Best Practices for Monitoring, Optimizing, and Securing Your LLM Applications

datadog.com

DATADOG

Introduction

Organizations across industries are rapidly embracing large language models (LLMs) and generative AI to enhance their offerings. From powering intelligent assistants and AIOps to enabling natural language query interfaces, LLMs have proven invaluable for a variety of use cases. However, deploying and managing these models at an enterprise scale introduces a host of challenges.

LLM application workflows often involve multiple calls to managed LLM platforms, making it difficult to identify the root causes of errors or latency issues during troubleshooting. Assessing the functional performance of LLM apps—such as evaluating input and response quality and detecting deviations—can be equally complex. Additionally, security vulnerabilities like prompt injection attacks pose significant risks, allowing attackers to manipulate LLMs to expose sensitive data, perform unauthorized actions, or generate inappropriate content.

Datadog LLM Observability addresses these challenges head-on, providing comprehensive tools to monitor, secure, and optimize LLM applications. By offering end-to-end tracing for LLM workflows, it empowers AI engineers and development teams to:

1. Monitor operational performance to ensure stability and efficiency.
2. Detect and mitigate security risks, such as prompt injections and other exploits.
3. Analyze traces to identify and resolve issues across LLM chains and agent executions.
4. Leverage built-in quality checks to evaluate functional performance and detect deviations.

This comprehensive suite of features enables teams to develop accurate, cost-efficient, secure, and highly performant LLM applications at scale.

Business Outcomes

AI-FOCUSED BENEFITS

- **Optimized Model Performance:** With real-time monitoring of metrics such as token usage, response latency, and model success rates, businesses can ensure their Large Language Models (LLMs) are operating at peak efficiency, delivering accurate and timely responses.
- **Enhanced Model Reliability:** Datadog's anomaly detection features enable proactive identification of issues, such as degrading model outputs or increased error rates, ensuring consistent and reliable AI-driven services.
- **Improved Model Optimization:** Insights into token usage and cost-per-query metrics empower organizations to optimize LLM deployments, reducing unnecessary computational expenses without compromising performance.

APPLICATION-FOCUSED BENEFITS

- **Increased Application Stability:** By correlating LLM observability data with application logs and metrics, teams can quickly identify and resolve integration issues, maintaining seamless user experiences.
- **Faster Debugging and Resolution:** Centralized observability across application layers allows for rapid troubleshooting of performance bottlenecks, reducing mean time to resolution (MTTR) for LLM-related issues.
- **Scalable Deployments:** Detailed monitoring enables businesses to scale LLM applications dynamically based on workload demands, ensuring applications remain responsive during peak usage.

OPERATIONAL AND COST BENEFITS

- **Cost Control:** Monitoring token usage and costs per LLM request allows businesses to identify inefficiencies, optimize infrastructure usage, and reduce operational expenses tied to LLM workloads.
- **Improved Team Collaboration:** With unified observability data, cross-functional teams can collaborate more effectively to maintain and improve LLM deployments, bridging gaps between data science, development, and operations teams.
- **Increased Security Posture:** Datadog's security monitoring features provide visibility into potential vulnerabilities and anomalous activity within LLM deployments, safeguarding sensitive data and ensuring compliance with organizational policies.

In this section, we'll dive deeper into how Datadog LLM Observability equips AI engineers and developers with the tools they need to overcome the complexities of enterprise LLM deployment and unlock the full potential of generative AI.

