Platmosphere

Milan | 20 May 2025

PLATFORM TECH

From AI Agents to LLMs as Judge

Reshaping Observability in the Era of Generative AI

Diana Todea

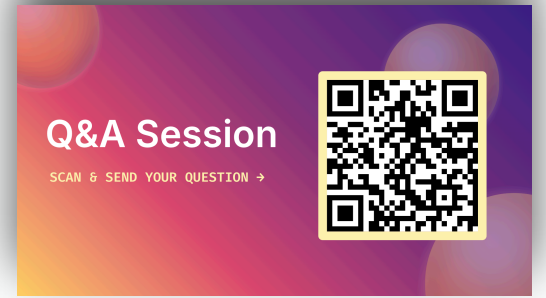
Technical Advocate

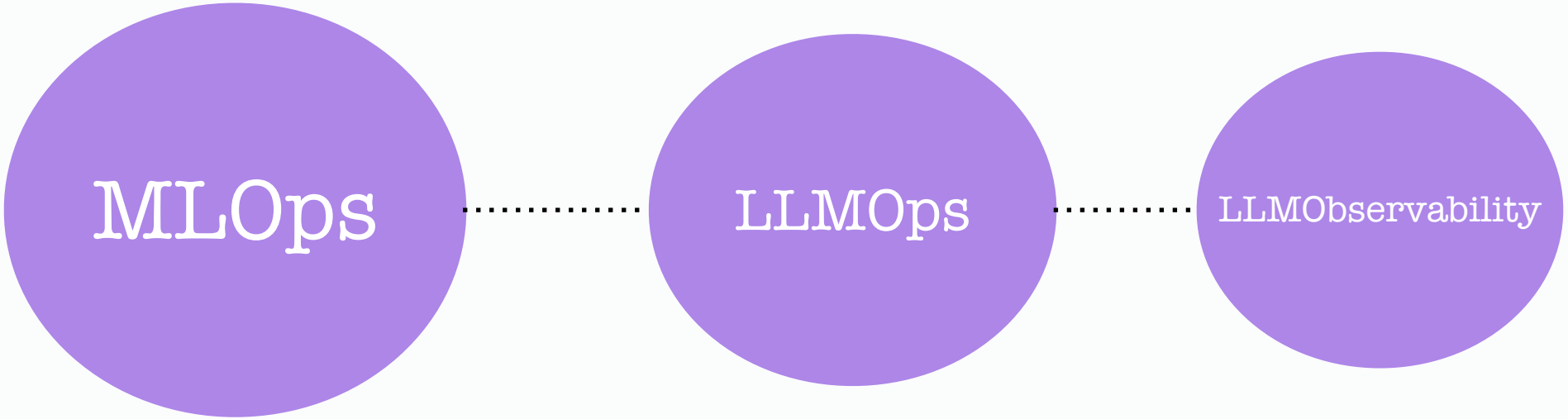
@ Aircall

Diana Todea

Developer Advocate, Observability
SME, OTEL contributor | Keynote & In…

- Concepts
- GenerativeAI architecture
- AI Assistants vs. AI Agents
- Olly & LLMs as Judge
- What's next?




