EVALUATION REPORT

Agentic Research Assistant

Author: Kishore Balaji
Course: Building Agentic Systems
Date: November 2025

---

1. Test Questions Used

ID	Question
Q1	What are the challenges in deploying large language models in healthcare?
Q2	How is AI used in cybersecurity threat detection?
Q3	What are the risks of autonomous vehicles?

---

2. Evaluation Metrics

We used a **1–5 scoring scale** for the following criteria:

- **Relevance** – how on-topic the response is
- **Correctness** – factual accuracy
- **Clarity** – readability and structure
- **Completeness** – coverage of key points
- **Robustness** – ability to recover from tool failures

---

3. Results

Individual Question Scores

Question	Relevance	Correctness	Clarity	Completeness	Robustness
Q1	5	4	5	4	5
Q2	4	4	5	4	5

Q3	5	4	4	4	5

Average Scores

Metric	Average Score
Relevance	4.67
Correctness	4.00
Clarity	4.67
Completeness	4.00
Robustness	5.00

---

4. Behavioral Observations

- System performs consistently even when Serper fails
- Custom tool improves high-quality analysis output
- Final writer agent produces clean and human-like answers

---

5. Limitations Identified

- Dependent on LLM internal knowledge when search fails
- No long-term memory across sessions
- Search tool reliability varies by API quota

---

6. Recommendations

- Add critic agent for auto-evaluation loop
- Add knowledge graph or vector DB for persistent memory
- Add parallel execution for speed

---

7. Conclusion

The Agentic Research Assistant demonstrates strong performance across all evaluation metrics, with

particularly excellent robustness (5.00) and high scores in relevance (4.67) and clarity (4.67). The system's ability to maintain consistent performance even during tool failures validates the effectiveness of the fallback mechanisms implemented in the ResearchController.

The areas of correctness and completeness, while still strong at 4.00, represent opportunities for enhancement through the recommended improvements, particularly the addition of a critic agent and persistent memory systems.

End of Evaluation Report