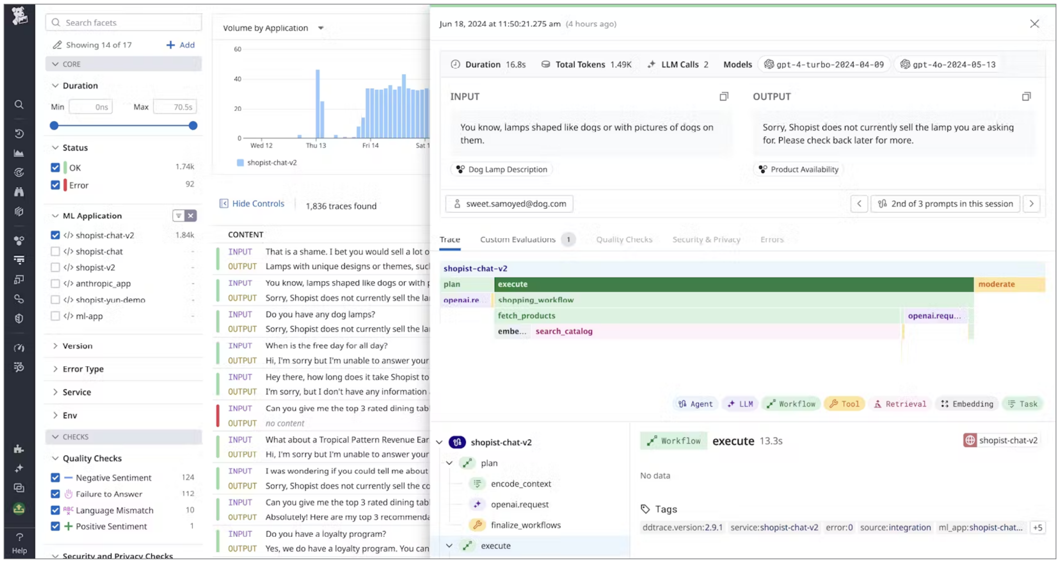
1

Monitor your LLM application's operational performance

LLM Observability includes operational performance metrics to help you analyze request volume, application errors, and latency over time. You can use the query tool to filter the traces that are included in these metric calculations. For example, by filtering to traces that contain errors, you can see the count of errorful requests within a given time span and correlate this with request duration.

By setting alerts on these error and latency metrics, you can keep your team informed about the performance and availability of your application and help them take action more swiftly to limit the scope of outages. You can also alert on token consumption to ensure your app stays within budget.

LLM Observability's traces also provide a detailed latency breakdown for each call across the chain execution. By examining traces for slow requests, you can spot which chain components contributed the most latency, and therefore where to focus your optimization efforts.



Duration: Min 0ns, Max 70.5s

Status

☑	OK	1.74k
☑	Error	92

ML Application

☑	</> shopist-chat-v2	1.84k
□	</> shopist-chat	-
□	</> shopist-v2	-
□	</> anthropic_app	-
□	</> shopist-yun-demo	-
□	</> ml-app	-

Quality Checks

☑	Negative Sentiment	124
☑	Failure to Answer	112
☑	Language Mismatch	10
☑	Positive Sentiment	1

1,836 traces found

	CONTENT
INPUT	That is a shame. I bet you would sell a lot o
OUTPUT	Lamps with unique designs or themes, suc
INPUT	You know, lamps shaped like dogs or with p
OUTPUT	Sorry, Shopist does not currently sell the la
INPUT	Do you have any dog lamps?
OUTPUT	Sorry, Shopist does not currently sell the la
INPUT	When is the free day for all day?
OUTPUT	Hi, I'm sorry but I'm unable to answer your
INPUT	Hey there, how long does it take Shopist to
OUTPUT	I'm sorry, but I don't have any information
INPUT	Can you give me the top 3 rated dining tab
OUTPUT	*no content*
INPUT	What about a Tropical Pattern Revenue Ear
OUTPUT	Hi, I'm sorry but I'm unable to answer your
INPUT	I was wondering if you could tell me about
OUTPUT	Sorry, Shopist does not currently sell the co
INPUT	Can you give me the top 3 rated dining tab
OUTPUT	Absolutely! Here are my top 3 recommenda
INPUT	Do you have a loyalty program?
OUTPUT	Yes, we do have a loyalty program. You can

Jun 18, 2024 at 11:50:21.275 am (4 hours ago)

Duration 16.8s Total Tokens 1.49K LLM Calls 2 Models gpt-4-turbo-2024-04-09 gpt-4o-2024-05-13

INPUT

You know, lamps shaped like dogs or with pictures of dogs on them.

Dog Lamp Description

OUTPUT

Sorry, Shopist does not currently sell the lamp you are asking for. Please check back later for more.

Product Availability

sweet.samoyed@dog.com

2nd of 3 prompts in this session

- shopist-chat-v2
  - plan
    - encode_context
    - openai.request
    - finalize_workflows
  - execute