MLOps
LLMOps
LLMObservability

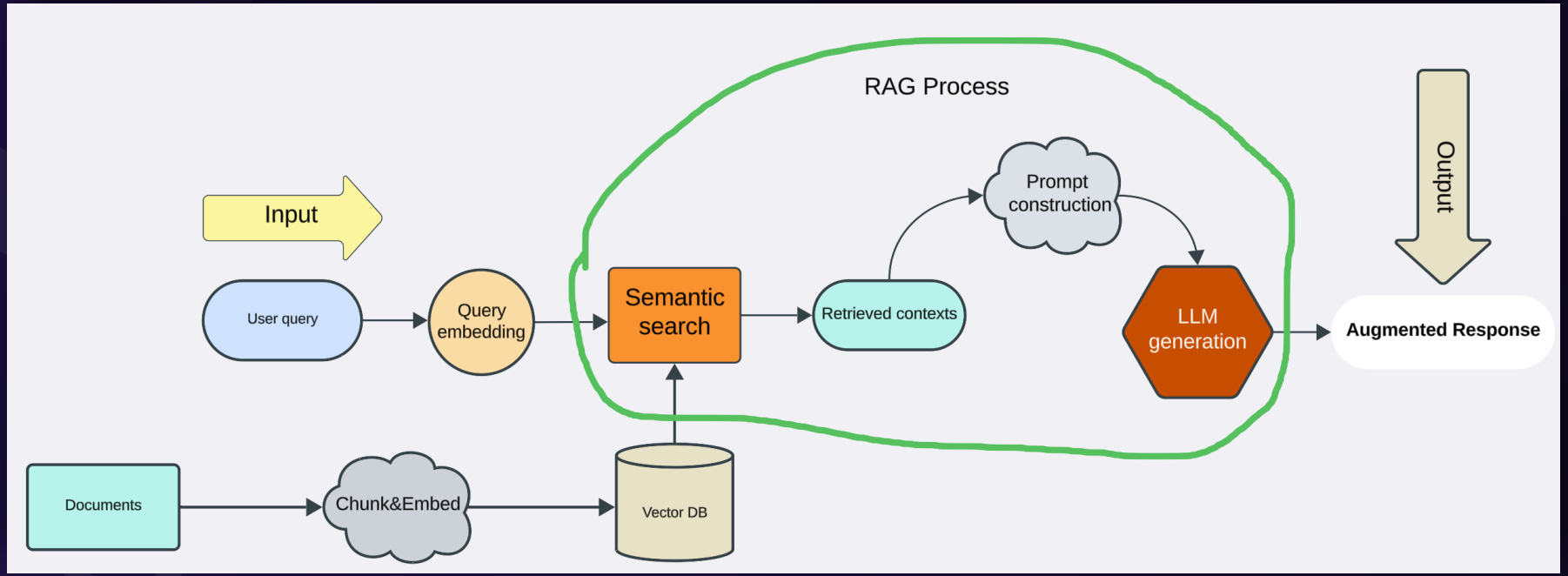
Attention mechanisms allow a neural network to selectively weigh the importance of different input features so the model can focus on the most relevant parts of the input for a given task.

Vector embedding representations are suitable for common machine learning tasks such as clustering, recommendation and classification.

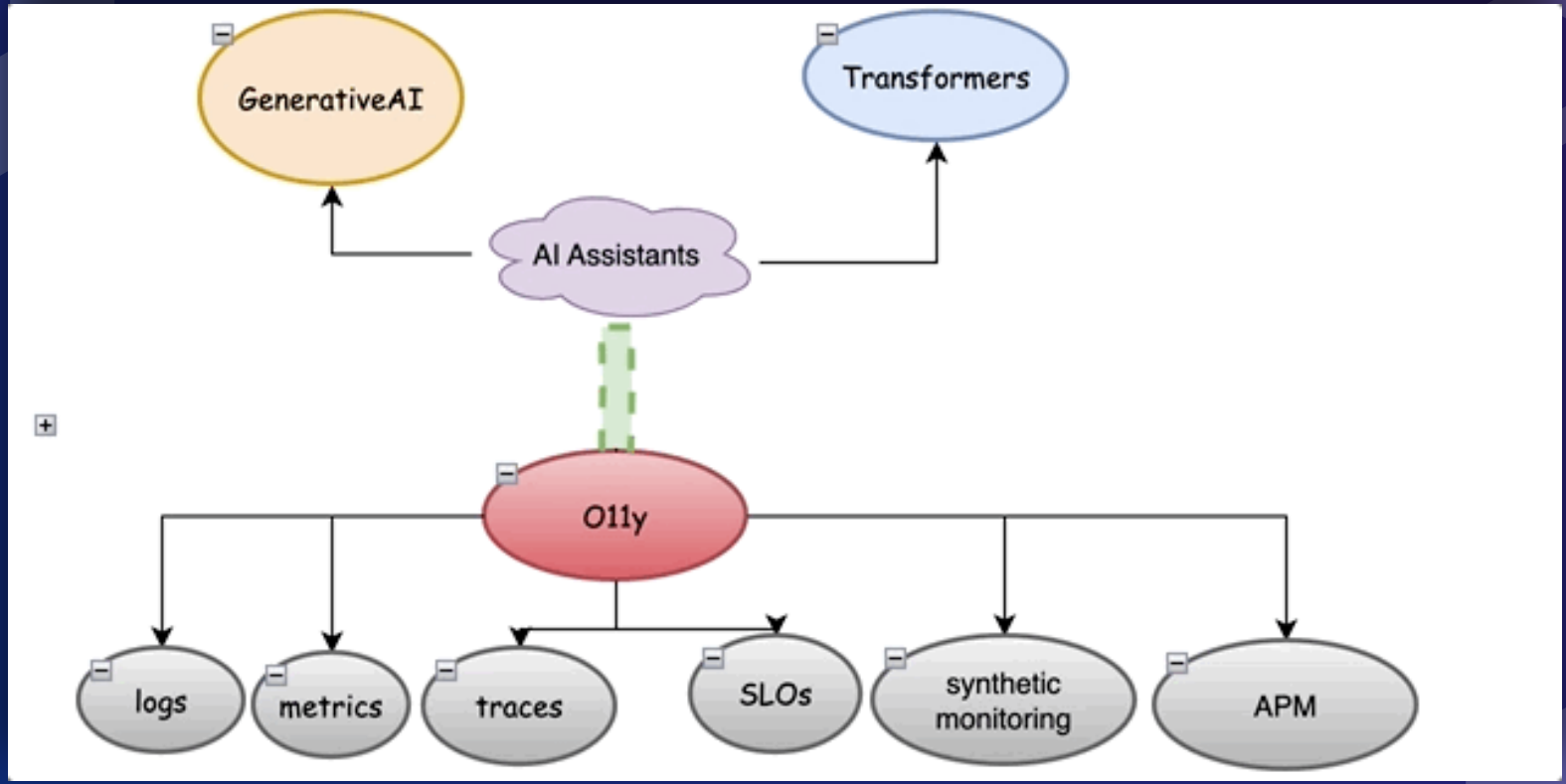
Vector database is used for storing, updating, retrieving. We refer to retrieving set of vectors that are most similar to a query in a form of a vector that is embedded in the same Latent space.

