LangSmith Evaluation Report: Multi-Source RAG Chatbot

Sami RAJICHI

April 18, 2025

1 Introduction

This report summarizes the results of an evaluation run performed on the Multi-Source RAG Chatbot (meta-llama/llama-4-maverick-17b-128e-instruct), utilizing LangSmith for tracing and metrics collection. The evaluation involved processing 50 benchmark questions designed to test the chatbot's dynamic routing capabilities across its different knowledge sources (LLM Native, Vectorstore, Web Search). The primary goal was to assess routing accuracy, response latency, and overall system stability.

2 Evaluation Summary Metrics

The following key metrics were observed during the evaluation run comprising 50 questions:

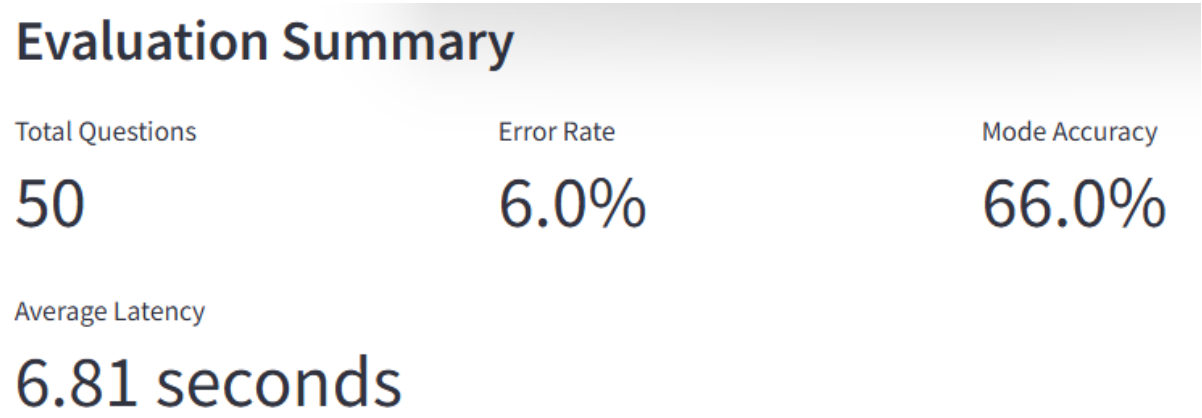

Figure 1: Evaluation Summary Metrics

- **Total Questions Processed:** 50
- **Error Rate:** 6.0%
  - This indicates that 3 out of the 50 questions resulted in an error during processing.
- **Mode Accuracy:** 66.0%
  - Calculated based on the 47 successful runs (50 total - 3 errors). This metric reflects how often the chatbot's dynamic router selected the pre-defined 'expected_mode' for a given question. While demonstrating a capability to route correctly most of the time, there is room for improvement in routing logic or prompt tuning.
- **Average Latency:** 6.81 seconds
  - The average time taken to process a query and generate a response.
- **P50 Latency (from LangSmith):** Approx. 4.4 seconds (Median Latency)
  - 50% of the queries were processed faster than this time.
- **P99 Latency (from LangSmith):** Approx. 24.6 seconds (99th Percentile Latency)
  - Indicates that 99% of queries completed within this time, highlighting that a small percentage of queries experienced significantly higher latency. This could be due to factors like cold starts, web search delays, or complex LLM generations.

3 Visualizations and Interpretation

Screenshots from the Streamlit evaluation dashboard and LangSmith provide visual insights into the performance.

3.1 Mode Accuracy

Figure 2 clearly shows the comparison between correctly and incorrectly routed queries among the successful runs.

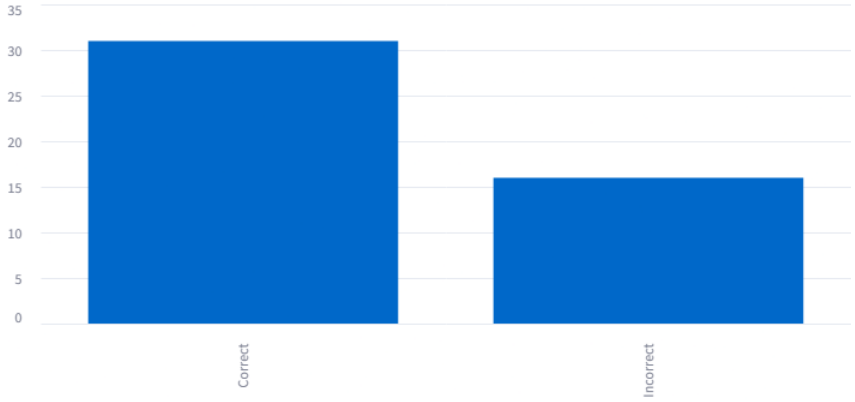

Figure 2: Routing Mode Correctness (Successful Runs)

Interpretation: The chart visually confirms the 66.0% accuracy, showing a higher number of correct routing decisions compared to incorrect ones, but also indicating a significant portion of queries where the router's choice did not match the expectation.

