MedMCP

Arshia Vadhani

Nov. 4 2025

Abstract

Recent advances to Large Language Models (LLM's) has enabled them to utilize external tools and interact with different data sources through the Model Context Protocol, allowing for more complex reasoning and the completion of long horizon tasks. However, current evaluations of such LLM's is underdeveloped and benchmarking the success of LLM's in domains such as biomedical research is unexplored. This project proposes MedMCP, a domain-specific benchmark framework for evaluating an LLM's efficacy in using MCP-Integrated tools medical literature search, comparison, and synthesis. MedMCP will simulate a biomedical research assistant by providing access to functions such as `search_pubmed`, `summarize_study`, and `compare_treatments`. Through a series of structured tasks, such as identifying key studies, summarizing clinical outcomes, and comparing treatment efficacy, the benchmark tests the model's ability to reason over complex, evidence-based information. We aim to analyze model performance across different areas: accuracy, tool usage efficiency, and reasoning depth. We seek to reveal where current LLMs succeed and where they lack in medical analysis. The findings aim to inform both future benchmark design and the development of more reliable, domain-aware AI research assistants.

1 Motivation

1.1 Problem Statement

The goal of this project is to evaluate LLMs' ability to handle complex, long-horizon tasks regarding biomedical research.

1.2 Background and Context

Recent advances in large language models (LLMs) have enabled them to process and generate natural language with increasing fluency and accuracy. Beyond text generation, LLMs are now being integrated with external tools and data sources through protocols such as the Model Context Protocol (MCP), which

allows models to perform grounded reasoning and execute real-world tasks. This shift represents a broader move toward tool-supplemented artificial intelligence, where models can interact with APIs, databases, and computational functions to extend their capabilities.

In the biomedical and clinical research domains, information retrieval and evidence synthesis remain human labor-intensive processes. Researchers typically rely on databases such as PubMed or ClinicalTrials.gov to locate studies, extract key findings, and compare treatment outcomes. Traditional text-mining pipelines and retrieval systems, while effective for keyword searches, struggle to perform multi-step reasoning or integrate diverse evidence sources [Lee et al.(2020)Lee, Yoon, Kim, Kim, Kim, So, and Kang, Beltagy et al.(2019)Beltagy, Lo, and Cohan]. As LLMs have demonstrated potential for summarizing medical literature and assisting with literature reviews [Singhal et al.(2023)Singhal, Azizi, Tu, and et al., Wang et al.(2023)Wang, Zhang, and Xu], questions remain about their reliability, factual grounding, and ability to utilize multiple tools effectively.

The emergence of MCP offers a new opportunity to evaluate these tools based on their accuracy and usability. By exposing domain-specific tools—such as `search_pubmed`, `summarize_study`, and `compare_treatments`—to an LLM through specific APIs, we can measure how well the model performs complex, tool-based tasks for biomedical reasoning. This is relevant as research moves toward multi-agent and multimodal AI systems that depend on accurate, interpretable reasoning over domain data.

1.3 Limitations of Existing Approaches

However, current approaches still suffer from several deep limitations when applied to biomedical research. Traditional literature search and text-mining tools assume that researchers can manually create queries and filter results, which becomes increasingly difficult as the number of publications grows. Even modern LLMs, without utilizing tool-integration, struggle with multi-step reasoning required to synthesize evidence across multiple studies, often producing made-up, inaccurate, or incomplete information. Existing evaluation frameworks are also limited, focusing on singular summarization or retrieval tasks rather than complex, dynamic reasoning over multiple sources. Additionally, domain-specific knowledge and tool coordination remain major challenges, particularly when models encounter unfamiliar data formats or new biomedical APIs. These gaps highlight the need for a benchmark that tests LLMs' ability to reason and act through structured, domain-specific tools.

1.4 Objectives

The objectives of this proposal are to...

1. Develop MedMCP a domain-specific benchmark for evaluating LLM performance on biomedical literature retrieval and synthesis tasks using MCP-integrated tools.

2. Implement a set of structured MCP tools, including `search_pubmed`, `summarize_study`, and `compare_treatments`, to simulate a biomedical research assistant.

3. Evaluate the model's accuracy, reasoning depth, and tool usage efficiency across multi-step tasks.

4. Analyze shortcomings and identify scenarios where LLMs struggle with tool coordination, long-horizon reasoning, or evidence integration.

2 Approach

2.1 Overview

Our proposed framework consists of the following components:

- **Input:** Task instructions specifying biomedical queries or comparative research questions.
- **LLM Agent:** A large language model integrated with the MCP framework, capable of reasoning over multiple steps and calling domain-specific tools.
- **MCP Tools:** Functions such as `search_pubmed`, `summarize_study`, and `compare_treatments` to retrieve, process, and analyze biomedical literature.
- **Output:** Structured responses including summarized findings, comparative tables, and citations from the simulated database.

2.2 Design Principle

The design emphasizes modular, tool-augmented reasoning. By separating the LLM from the tool modules, we can systematically test the model's ability to plan, integrate, and execute outputs based on the results from multiple sources. This approach also ensures reproducibility and allows benchmarking across different LLMs or prompting strategies. It mirrors real-world scientific workflows, where research requires consistent retrieval, analysis, and summarization.

3 Experiments

3.1 Experimental Setup

Datasets / Benchmarks: We will create a synthetic biomedical knowledge base simulating PubMed articles, clinical trial summaries, and treatment outcomes. The dataset will include 500+ studies across several medical domains (e.g., cardiology, oncology, endocrinology). This dataset will be sufficient to

test multi-step reasoning and tool usage while avoiding privacy issues with real patient data.

Environment: Experiments will be run on a workstation with a GPU using PyTorch and the MCP framework. The LLM agent will be connected to the MCP server hosting the synthetic tools.

Baselines and Metrics: - Compare performance with plain LLM prompting without MCP integration. - Metrics: task success rate, number of tool calls, accuracy of synthesized information, and reasoning depth (measured by multi-step completion). - Ablations: testing each tool individually to assess contribution to overall performance.

Timeline

- **Month 1:** Build synthetic biomedical dataset and implement MCP tool functions (`search_pubmed`, `summarize_study`, `compare_treatments`).
- **Month 2:** Integrate the LLM agent with the MCP framework and develop structured task instructions for evaluation.
- **Month 3:** Conduct experiments, collect logs, and perform initial analysis on task success, reasoning depth, and tool usage.
- **Month 4:** Finalize results, create visualizations, write the report, and summarize key findings and contributions.

Expected Contributions

- A benchmark framework, MedMCP, for evaluating LLM reasoning over domain-specific biomedical tasks.
- A set of MCP tools that simulate a biomedical research assistant for literature search and synthesis.
- Insights into the limitations of LLMs in long-horizon, tool-augmented reasoning in the biomedical domain.
- Recommendations for designing more reliable and interpretable LLM-based research assistants.

References

[Beltagy et al.(2019)Beltagy, Lo, and Cohan] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[Lee et al.(2020)Lee, Yoon, Kim, Kim, Kim, So, and Kang] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[Singhal et al.(2023)Singhal, Azizi, Tu, and et al.] Karan Singhal, Shekoofeh Azizi, Thang M Tu, and et al. Large language models encode clinical knowledge. *Nature*, 620(7976):172–180, 2023.

[Wang et al.(2023)Wang, Zhang, and Xu] Li Wang, Yue Zhang, and Jin Xu. Evaluating large language models on medical evidence summarization. *arXiv preprint arXiv:2305.11212*, 2023.