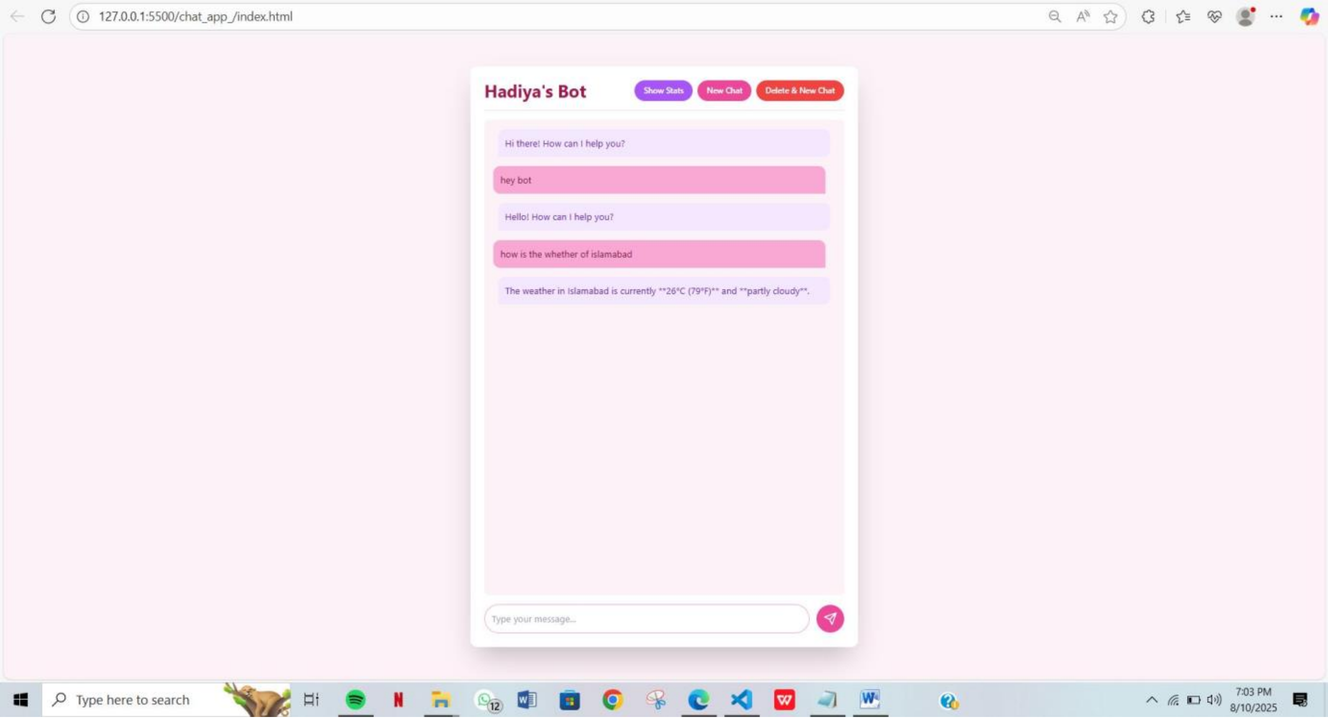
HADIYA’S BOT

1. Simple HTML page with streaming response

Implemented a single-page HTML client that opens a streaming connection to the backend and renders incremental chat/response tokens in real-time.



Hi there! How can I help you?