In machine learning, a transformer is a neural network that learns context and meaning by tracking relationships in sequential data like the words in the sentence.

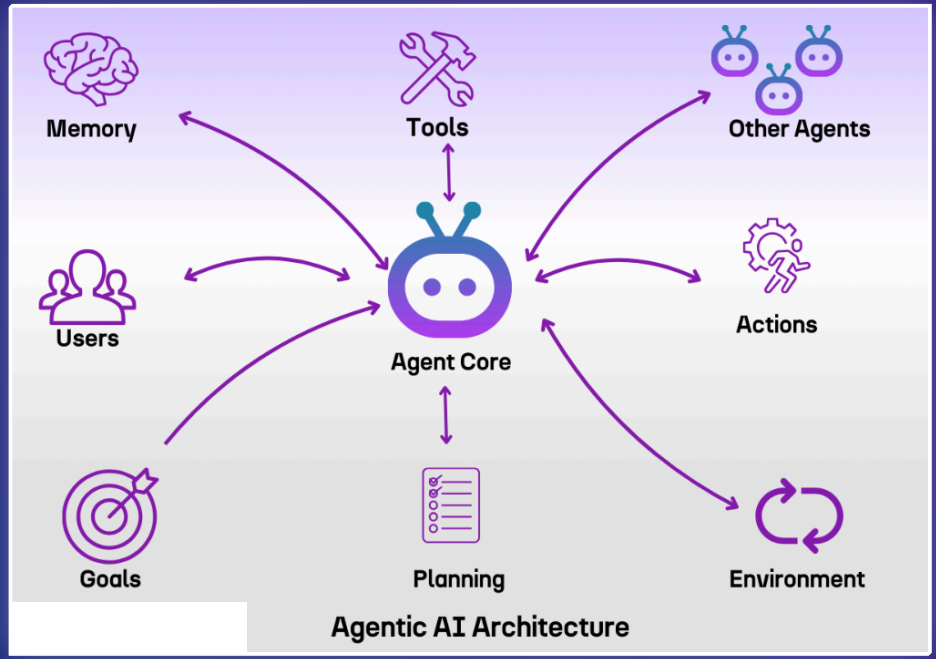
The retrieval procedure is called ANN (approximate nearest neighbor) search.