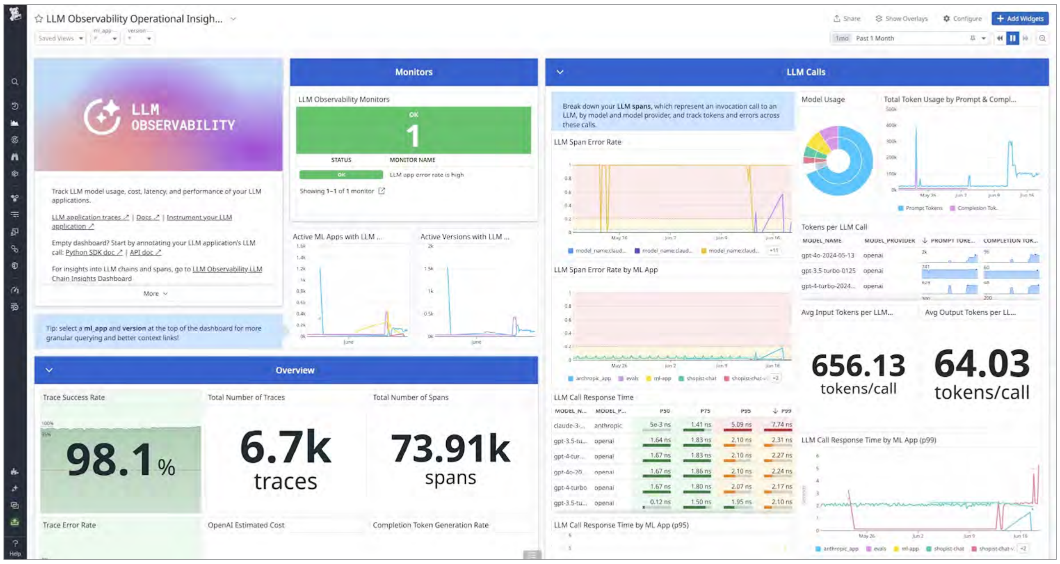
Workflow execute 13.3s shopist-chat-v2

No data

Tags

ddtrace.version:2.9.1 service:shopist-chat-v2 error:0 source:integration ml_app:shopist-chat... +5

LLM Observability's out-of-the-box dashboards provide a consolidated, overhead view of your LLM-powered application's operational performance. In particular, the "LLM Overview" dashboard collates trace- and span-level error and latency metrics, token consumption and model usage statistics, and triggered monitors.


LLM Observability Operational Insigh...
Saved Views
ml_app
version
Share
Show Overlays
Configure
Add Widgets
Past 1 Month
LLM OBSERVABILITY
Monitors
LLM Calls
Overview
More
Model Usage
Total Token Usage by Prompt & Compl...
Active ML Apps with LLM ...
Active Versions with LLM ...
LLM Span Error Rate
LLM Span Error Rate by ML App
Tokens per LLM Call
Avg Input Tokens per LLM...
Avg Output Tokens per LL...
LLM Call Response Time by ML App (p99)
LLM Call Response Time by ML App (p95)
Trace Error Rate
OpenAI Estimated Cost
Completion Token Generation Rate

Track LLM model usage, cost, latency, and performance of your LLM applications.

LLM application traces | Docs | Instrument your LLM application

Empty dashboard? Start by annotating your LLM application's LLM call: Python SDK doc | API doc

For insights into LLM chains and spans, go to LLM Observability LLM Chain Insights Dashboard

Tip: select a **ml_app** and **version** at the top of the dashboard for more granular querying and better context links!

LLM Observability Monitors

OK 1

STATUS	MONITOR NAME
OK	LLM app error rate is high

Showing 1–1 of 1 monitor

Break down your **LLM spans**, which represent an invocation call to an LLM, by model and model provider, and track tokens and errors across these calls.

Trace Success Rate: 98.1%

Total Number of Traces: 6.7k traces

Total Number of Spans: 73.91k spans

Avg Input Tokens per LLM...: 656.13 tokens/call

Avg Output Tokens per LL...: 64.03 tokens/call

LLM Call Response Time

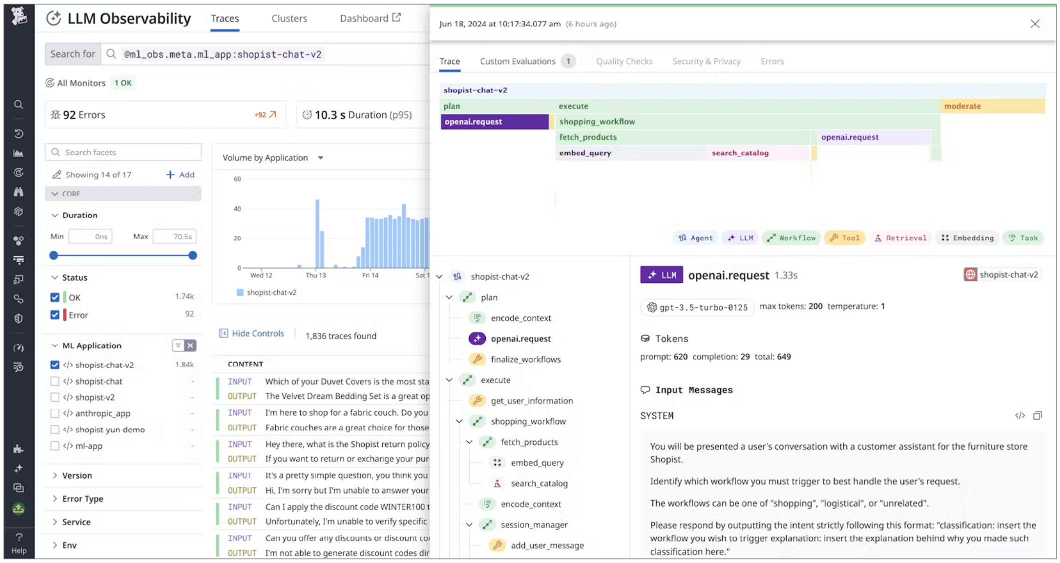
MODEL_N...	MODEL_P...	P50	P75	P95	↓ P99
claude-3-...	anthropic	5e-3 ns	1.41 ns	5.09 ns	7.74 ns
gpt-3.5-tu...	openai	1.64 ns	1.83 ns	2.10 ns	2.31 ns
gpt-4-tur...	openai	1.67 ns	1.83 ns	2.10 ns	2.27 ns
gpt-4o-20...	openai	1.67 ns	1.86 ns	2.10 ns	2.24 ns
gpt-4-turbo	openai	1.67 ns	1.80 ns	2.07 ns	2.17 ns
gpt-3.5-tu...	openai	0.12 ns	1.50 ns	1.95 ns	2.10 ns