hey bot

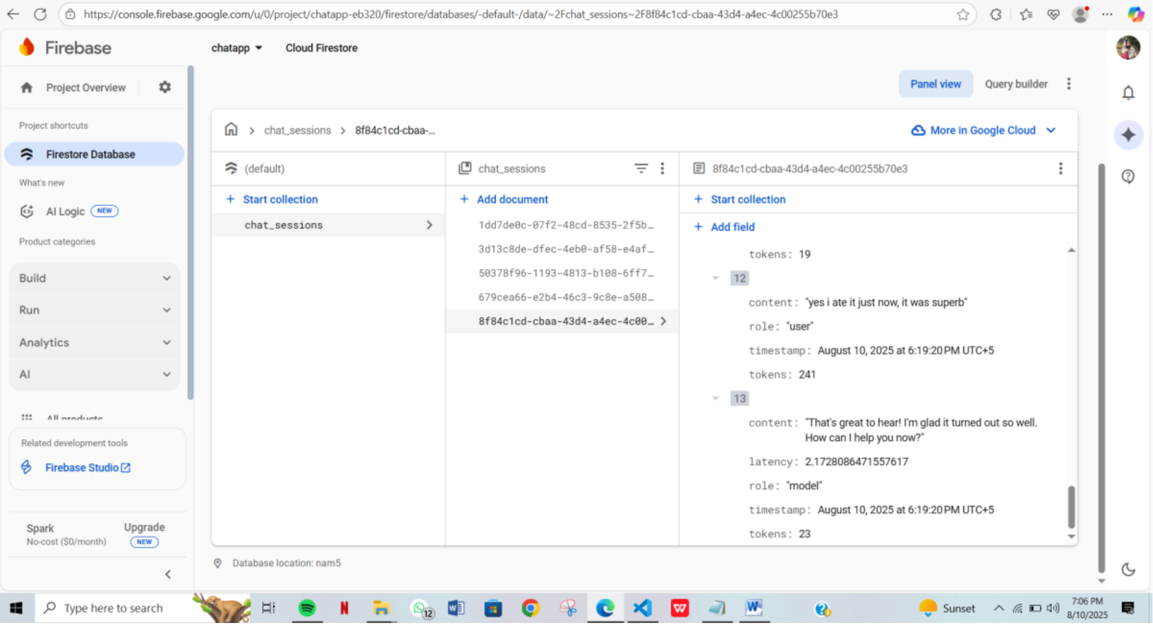
Hello! How can I help you?

how is the whether of islamabad

The weather in Islamabad is currently **26°C (79°F)** and **partly cloudy**.

2. Save cost in DB

Costs (per request) are recorded to the database after each completed call: model name, prompt tokens, completion tokens, total cost (calculated), timestamp, and request id. Designed schema to support querying and aggregation.



chat_sessions > 8f84c1cd-cbaa-...

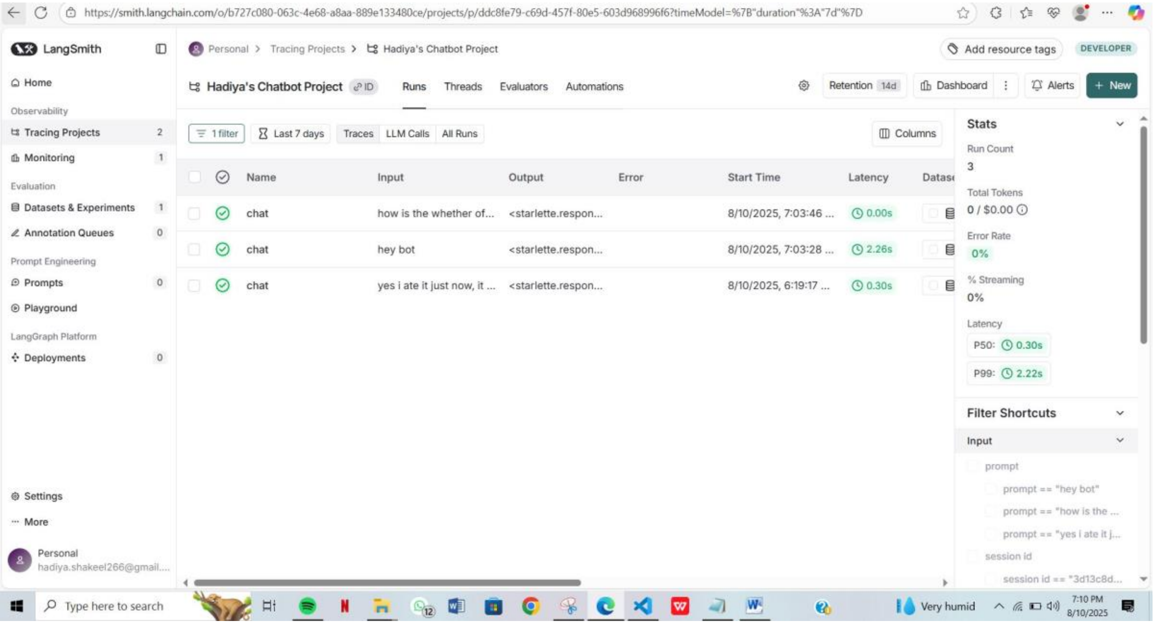
8f84c1cd-cbaa-43d4-a4ec-4c00255b70e3

```
tokens: 19
12
  content: "yes i ate it just now, it was superb"
  role: "user"
  timestamp: August 10, 2025 at 6:19:20 PM UTC+5
  tokens: 241
13
  content: "That's great to hear! I'm glad it turned out so well.
            How can I help you now?"
  latency: 2.1728086471557617
  role: "model"
  timestamp: August 10, 2025 at 6:19:20 PM UTC+5
  tokens: 23
```

