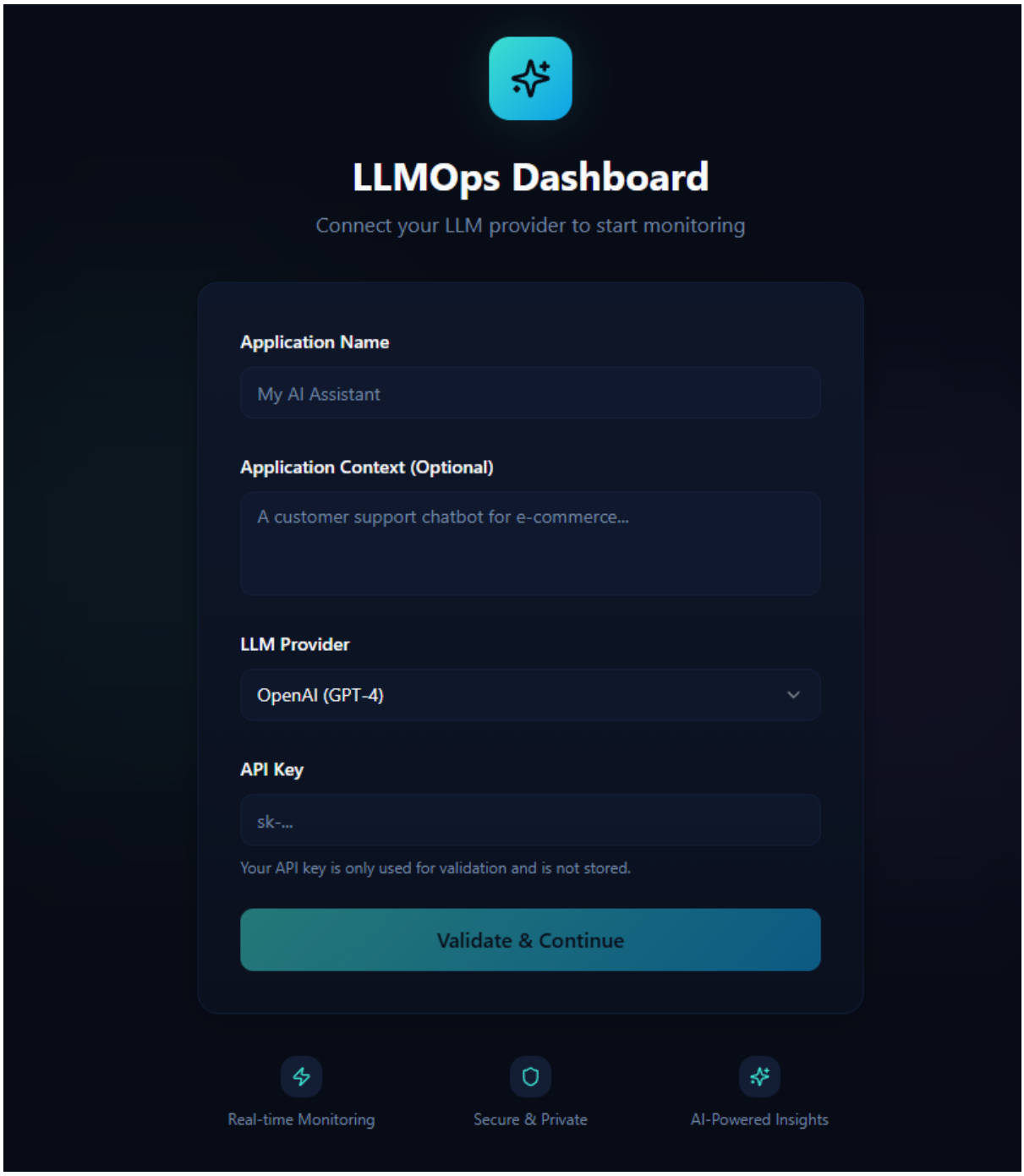
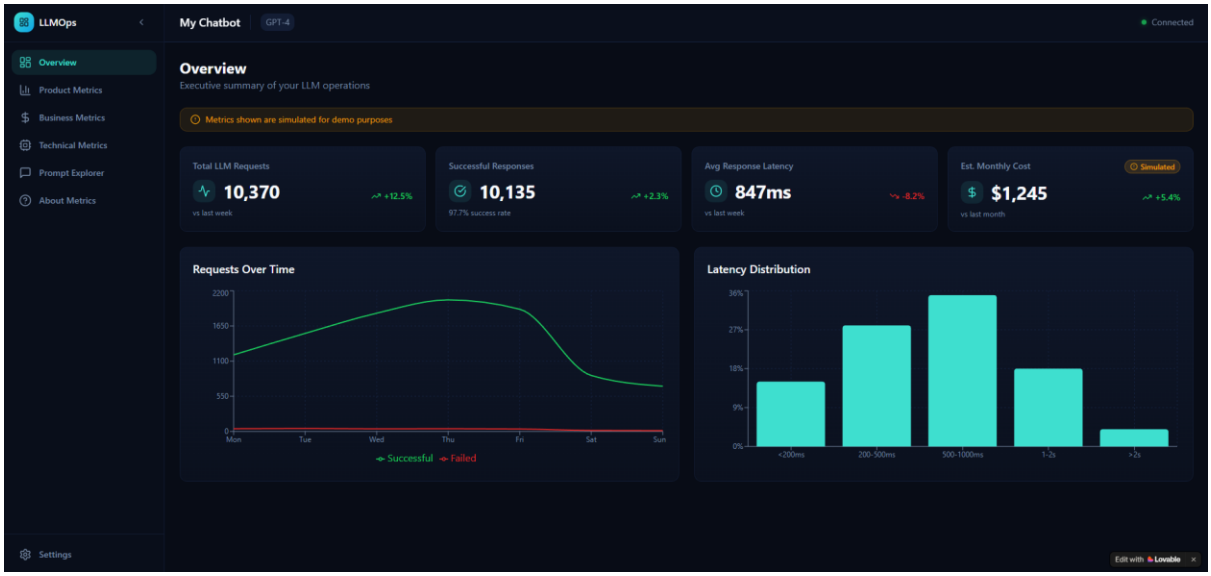
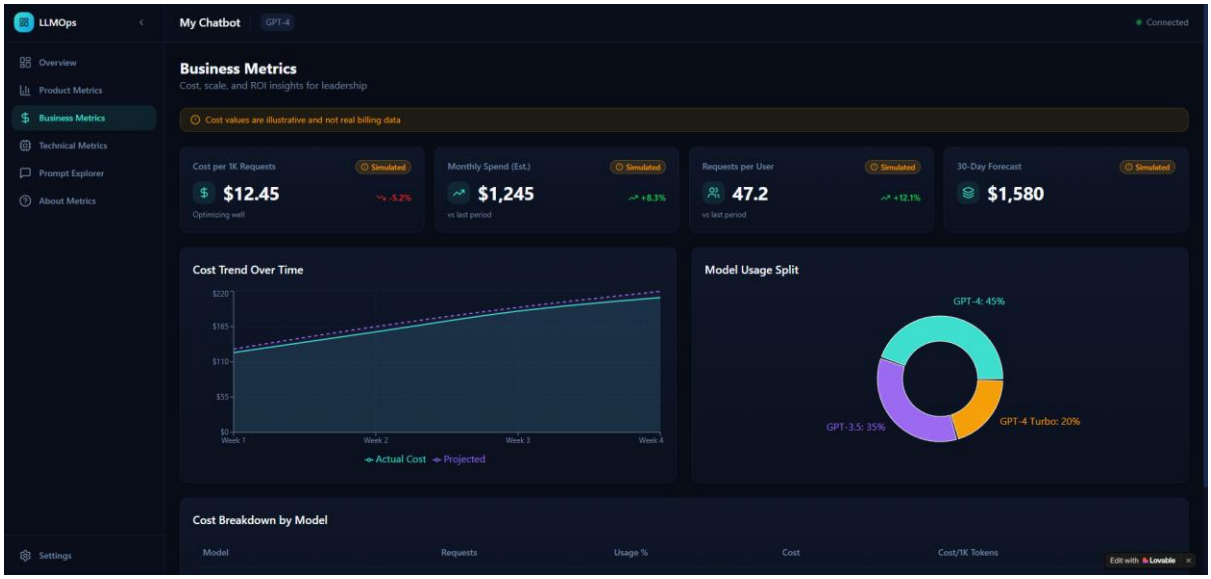
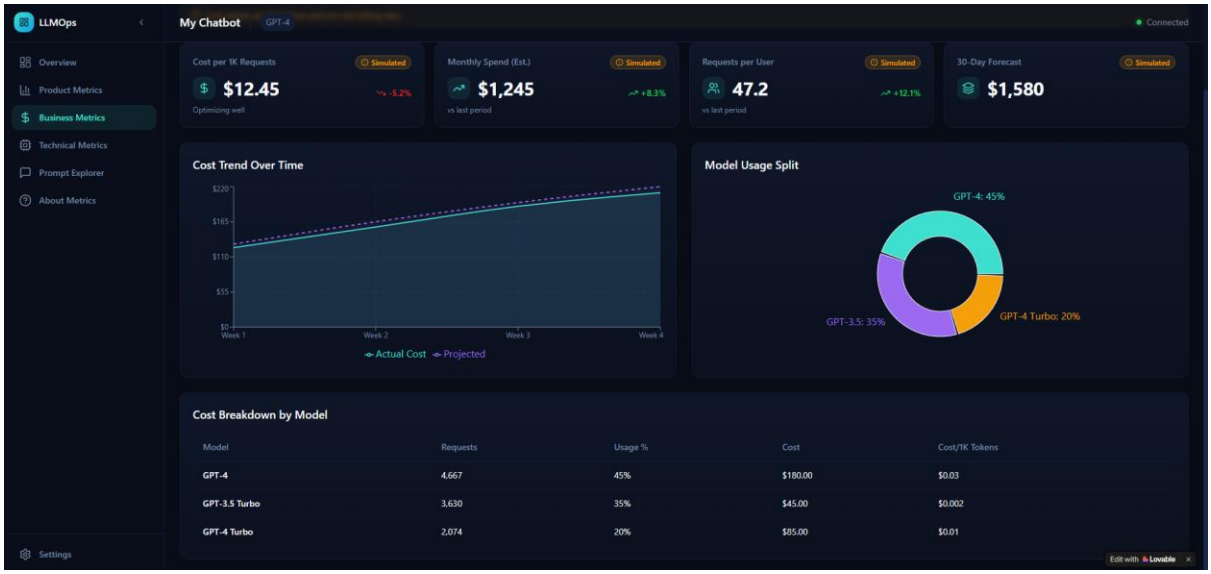
Link: https://llmo-insight-flow.lovable.app

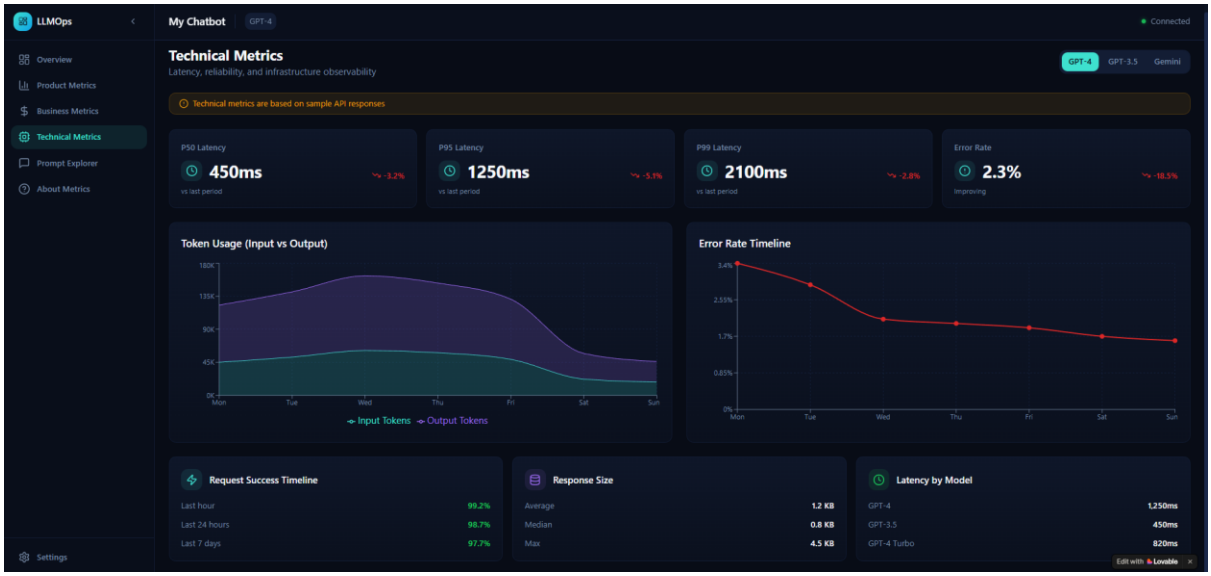


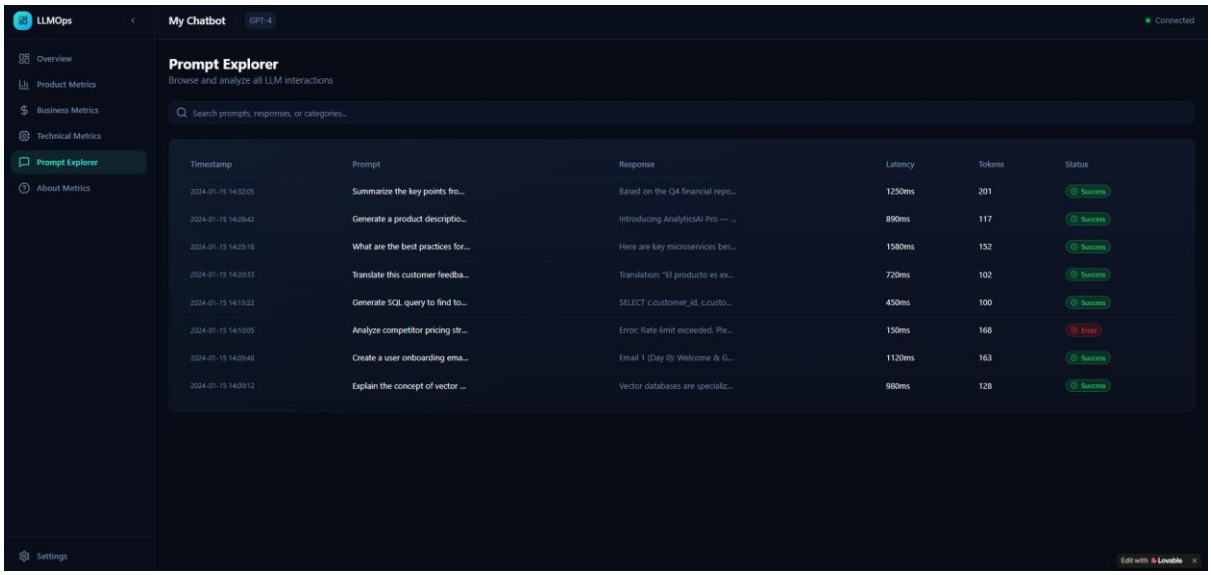

LLMOps
My Chatbot
GPT-4
Connected
Overview
Product Metrics
Business Metrics
Technical Metrics
Prompt Explorer
About Metrics
Settings
Overview
Executive summary of your LLM operations
Metrics shown are simulated for demo purposes
Total LLM Requests
10,370
+12.5%
vs last week