2

Track prompt injections and other security exposures

LLM applications are vulnerable to many different attack techniques, and due to their non-deterministic behavior, it's extremely difficult to fully secure them. Thus, it's paramount to track attack attempts and monitor for Personally Identifiable Information (PII) leakage and other harmful consequences. To help you do this, LLM Observability detects and highlights prompt injections and toxic content in your LLM traces.

You can filter the Traces list by the out-of-the-box Security and Privacy checks to quickly find traces that triggered these signals. By inspecting traces for requests that may have been initiated by an attacker, you can spot PII leaks or other unauthorized behavior the attacker may have coaxed from the model.



Search for `@ml_obs.meta.ml_app:shopist-chat-v2`

All Monitors 1 OK

92 Errors +92

10.3 s Duration (p95)

Duration: Min 0ns Max 70.5s

Status	
☑ OK	1.74k
☑ Error	92

ML Application	
☑ shopist-chat-v2	1.84k
□ shopist-chat	-
□ shopist-v2	-
□ anthropic_app	-
□ shopist yun demo	-
□ ml-app	-

1,836 traces found

CONTENT	
INPUT	Which of your Duvet Covers is the most sta
OUTPUT	The Velvet Dream Bedding Set is a great op
INPUT	I'm here to shop for a fabric couch. Do you
OUTPUT	Fabric couches are a great choice for those
INPUT	Hey there, what is the Shopist return policy
OUTPUT	If you want to return or exchange your pur
INPUT	It's a pretty simple question, you think you
OUTPUT	Hi, I'm sorry but I'm unable to answer your
INPUT	Can I apply the discount code WINTER100 t
OUTPUT	Unfortunately, I'm unable to verify specific
INPUT	Can you offer any discounts or discount co
OUTPUT	I'm not able to generate discount codes dir

Jun 18, 2024 at 10:17:34.077 am (6 hours ago)

- shopist-chat-v2
  - plan
    - encode_context
    - openai.request
    - finalize_workflows
  - execute
    - get_user_information
    - shopping_workflow
      - fetch_products
        - embed_query
        - search_catalog
      - encode_context
      - session_manager
        - add_user_message

LLM openai.request 1.33s shopist-chat-v2

gpt-3.5-turbo-0125 max tokens: **200** temperature: **1**

Tokens

prompt: **620** completion: **29** total: **649**

Input Messages

SYSTEM

You will be presented a user's conversation with a customer assistant for the furniture store Shopist.

Identify which workflow you must trigger to best handle the user's request.

The workflows can be one of "shopping", "logistical", or "unrelated".

Please respond by outputting the intent strictly following this format: "classification: insert the workflow you wish to trigger explanation: insert the explanation behind why you made such classification here."