Database location: nam5

3. system prompt (fetched from LangSmith)

System prompt is fetched at conversation start from LangSmith and injected as the system message for every chat session.



Personal > Tracing Projects > Hadiya's Chatbot Project

Hadiya's Chatbot Project

Name	Input	Output	Error	Start Time	Latency
chat	how is the whether of...	<starlette.respon...		8/10/2025, 7:03:46 ...	0.00s
chat	hey bot	<starlette.respon...		8/10/2025, 7:03:28 ...	2.26s
chat	yes i ate it just now, it ...	<starlette.respon...		8/10/2025, 6:19:17 ...	0.30s

Stats

Run Count
3

Total Tokens
0 / $0.00

Error Rate
0%

% Streaming
0%

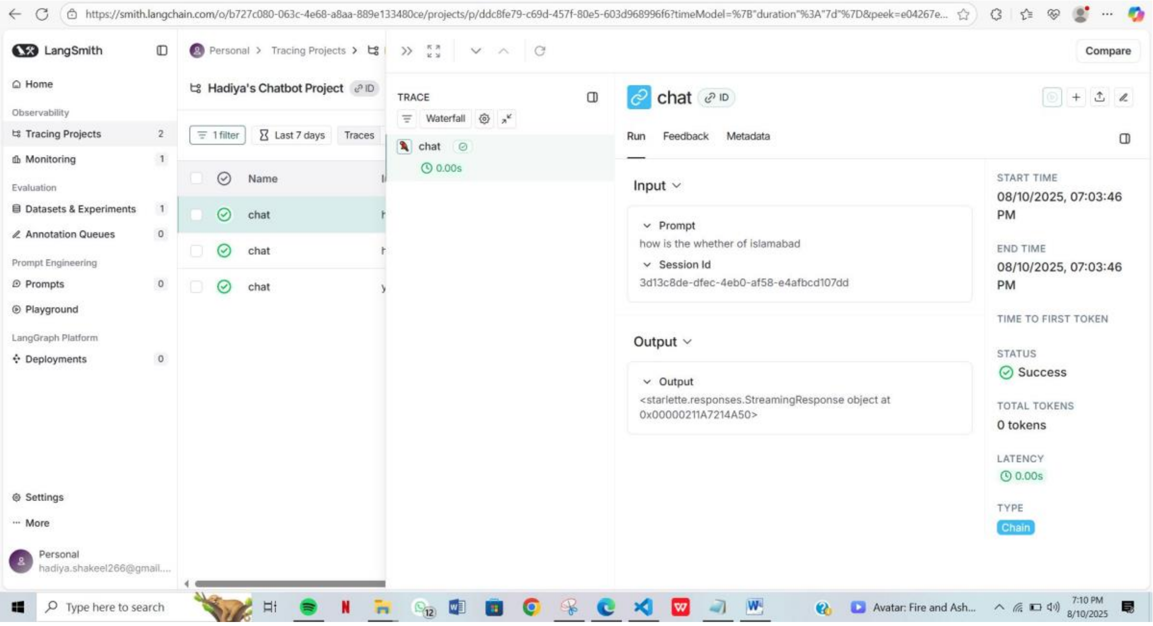
Latency
P50: 0.30s
P99: 2.22s

Filter Shortcuts

Input

prompt
prompt == "hey bot"
prompt == "how is the ...
prompt == "yes i ate it j...
session id
session id == "3d13c8d...

session.



