-ended or long-text QA with multiple valid answers
  + Does *not* capture semantic similarity.
* ROUGE (Recall-Oriented Understudy for Gisting Evaluation):
  + Evaluates how much of the ground truth is covered in the model's response, focusing on word or full sentence matches.
  + Particularly useful for long-answer QA systems like Med-PaLM 2.
  + It provides sentence-level similarity assessment.

5- Survey papers on medical LLMs and QA to identify foundational models we can use and state-of-the-art approaches:

* Large Language Models in the Medical Domain: A Survey
  + RAG models offer more reliable results than standard fine-tuning.
  + Best QA strategy: CPT + SFT + Human Evaluation.
* Med-LLMs: A Comprehensive Survey on Foundation Models in Healthcare
  + Resource-efficient techniques like PEFT and QLoRA offer strong performance.
  + MeLLaMA stands out with open access and high performance.
  + Most models are trained using a multitasking approach (QA, summarization, diagnosis).
* Trustworthy Medical LLMs: Taxonomy, Challenges, and Solutions
  + Key evaluation criteria: factuality, bias, explainability.
  + Recommends retrieval-based hybrid systems (RAG, CoR, Re-RAG) to prevent hallucinations.
  + Advocates for rule-based response filtering over RLHF.
* Medical QA Systems: Datasets, Evaluation, and Trends
  + Recommend combined use of automatic and expert evaluation.
  + Emphasizes support for multi-turn QA (dialogue systems).