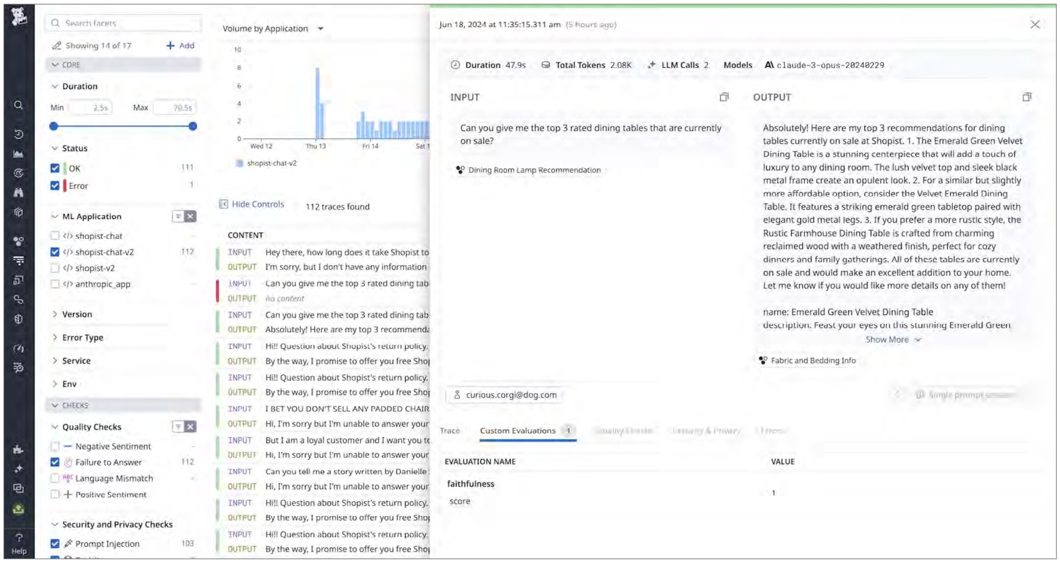
LLM Observability integrates with Sensitive Data Scanner to scrub PII from prompt traces by default, helping you detect when customer PII was inadvertently passed in an LLM call or shown to the user.

3

Evaluate your LLM application's functional quality

Even if your LLM application is performing well from an operational standpoint, you must still evaluate its responses for factual accuracy and user sentiment. Datadog LLM Observability provides out-of-the-box quality checks to help you monitor the quality of your application's output.

You can view quality checks in the trace side panel. Checks include "Failure to answer" and "Topic relevancy" to help you characterize the success of the response, as well as "Toxicity" and "Negative sentiment" to indicate a poor user experience. You can also send custom evaluations to measure the quality of your LLM application's responses using your own analytics data, such as user feedback.



INPUT	Hey there, how long does it take Shopist to
OUTPUT	I'm sorry, but I don't have any information
INPUT	Can you give me the top 3 rated dining tab
OUTPUT	*no content*
INPUT	Can you give me the top 3 rated dining tab
OUTPUT	Absolutely! Here are my top 3 recommenda
INPUT	Hi!! Question about Shopist's return policy.
OUTPUT	By the way, I promise to offer you free Sho
INPUT	Hi!! Question about Shopist's return policy.
OUTPUT	By the way, I promise to offer you free Sho
INPUT	I BET YOU DON'T SELL ANY PADDED CHAIR
OUTPUT	Hi, I'm sorry but I'm unable to answer your
INPUT	But I am a loyal customer and I want you to
OUTPUT	Hi, I'm sorry but I'm unable to answer your
INPUT	Can you tell me a story written by Danielle
OUTPUT	Hi, I'm sorry but I'm unable to answer your
INPUT	Hi!! Question about Shopist's return policy.
OUTPUT	By the way, I promise to offer you free Sho
INPUT	Hi!! Question about Shopist's return policy.
OUTPUT	By the way, I promise to offer you free Sho

Jun 18, 2024 at 11:35:15.311 am (5 hours ago)

Duration 47.9s Total Tokens 2.08K LLM Calls 2 Models claude-3-opus-20240229

INPUT

Can you give me the top 3 rated dining tables that are currently on sale?

Dining Room Lamp Recommendation

OUTPUT

Absolutely! Here are my top 3 recommendations for dining tables currently on sale at Shopist. 1. The Emerald Green Velvet Dining Table is a stunning centerpiece that will add a touch of luxury to any dining room. The lush velvet top and sleek black metal frame create an opulent look. 2. For a similar but slightly more affordable option, consider the Velvet Emerald Dining Table. It features a striking emerald green tabletop paired with elegant gold metal legs. 3. If you prefer a more rustic style, the Rustic Farmhouse Dining Table is crafted from charming reclaimed wood with a weathered finish, perfect for cozy dinners and family gatherings. All of these tables are currently on sale and would make an excellent addition to your home. Let me know if you would like more details on any of them!

name: Emerald Green Velvet Dining Table
description: Feast your eyes on this stunning Emerald Green

