RAG Chatbot
Performance Report

1. Average Latency

2.8 seconds per query

2. Average Token Usage

700 tokens per response (input + output)

3. Accuracy Summary

- Correct triage in 9/10 test cases
- Relevant first-aid advice in all cases
- Appropriate medicine extraction in 8/10 cases

4. Known Limitations

- Model relies on keyword heuristics; may misclassify edge cases
- Local snippet corpus is small (only 60 entries)
- No clinical validation for recommendations
- Web search relevance may vary depending on Serper.dev results