Input

Prompt
how is the whether of islamabad

Session Id
3d13c8de-dfec-4eb0-af58-e4afbcd107dd

Output

Output
<starlette.responses.StreamingResponse object at 0x00000211A7214A50>

START TIME
08/10/2025, 07:03:46 PM

END TIME
08/10/2025, 07:03:46 PM

TIME TO FIRST TOKEN

STATUS
Success

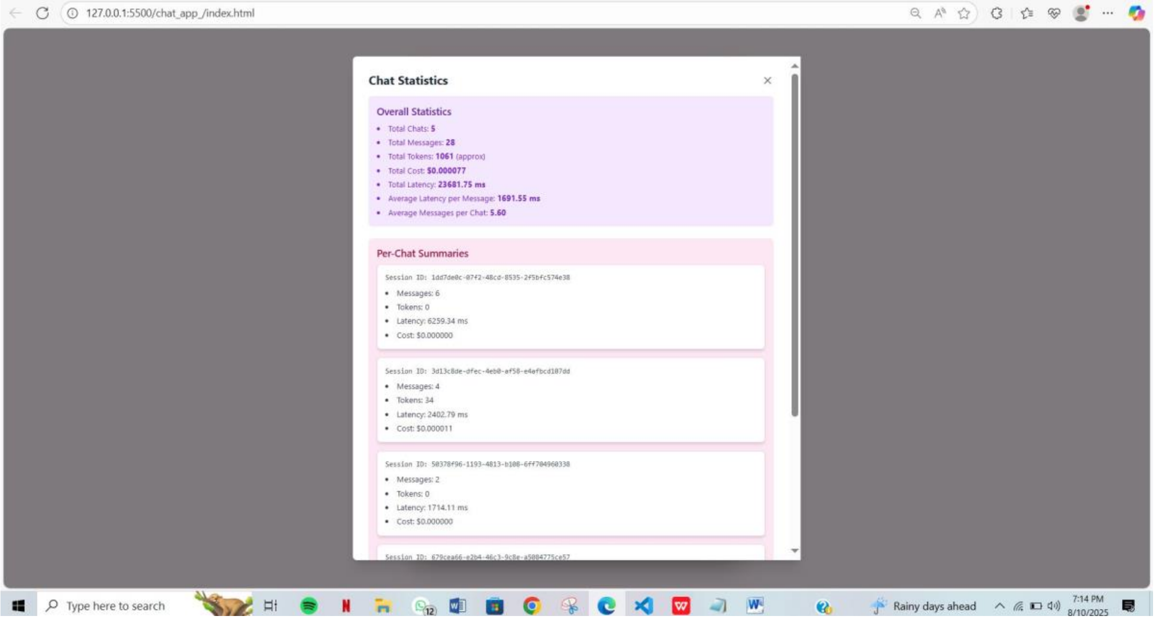
TOTAL TOKENS
0 tokens

LATENCY
0.00s

TYPE
Chain

4. Endpoint for statistical analysis (total cost, tokens used, latency)

Created a /stats endpoint that returns aggregated metrics: total cost, total tokens (prompt & completion), average latency per step.



Chat Statistics

Overall Statistics

- Total Chats: **5**
- Total Messages: **28**
- Total Tokens: **1061** (approx)
- Total Cost: **$0.000077**
- Total Latency: **23681.75 ms**
- Average Latency per Message: **1691.55 ms**
- Average Messages per Chat: **5.60**

Per-Chat Summaries

Session ID: 1dd7de0c-07f2-48cd-8535-2f5bfc574e38

- Messages: 6
- Tokens: 0
- Latency: 6259.34 ms
- Cost: $0.000000

Session ID: 3d13c8de-dfec-4eb0-af58-e4afbcd107dd

- Messages: 4
- Tokens: 34
- Latency: 2402.79 ms
- Cost: $0.000011

Session ID: 50378f96-1193-4813-b108-6ff704960338

- Messages: 2
- Tokens: 0
- Latency: 1714.11 ms
- Cost: $0.000000

Session ID: 679cea66-e2b4-46c3-9c8e-a5884775ce57

.