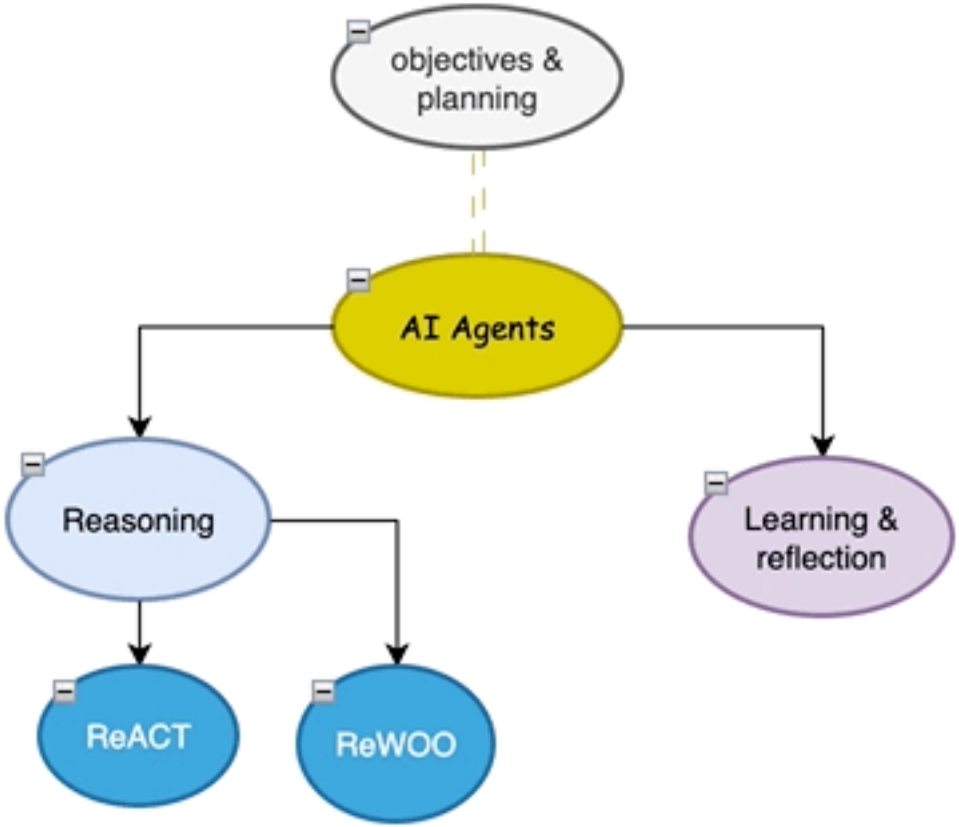


generativeAI architecture + RAG


GenerativeAI
Transformers
AI Assistants
O11y
logs
metrics
traces
SLOs
synthetic monitoring
APM

AI Assistants vs. AI Agents




objectives & planning
AI Agents
Reasoning
Learning & reflection
ReACT
ReWOO

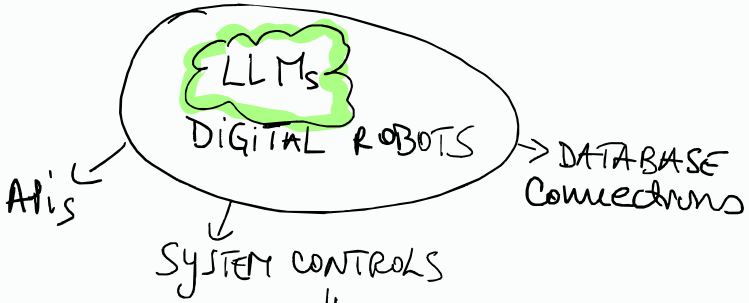
SPAR → LEVEL 5 Fully autonomous agents

SENSE
PLAN
ACT
REFLECT

higher capabilities

TOOLS

- tools identity
- input parameters
- output specifications
- operational constraints
- error handling