Successful Responses
10,135
+2.3%
97.7% success rate
Avg Response Latency
847ms
-8.2%
vs last week
Est. Monthly Cost
Simulated
$1,245
+5.4%
vs last month
Requests Over Time
2200
1650
1100
550
0
Mon
Tue
Wed
Thu
Fri
Sat
Sun
Successful
Failed
Latency Distribution
36%
27%
18%
9%
0%
<200ms
200-500ms
500-1000ms
1-2s
>2s
Edit with Lovable

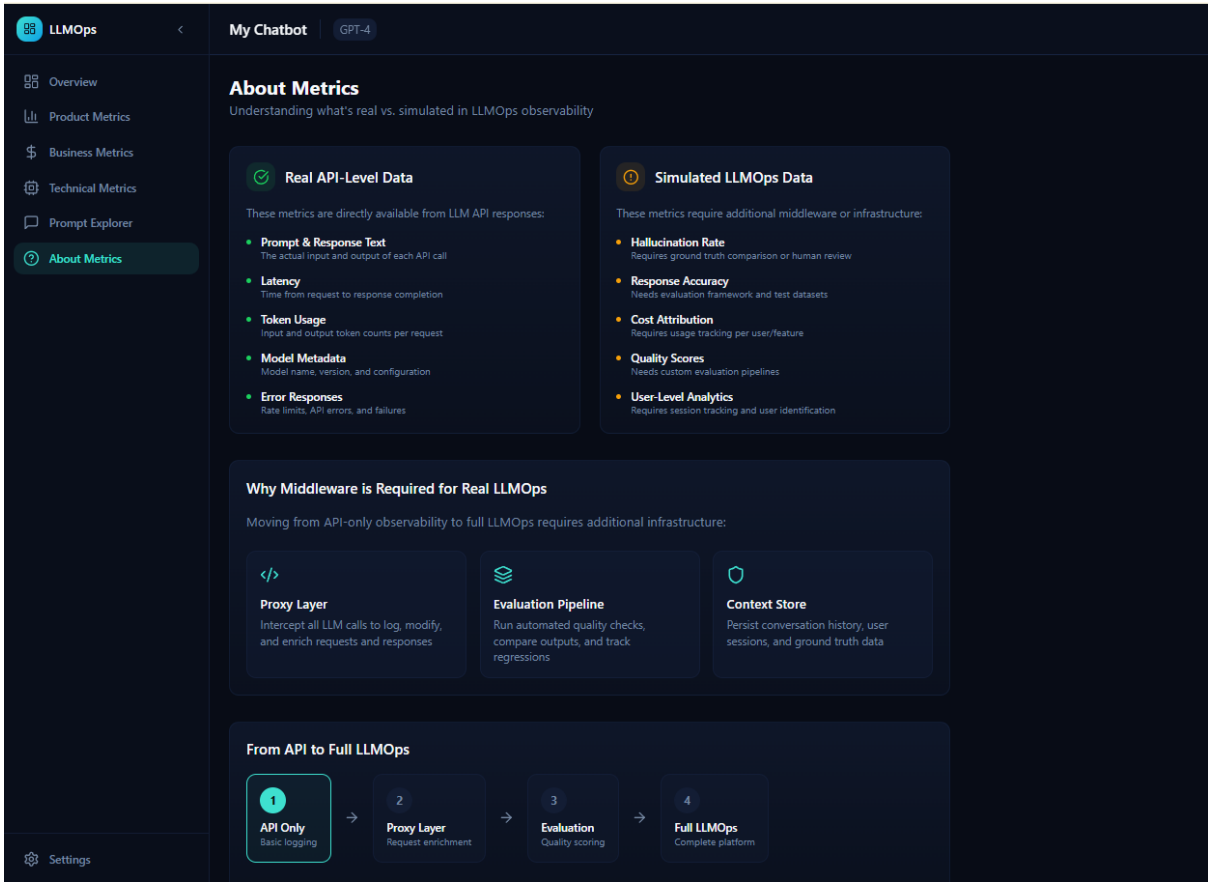

LLMOps
My Chatbot
GPT-4
Connected
Overview
Product Metrics
Business Metrics
Technical Metrics
Prompt Explorer
About Metrics
Settings
Product Metrics
Insights for Product Managers