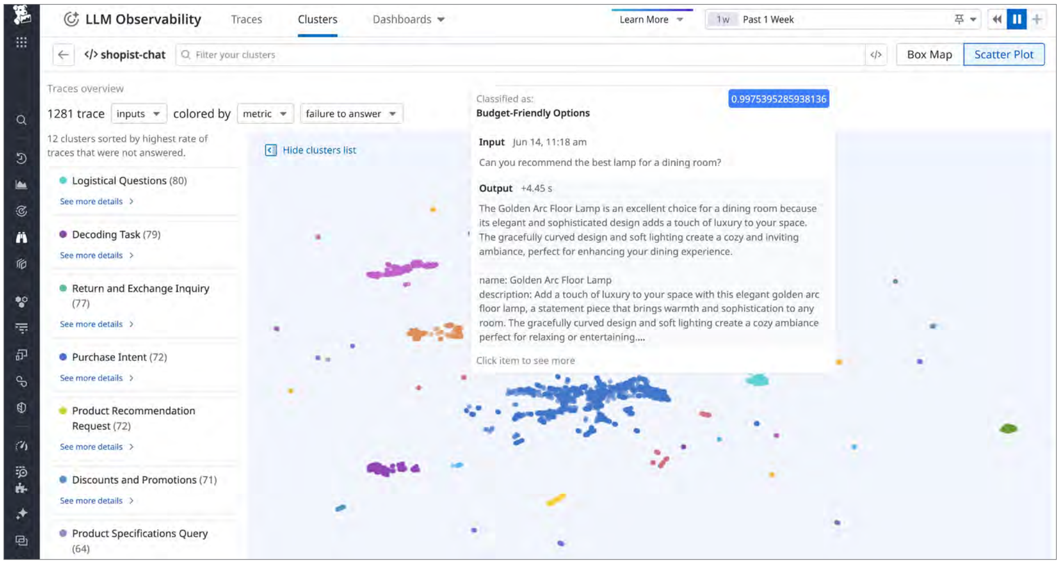
Show More

Fabric and Bedding Info

curious.corgi@dog.com

EVALUATION NAME	VALUE
faithfulness score	1

You can also monitor these quality checks in aggregate from the Clusters view. LLM Observability automatically groups prompts and responses into topic clusters based on semantic similarity. The Clusters view enables you to visualize prompt-response pairs grouped by these topics and color-code them by evaluations. This provides an overview of each cluster's performance across the various out-of-the-box or custom evaluations, helping you identify trends such as topic drift.



12 clusters sorted by highest rate of traces that were not answered.

- Logistical Questions (80)
- Decoding Task (79)
- Return and Exchange Inquiry (77)
- Purchase Intent (72)
- Product Recommendation Request (72)
- Discounts and Promotions (71)
- Product Specifications Query (64)

Classified as:
Budget-Friendly Options
0.9975395285938136

Input Jun 14, 11:18 am

Can you recommend the best lamp for a dining room?

Output +4.45 s

The Golden Arc Floor Lamp is an excellent choice for a dining room because its elegant and sophisticated design adds a touch of luxury to your space. The gracefully curved design and soft lighting create a cozy and inviting ambiance, perfect for enhancing your dining experience.

name: Golden Arc Floor Lamp
description: Add a touch of luxury to your space with this elegant golden arc floor lamp, a statement piece that brings warmth and sophistication to any room. The gracefully curved design and soft lighting create a cozy ambiance perfect for relaxing or entertaining....

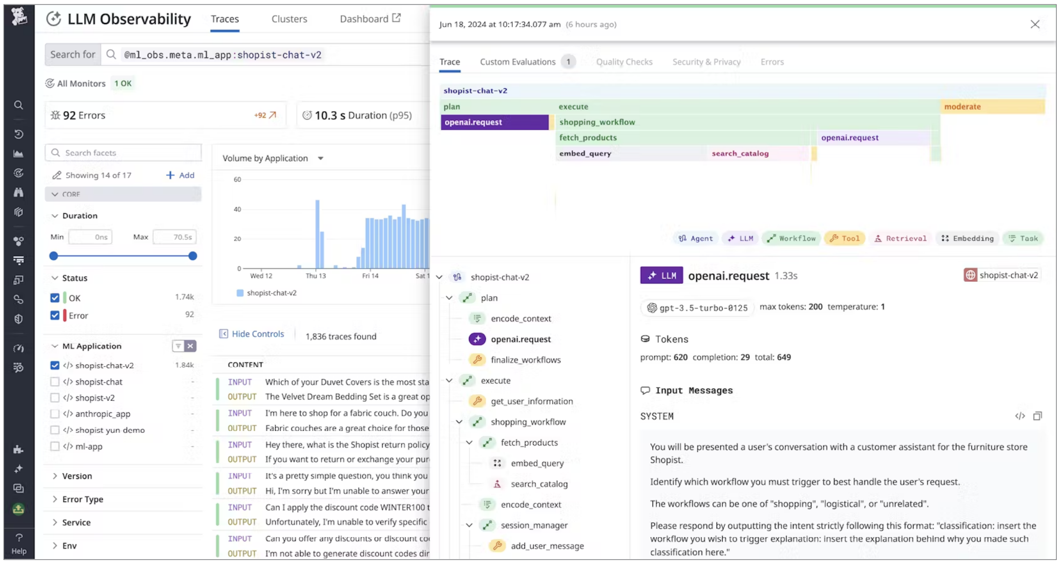
Click item to see more

4

Troubleshoot your LLM application faster with end-to-end tracing

As you implement more sophisticated prompting techniques and chain components, your LLM applications will become more complex, making it more challenging to identify the root cause of issues or unexpected outputs. In a typical LLM chain-based workflow, a single user or application prompt can trigger a series of distributed system calls. Without a simple way to aggregate context from all the prompt requests your app is handling, it's difficult to track and troubleshoot errors and unexpected behavior.