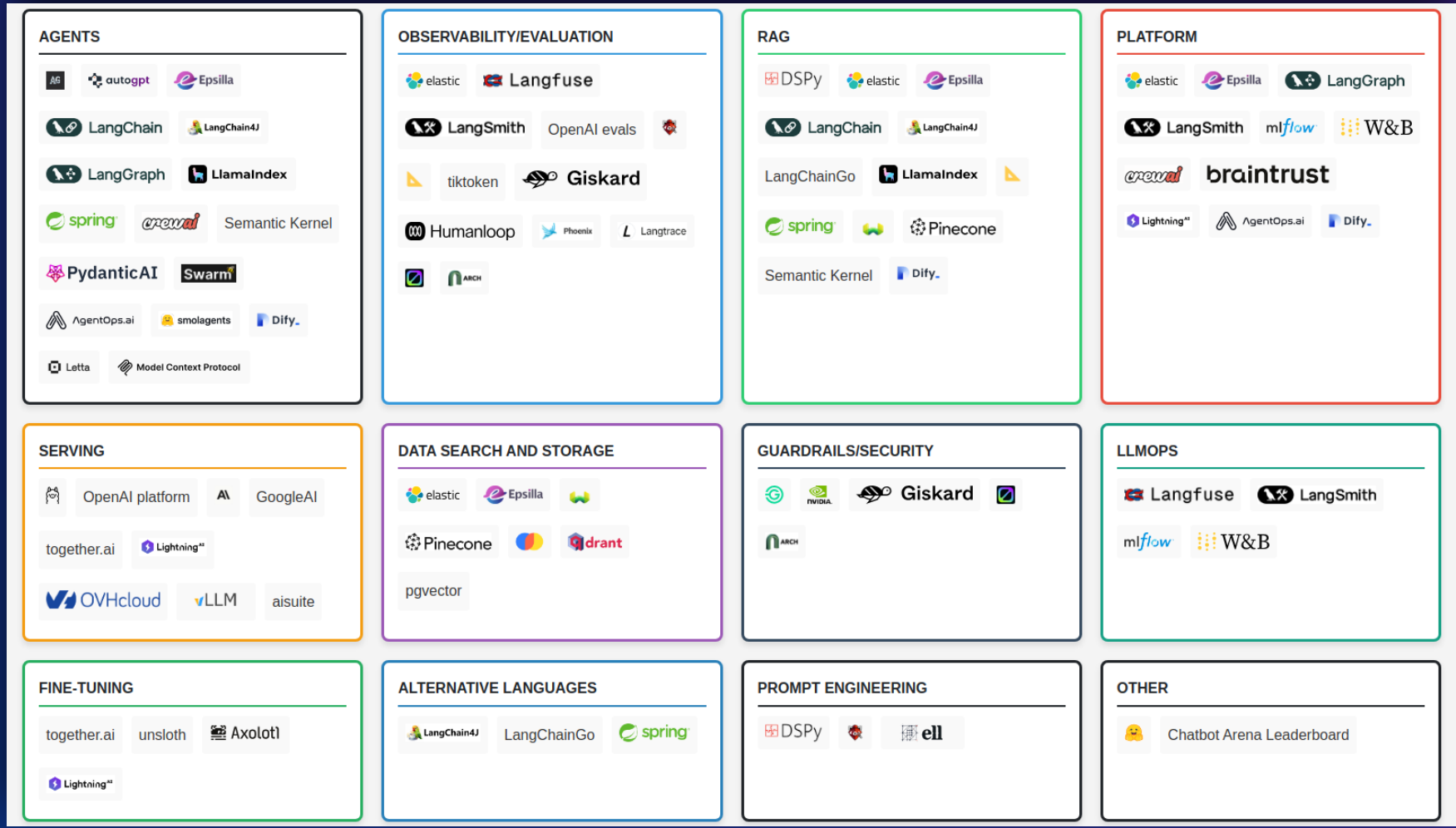
AI AGENTS => enhance decision-making in complex environments
=> improved error detection
=> adaptability

ISSUES

Reasoning errors multiply through a network effect

improve !

- ✓ reasoning checkpoints
- ✓ human validation
- ✓ monitor for signs of error amplification across the system
- ✓ regularly ask agents to explain their dependencies on other agents' outputs


AGENTS
autogpt
Epsilla
LangChain
LangChain4J
LangGraph
LlamaIndex
spring
crewai
Semantic Kernel
PydanticAI
Swarm
AgentOps.ai
smolagents
Dify.
Letta
Model Context Protocol
OBSERVABILITY/EVALUATION
elastic
Langfuse
LangSmith
OpenAI evals
tiktoken
Giskard
Humanloop
Phoenix
Langtrace
ARCH
RAG
DSPy
elastic
Epsilla
LangChain
LangChain4J
LangChainGo
LlamaIndex
spring
Pinecone
Semantic Kernel
Dify.
PLATFORM
elastic
Epsilla
LangGraph
LangSmith
mlflow
W&B
crewai
braintrust
Lightning
AgentOps.ai
Dify.
SERVING
OpenAI platform
GoogleAI
together.ai
Lightning
OVHcloud
vLLM
aisuite
DATA SEARCH AND STORAGE
elastic
Epsilla
Pinecone
drant
pgvector
GUARDRAILS/SECURITY
NVIDIA
Giskard
ARCH
LLMOPS
Langfuse
LangSmith
mlflow
W&B
FINE-TUNING
together.ai
unsloth
Axolotl
Lightning
ALTERNATIVE LANGUAGES
LangChain4J
LangChainGo
spring
PROMPT ENGINEERING
DSPy
ell
OTHER
Chatbot Arena Leaderboard