3.2 Latency Distribution

Figure 3 illustrates the latency for each question processed during the evaluation run.

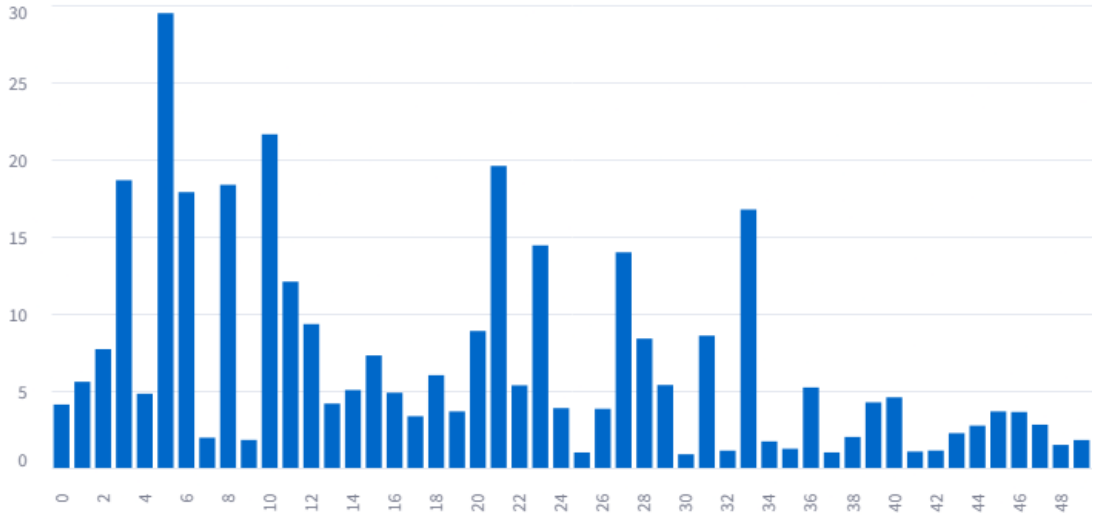

Figure 3: Latency per Question (Seconds)

Interpretation: The latency varies considerably between questions, aligning with the difference between average (6.81s) and P99 (24.6s) latencies. Peaks suggest certain queries triggered more time-consuming operations (potentially web search or Vectorstore).

3.3 LangSmith Run Monitoring

Figure 4 displays the trace count and success rates monitored within LangSmith during the evaluation period.

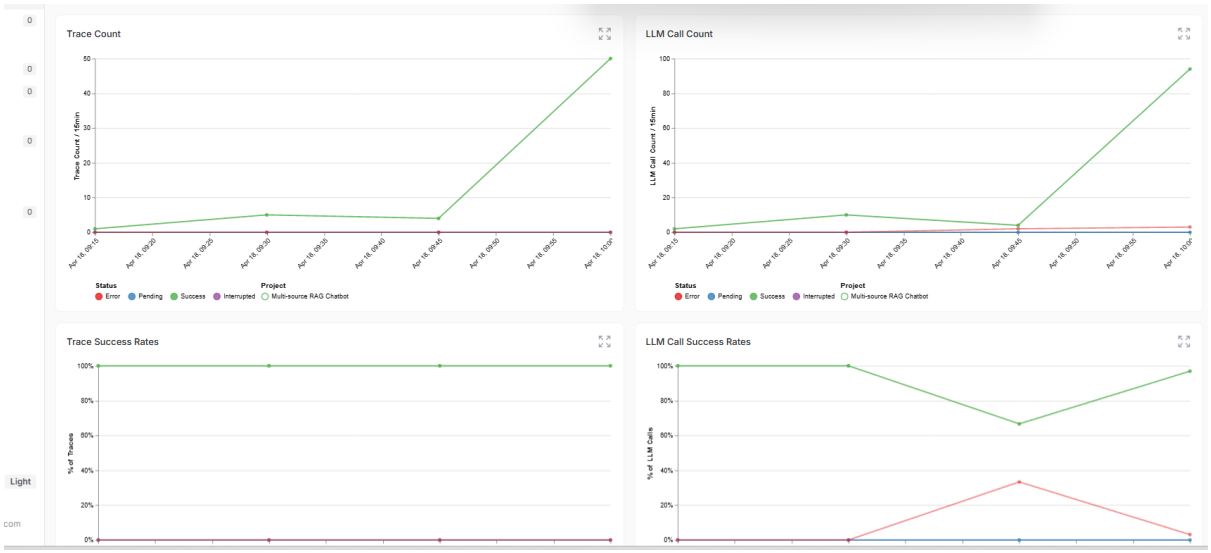

Figure 4: LangSmith Monitoring: Trace Count and Success Rates

Interpretation: The top charts show processing activity (traces and LLM calls) concentrated during the evaluation run time. The bottom charts indicate a high Trace Success Rate (100%) and a generally high LLM Call Success Rate, although the slight dip in LLM success might correlate with the observed 6.0% overall error rate, potentially indicating transient LLM API issues.