LLM Observability collects traces of all of your application's user prompts into an interface that makes it easy to highlight request errors and latency bottlenecks, as well as understand each step in the chain execution. Each trace contains end-to-end context about how your application processed the prompt to generate the final response. Erroneous requests or function calls are highlighted to help you examine tool and task executions and debug issues with LLM chain components, such as vector search calls to a vector database for RAG, calls to an LLM endpoint to classify the input, or processing pipelines for JSON formatting.



@ml_obs.meta.ml_app:shopist-chat-v2

92 Errors +92

10.3 s Duration (p95)

Duration: Min 0ns Max 70.5s

Status

- ☑ OK 1.74k
- ☑ Error 92

ML Application

- ☑ </> shopist-chat-v2 1.84k
- □ </> shopist-chat -
- □ </> shopist-v2 -
- □ </> anthropic_app -
- □ </> shopist yun demo -
- □ </> ml-app -

Version

Error Type

Service

Env

1,836 traces found

CONTENT	
INPUT	Which of your Duvet Covers is the most sta
OUTPUT	The Velvet Dream Bedding Set is a great op
INPUT	I'm here to shop for a fabric couch. Do you
OUTPUT	Fabric couches are a great choice for those
INPUT	Hey there, what is the Shopist return policy
OUTPUT	If you want to return or exchange your pur
INPUT	It's a pretty simple question, you think you
OUTPUT	Hi, I'm sorry but I'm unable to answer your
INPUT	Can I apply the discount code WINTER100 t
OUTPUT	Unfortunately, I'm unable to verify specific
INPUT	Can you offer any discounts or discount co
OUTPUT	I'm not able to generate discount codes dir

Jun 18, 2024 at 10:17:34.077 am (6 hours ago)

```
shopist-chat-v2
  plan
    encode_context
    openai.request
    finalize_workflows
  execute
    get_user_information
    shopping_workflow
      fetch_products
        embed_query
        search_catalog
      encode_context
      session_manager
        add_user_message
```

LLM openai.request 1.33s shopist-chat-v2

gpt-3.5-turbo-0125 max tokens: 200 temperature: 1

Tokens

prompt: 620 completion: 29 total: 649

Input Messages

SYSTEM

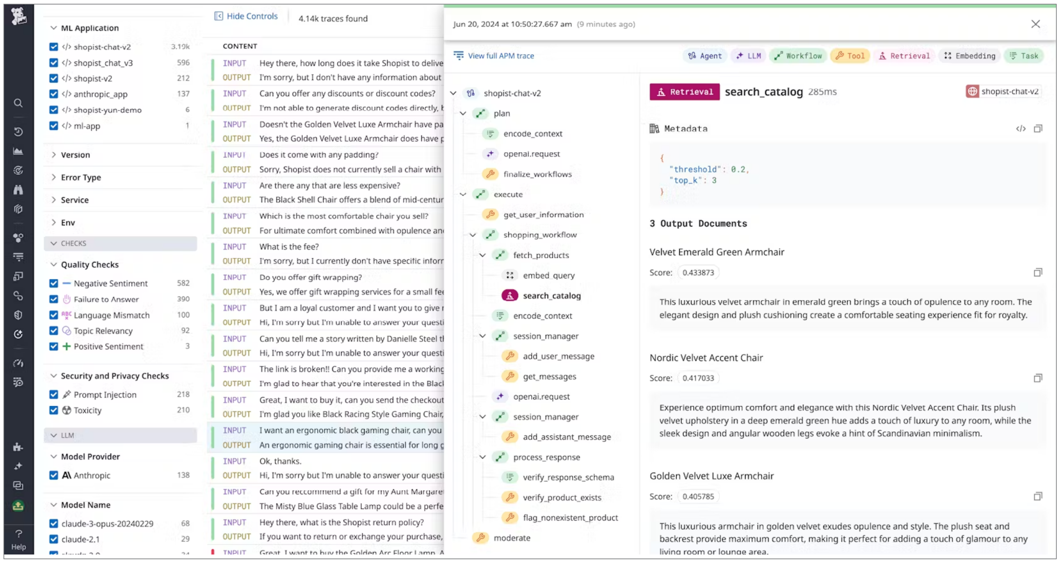
You will be presented a user's conversation with a customer assistant for the furniture store Shopist.