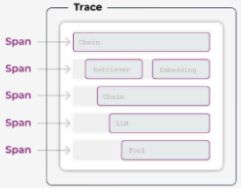
observability and its 5 pillars

Evaluation

Evaluations of LLM outputs by using a separate evaluation LLM

Traces & Spans

Visibility into where the agentic workflow broke



Prompt Engineering

Iterating on prompt templates for improved results

Search & Retrieval

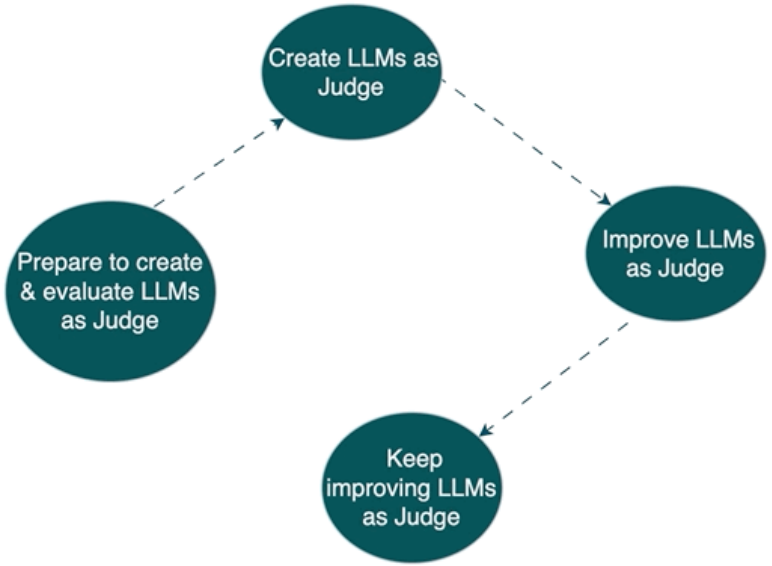
Locate and improve retrieved context

Fine-tuning

Re-train LLM on use case / company data

Source: ArizeAI

evaluation cycle of LLMs as Judge



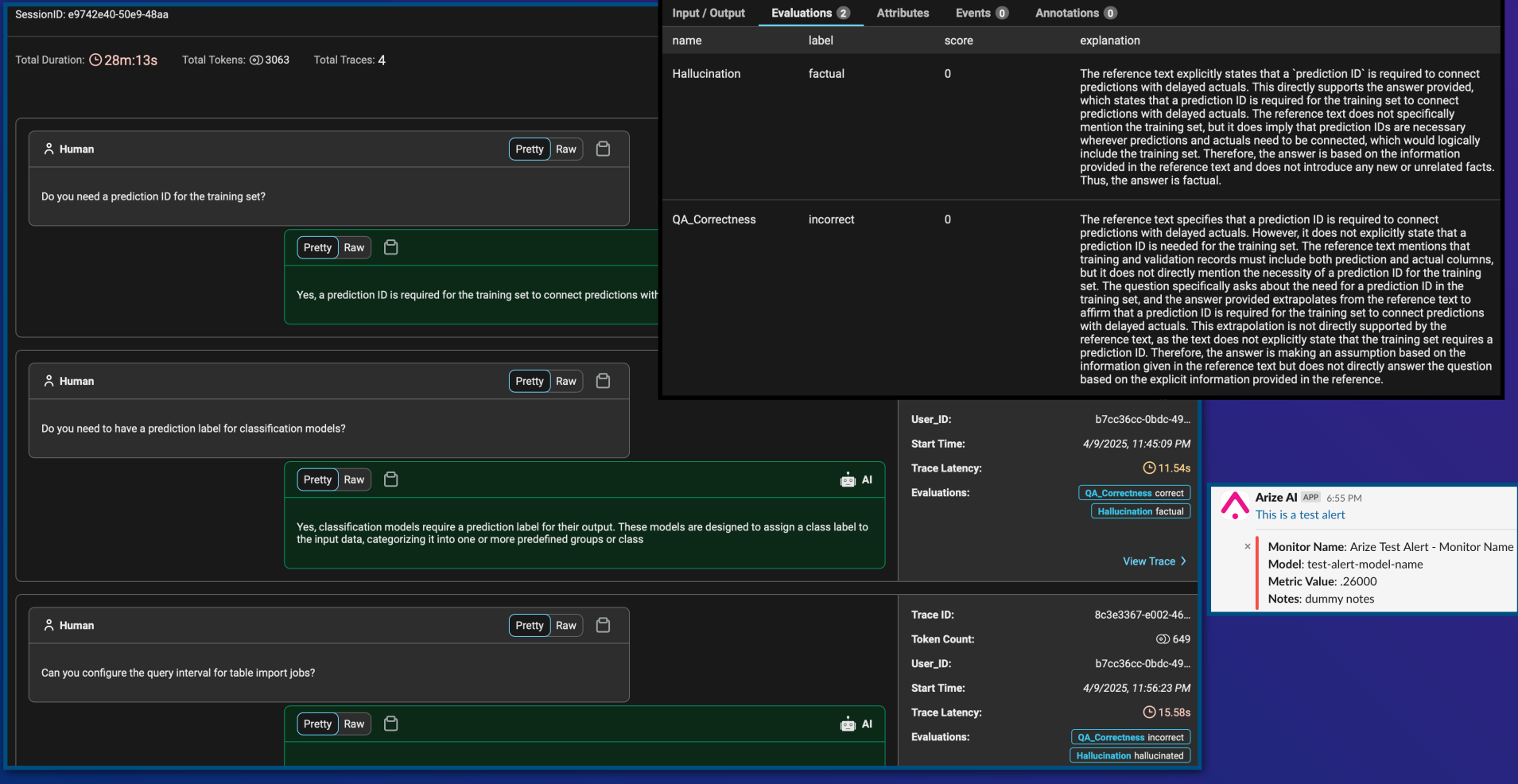

```
ADDITIVE_PROMPT = """
(...)
- Award 1 point if the answer is related to the question.
- Give 1 additional point if the answer is clear and precise.
- Provide 1 further point if the answer is true.
- One final point should be awarded if the answer provides additional resources to support the user.
...
"""
```

```
# Sample examples
ratings_where_raters_agree = ratings.loc[ratings["score_1"] == ratings["score_2"]]
examples = ratings_where_raters_agree.groupby("score_1").sample(7, random_state=1214)
examples["human_score"] = examples["score_1"]

# Visualize 1 sample for each score
display(examples.groupby("human_score").first())
```

```
JUDGE_PROMPT = """
You will be given a user_question and system_answer couple.
Your task is to provide a 'total rating' scoring how well the system_answer answers the user concerns exp
Give your answer as a float on a scale of 0 to 10, where 0 means that the system_answer is not helpful at

Provide your feedback as follows:

Feedback:::