3.4 LangSmith Run Details

Figure 5 provides a snapshot of the detailed run table within LangSmith, showcasing individual query traces and logged feedback.

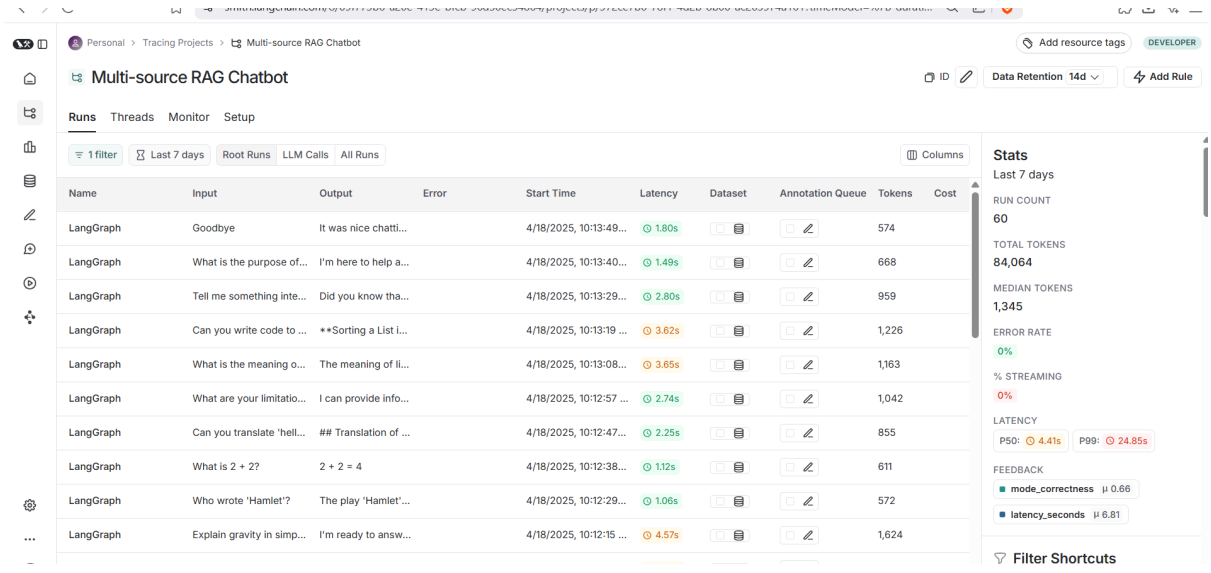

Figure 5: LangSmith Runs Table with Logged Feedback

Interpretation: This view demonstrates the power of LangSmith for detailed analysis. Each

row represents a single chatbot invocation. Key information like the input query, generated output, latency, token usage, and crucially, the programmatically logged feedback (‘mode_correctness‘ score average ≈ 0.66, ‘latency_seconds‘ average ≈ 6.81s) are visible. This allows for granular debugging and performance tracking.

4 Key Observations and Notes

- **Error Rate Impact:** As noted, the Mode Accuracy calculation (66.0%) correctly excludes the runs that resulted in an error (6.0% of total runs). Errored runs are treated as failures, not incorrect routing decisions.
- **Potential Error Sources:** The observed errors are likely attributable to external factors, primarily potential ‘503 Internal Server Error‘ responses from the Groq API. This is a common characteristic when relying on free-tier API services, which may experience high load or temporary outages.
- **Model Variability:** These results are specific to the LLM model selected during the evaluation run (meta-llama/llama-4-maverick-17b-128e-instruct). Users are encouraged to utilize the application's sidebar configuration to select different Groq models and re-run the evaluation to compare performance variations in terms of latency, accuracy, and potentially error rates across models.
- **LangSmith for Deeper Analysis:** The LangSmith platform offers capabilities beyond this summary. Users can interactively explore individual traces, examine the inputs and outputs of each step within the LangGraph, filter runs based on latency or feedback scores, and identify specific queries that failed or were routed incorrectly.

5 Conclusion

The LangSmith evaluation provides valuable quantitative insights into the Multi-Source RAG Chatbot's performance. While demonstrating reasonable routing accuracy and average latency, the results also highlight areas for potential improvement, particularly in routing consistency and handling latency variations. The 6.0% error rate underscores the dependency on external API stability (like Groq). LangSmith proved essential in capturing detailed traces and metrics, enabling effective monitoring, debugging, and performance analysis.