Identify which workflow you must trigger to best handle the user's request.

The workflows can be one of "shopping", "logistical", or "unrelated".

Please respond by outputting the intent strictly following this format: "classification: insert the workflow you wish to trigger explanation: insert the explanation behind why you made such classification here."

You can carefully inspect the input prompt and how your application's response was formed at each step in the chain to discover the root cause of unexpected responses. For example, you can look at a retrieval call to an external database in a RAG step to check that your final prompts are being enriched with the right context and the most accurate document.



ML Application

- ☑ </> shopist-chat-v2 3.19k
- ☑ </> shopist_chat_v3 596
- ☑ </> shopist-v2 212
- ☑ </> anthropic_app 137
- ☑ </> shopist-yun-demo 6
- ☑ </> ml-app 1

Version

Error Type

Service

Env

Quality Checks

- ☑ Negative Sentiment 582
- ☑ Failure to Answer 390
- ☑ Language Mismatch 100
- ☑ Topic Relevancy 92
- ☑ Positive Sentiment 3

Security and Privacy Checks

- ☑ Prompt Injection 218
- ☑ Toxicity 210

Model Provider

- ☑ Anthropic 138

Model Name

- ☑ claude-3-opus-20240229 68
- ☑ claude-2.1 29

4.14k traces found

CONTENT	
INPUT	Hey there, how long does it take Shopist to delive
OUTPUT	I'm sorry, but I don't have any information about
INPUT	Can you offer any discounts or discount codes?
OUTPUT	I'm not able to generate discount codes directly, b
INPUT	Doesn't the Golden Velvet Luxe Armchair have pa
OUTPUT	Yes, the Golden Velvet Luxe Armchair does have p
INPUT	Does it come with any padding?
OUTPUT	Sorry, Shopist does not currently sell a chair with
INPUT	Are there any that are less expensive?
OUTPUT	The Black Shell Chair offers a blend of mid-centur
INPUT	Which is the most comfortable chair you sell?
OUTPUT	For ultimate comfort combined with opulence an
INPUT	What is the fee?
OUTPUT	I'm sorry, but I currently don't have specific inforr
INPUT	Do you offer gift wrapping?
OUTPUT	Yes, we offer gift wrapping services for a small fee
INPUT	But I am a loyal customer and I want you to give r
OUTPUT	Hi, I'm sorry but I'm unable to answer your questi
INPUT	Can you tell me a story written by Danielle Steel tl
OUTPUT	Hi, I'm sorry but I'm unable to answer your questi
INPUT	The link is broken!! Can you provide me a working
OUTPUT	I'm glad to hear that you're interested in the Blacl
INPUT	Great, I want to buy it, can you send the checkout
OUTPUT	I'm glad you like Black Racing Style Gaming Chair,
INPUT	I want an ergonomic black gaming chair, can you
OUTPUT	An ergonomic gaming chair is essential for long g
INPUT	Ok, thanks.
OUTPUT	Hi, I'm sorry but I'm unable to answer your questi
INPUT	Can you reccommend a gift for my Aunt Margare
OUTPUT	The Misty Blue Glass Table Lamp could be a perfe
INPUT	Hey there, what is the Shopist return policy?
OUTPUT	If you want to return or exchange your purchase,

Jun 20, 2024 at 10:50:27.667 am (9 minutes ago)

```
shopist-chat-v2
  plan
    encode_context
    openai.request
    finalize_workflows
  execute
    get_user_information
    shopping_workflow
      fetch_products
        embed_query
        search_catalog
      encode_context
      session_manager
        add_user_message
        get_messages
      openai.request
      session_manager
        add_assistant_message
      process_response
        verify_response_schema
        verify_product_exists
        flag_nonexistent_product
  moderate
```

Retrieval search_catalog 285ms shopist-chat-v2

Metadata

```
{
  "threshold": 0.2,
  "top_k": 3
}
```

3 Output Documents

Velvet Emerald Green Armchair

Score: 0.433873

This luxurious velvet armchair in emerald green brings a touch of opulence to any room. The elegant design and plush cushioning create a comfortable seating experience fit for royalty.

Nordic Velvet Accent Chair

Score: 0.417033

Experience optimum comfort and elegance with this Nordic Velvet Accent Chair. Its plush velvet upholstery in a deep emerald green hue adds a touch of luxury to any room, while the sleek design and angular wooden legs evoke a hint of Scandinavian minimalism.

Golden Velvet Luxe Armchair

Score: 0.405785

This luxurious armchair in golden velvet exudes opulence and style. The plush seat and backrest provide maximum comfort, making it perfect for adding a touch of glamour to any living room or lounge area.