Analytics shown are simulated for demonstration
Unique Prompts
2,847
+18.3%
vs last period
Regeneration Rate
Simulated
8.2%
-12.5%
Lower is better
Hallucination Risk
Simulated
Low
Intent Match Rate
Simulated
94.2%
+3.1%
vs last period
Prompt Categories
Simulated
Content Generation
Analysis
Code Generation
Translation
Other
0%
9%
18%
27%
36%
Top User Prompts
Summarize the key points from the Q4 financial report and highlight any concerning trends.
Analysis • 1250ms
Success
Generate a product description for our new AI-powered analytics dashboard targeting enterprise customers.
Content Generation • 890ms
Success
What are the best practices for implementing microservices architecture?
Technical • 1580ms
Success
Translate this customer feedback to Spanish and identify sentiment.
Success
Edit with Lovable


LLMOps
My Chatbot
GPT-4
Connected
Overview
Product Metrics
Business Metrics
Technical Metrics
Prompt Explorer
About Metrics
Settings
Business Metrics
Cost, scale, and ROI insights for leadership
Cost values are illustrative and not real billing data
Cost per 1K Requests
Simulated
$12.45
-5.2%
Optimizing well
Monthly Spend (Est.)
Simulated
$1,245
+8.3%
vs last period
Requests per User
Simulated
47.2
+12.1%
vs last period
30-Day Forecast
Simulated
$1,580
Cost Trend Over Time
$220
$165
$110
$55
$0
Week 1
Week 2
Week 3
Week 4
Actual Cost
Projected
Model Usage Split
GPT-4: 45%
GPT-3.5: 35%
GPT-4 Turbo: 20%
Cost Breakdown by Model
Model
Requests
Usage %
Cost
Cost/1K Tokens
Edit with Lovable


LLMOps
My Chatbot
GPT-4
Connected
Overview
Product Metrics
Business Metrics
Technical Metrics
Prompt Explorer
About Metrics
Settings
Cost per 1K Requests
Simulated
$12.45
-5.2%
Optimizing well
Monthly Spend (Est.)
