MARRVEL-MCP Revision

Hyun-Hwan Jeong

Liu Lab Meeting

February 4, 2026

Reviewer #1: General Comments

> *This manuscript describes MARRVEL-MCP, an AI-agent system that automatically navigates appropriate biomedical resources to answer rare disease questions. This is a* ***very timely and important topic*** *for Mendelian disease discovery in the era of AI. The reviewer has several suggestions for improving the manuscript.*

6 specific comments:

1. Related work on biomedical AI agents
2. Clarification of "context engineering"
3. Benchmark dataset and evaluation details
4. Comparison with commercial AI tools (GPT, Gemini)
5. Title wording clarification
6. Usability improvement evidence

R1.1: Related Work on Biomedical AI Agents

Reviewer comment:

> *The authors should consider including more relevant work on biomedical AI agents (e.g., BioMNI). It would also be helpful to compare MARRVEL-MCP with existing biomedical AI agents in terms of architecture, performance, speed, etc., to better highlight the uniqueness of the proposed system for rare diseases.*

Action items:

- Add BioMNI and other biomedical AI agent references
- Include comparison table: architecture, performance, speed
- Emphasize rare disease specialization as differentiator

R1.2: Clarification of “Context Engineering”

Reviewer comment:

> *The term context engineering has been used to describe the next generation of RAG systems. Its definition typically extends beyond selecting appropriate input resources, which appears to be the main definition used in this study. Please ensure that the usage in the manuscript is precise and consistent with existing terminology.*

Action items:

- Review broader definition of context engineering (RAG 2.0)
- Clarify our usage vs. the general definition
- Ensure consistent terminology throughout manuscript

R1.3: Benchmark Dataset and Evaluation Details

Reviewer comment:

> *Please provide additional details about the benchmark dataset and evaluation pipeline. What criteria were used to define "answerable" questions? Were answers manually generated by experts? When Claude Sonnet was used as an external judge, did it compare with expert-generated reference answers, or rely on its own internal knowledge?*

Action items:

- Detail criteria for "answerable" questions
- Clarify expert answer generation process
- Explain Claude Sonnet judge: reference answers vs. internal knowledge
- Add evaluation pipeline diagram or flowchart

R1.4: Comparison with Commercial AI Tools

Reviewer comment:

> *The authors should include the performance of state-of-the-art commercial AI tools such as GPT and Gemini on the benchmark dataset. If these systems already achieve strong performance, please justify the added value of MARRVEL-MCP (e.g., smaller models, lower cost, domain customization, transparency, or reproducibility).*

Action items:

- Benchmark GPT-4o / GPT-o1 and Gemini on evaluation set
- Report performance comparison table
- Justify added value: domain tools, cost, transparency, reproducibility

R1.5: Title Wording Clarification

Reviewer comment:

> *Is MARRVEL-MCP primarily a question-answering system, or does it function as a chatbot that supports multi-turn conversations? I suggest replacing “NATURAL-LANGUAGE QUERY-TO-RESPONSE INTERFACE” in the title with “Question-Answering System” or “Chatbot” to improve clarity and precision.*

Action items:

- ▶ Decide: QA system vs. chatbot vs. current wording
- ▶ Clarify single-turn vs. multi-turn capability in manuscript
- ▶ Revise title accordingly

R1.6: Usability Improvement Evidence

Reviewer comment:

> *The authors have noted that user-friendliness is one of the motivations for developing MARRVEL-MCP. However, is there any evidence showing that MARRVEL-MCP provides outputs that are more beneficial to users? If there has not been a formal user study, the authors should consider addressing this point briefly in the discussion section.*

Action items:

- Acknowledge lack of formal user study (if applicable)
- Add discussion paragraph on usability benefits
- Consider citing indirect evidence or anecdotal feedback
- Mention user study as future work

Reviewer #2: General Comments

> *I think this paper is on the right track, but right now it feels more like a **solid engineering effort** than a big conceptual leap. The MCP tools are genuinely useful, but the manuscript could be more honest about what is new versus what is just well-integrated. The MCP layers the authors built are probably the **most valuable part** of the work. They make it much easier for LLMs to interact with biological databases like dbNSFP, which is great for the community. The tooling infrastructure is the most compelling contribution, but the **system-level claims need stronger justification** and clearer evidence.*

6 specific comments:

1. Context engineering already in most agent systems
2. “Hard-coded” vs. non-hard-coded clarification
3. Evaluation is thin; hand-picked examples
4. Which MCP tools help the most?
5. Error types count mismatch (Sec. 3.3)
6. Gold standard reliability concern

R2.1: Context Engineering in Agent Systems

Reviewer comment:

> *In most agent systems, context engineering is already part of the architecture, so the comparison in the intro is confusing.*

Action items:

- Clarify what is novel about our context engineering approach
- Distinguish from standard agent-system context handling
- Rewrite intro comparison to avoid confusion

R2.2: Hard-Coded vs. Non-Hard-Coded

Reviewer comment:

> *Authors emphasized not "hard-coded"; I partially agree. But given many coding is implemented in the MCP layer (providing interfaces), I won't say it is completely non hard-coded. The improvement of MCP over the baseline is because of the "hard-coding" that exposes the abstract layer usable by LLM agents. A better way to describe: MCP is certainly hard-coded, but the workflow is not.*

Action items:

- Adopt nuanced framing: MCP tools are hard-coded interfaces, but the **workflow/reasoning** is not hard-coded
- Revise manuscript language accordingly

R2.3: Evaluation Concerns

Reviewer comment:

> *The current evaluation is a bit thin. The examples in Section 3.1 are hand-picked and likely biased. For the 45 questions, it's not clear which ones actually need MCP tools and which ones don't. I'd really like to see cases when questions do not need MCP tools, how it performed comparing to the baseline.*

Action items:

- Categorize 45 questions: MCP-required vs. MCP-optional
- Report performance stratified by category
- Show baseline performance on non-MCP questions
- Address selection bias in Section 3.1 examples

R2.4: Tool Usage Analysis

Reviewer comment:

> *It would be helpful to know which MCP tools actually help the most? How many tool calls are needed per question? Whether performance drops as tool usage increases?*

Action items:

- Report per-tool contribution to answer quality
- Analyze distribution of tool calls per question
- Investigate performance vs. number of tool calls
- Add tool usage breakdown figure or table

R2.5: Error Types Count Mismatch

Reviewer comment:

> *There’s a small error where the paper says there are three error types but only lists two. (Section 3.3: “We categorized errors into three types based on their underlying cause.”)*

Action items:

- Fix: either add the missing third error type or correct the count
- Proofread Section 3.3 thoroughly

R2.6: Gold Standard Reliability

Reviewer comment:

> *As LLMs become stronger, the reliability of purely human-curated “gold standard” answers becomes less clear. It would be helpful to know whether the authors manually reviewed LLM-generated “error” answers to confirm they are truly incorrect, rather than alternative valid answers.*

Action items:

- Manually review LLM answers marked as “errors”
- Report how many “errors” were actually valid alternative answers
- Discuss limitations of human-curated gold standards
- Consider inter-annotator agreement or adjudication process