Simulated
$1,245
+8.3%
vs last period
Requests per User
Simulated
47.2
+12.1%
vs last period
30-Day Forecast
Simulated
$1,580
Cost Trend Over Time
$220
$165
$110
$55
$0
Week 1
Week 2
Week 3
Week 4
Actual Cost
Projected
Model Usage Split
GPT-4: 45%
GPT-3.5: 35%
GPT-4 Turbo: 20%
Cost Breakdown by Model
Model
Requests
Usage %
Cost
Cost/1K Tokens
GPT-4
4,667
45%
$180.00
$0.03
GPT-3.5 Turbo
3,630
35%
$45.00
$0.002
GPT-4 Turbo
2,074
20%
$85.00
$0.01
Edit with Lovable


LLMOps
My Chatbot
GPT-4
Connected
Overview
Product Metrics
Business Metrics
Technical Metrics
Prompt Explorer
About Metrics
Settings
Technical Metrics
Latency, reliability, and infrastructure observability
GPT-4
GPT-3.5
Gemini
Technical metrics are based on sample API responses
P50 Latency
450ms
-3.2%
vs last period
P95 Latency
1250ms
-5.1%
vs last period
P99 Latency
2100ms
-2.8%
vs last period
Error Rate
2.3%
-18.5%
Improving
Token Usage (Input vs Output)
180K
135K
90K
45K
0K
Mon
Tue
Wed
Thu
Fri
Sat
Sun
Input Tokens
Output Tokens
Error Rate Timeline
3.4%
2.55%
1.7%
0.85%
0%
Request Success Timeline
Last hour
99.2%
Last 24 hours
98.7%
Last 7 days
97.7%
Response Size
Average
1.2 KB
Median
0.8 KB
Max
4.5 KB
Latency by Model
GPT-4
1,250ms
GPT-3.5
450ms
GPT-4 Turbo
820ms
Edit with Lovable


LLMOps
My Chatbot
GPT-4
Connected
Overview
Product Metrics
Business Metrics
Technical Metrics
Prompt Explorer
About Metrics
Settings
Prompt Explorer
Browse and analyze all LLM interactions
Search prompts, responses, or categories...
Timestamp
Prompt
Response
Latency
Tokens
Status
2024-01-15 14:32:05
Summarize the key points fro...
Based on the Q4 financial repo...
1250ms
201
Success
2024-01-15 14:28:42
Generate a product descriptio...
Introducing AnalyticsAI Pro — ...
890ms
117
Success
2024-01-15 14:25:18
What are the best practices for...
Here are key microservices bes...
1580ms
152
Success
2024-01-15 14:20:33
Translate this customer feedba...
Translation: "El producto es ex...
720ms
102
Success
2024-01-15 14:15:22
Generate SQL query to find to...
SELECT c.customer_id, c.custo...
450ms
100
Success
2024-01-15 14:10:05
Analyze competitor pricing str...
Error: Rate limit exceeded. Ple...
150ms
168
Error
2024-01-15 14:05:48
Create a user onboarding ema...
Email 1 (Day 0): Welcome & G...
1120ms
163
Success
2024-01-15 14:00:12
Explain the concept of vector ...
Vector databases are specializ...
980ms
128
Success
Edit with Lovable


LLMOps
My Chatbot
GPT-4
Overview
Product Metrics
Business Metrics
Technical Metrics
Prompt Explorer
About Metrics
Settings
About Metrics
Understanding what's real vs. simulated in LLMOps observability
Real API-Level Data
These metrics are directly available from LLM API responses:
Prompt & Response Text
The actual input and output of each API call
Latency
Time from request to response completion
Token Usage
Input and output token counts per request
Model Metadata
Model name, version, and configuration
Error Responses
Rate limits, API errors, and failures
Simulated LLMOps Data
These metrics require additional middleware or infrastructure:
Hallucination Rate
Requires ground truth comparison or human review
Response Accuracy
Needs evaluation framework and test datasets
Cost Attribution
Requires usage tracking per user/feature
Quality Scores
Needs custom evaluation pipelines
User-Level Analytics
Requires session tracking and user identification
Why Middleware is Required for Real LLMOps
Moving from API-only observability to full LLMOps requires additional infrastructure:
Proxy Layer
Intercept all LLM calls to log, modify, and enrich requests and responses
Evaluation Pipeline
Run automated quality checks, compare outputs, and track regressions
Context Store
Persist conversation history, user sessions, and ground truth data
From API to Full LLMOps
1
API Only
Basic logging
2
Proxy Layer
Request enrichment
3
Evaluation
Quality scoring
4
Full LLMOps
Complete platform