Total rating: (your rating, as a float between 0 and 10)

Now here are the question and answer.

Question: {question}
Answer: {answer}

Feedback:::
Total rating: """
```

```
errors = pd.concat(
    [
        examples.loc[examples["llm_judge_improved_score"] > examples["human_score"]].head(1),
        examples.loc[examples["llm_judge_improved_score"] < examples["human_score"]].head(2),
    ]
)

display(
    errors[
        [
            "question",
            "answer",
            "human_score",
            "explanation_1",
            "llm_judge_improved_score",
            "llm_judge_improved",
        ]
    ]
)
```


Traces
Spans
Total Spans 6.16k
Total Tokens 778.3k
Latency P50 3.709s
Latency P99 49.37s
Statu: Kind Name Input Output Session ID User ID Evaluations Annotations Start Timestamp Latency Total Tokens
CHAIN query Can you configure the... Yes, you can configure... e9742e4... b7cc36c... Hallucinati... hallucinated +1 more -- 4/9/2025, 11:56... PM 15.58s 649
CHAIN query Do you need to have a... Yes, classification... e9742e4... b7cc36c... Hallucinati... factual +1 more -- 4/9/2025, 11:45... PM 11.54s 908
CHAIN query Do you need a... Yes, a prediction ID is... e9742e4... b7cc36c... Hallucinati... factual +1 more -- 4/9/2025, 11:35... PM 4.96s 464
CHAIN query How do you set up... To set up PagerDuty... 25182dc... 1b49075... Hallucinati... factual +1 more -- 4/9/2025, 11:26... PM 11.04s 784
CHAIN query Does the ingestion job... The ingestion job for... 25182dc... 1b49075
CHAIN query How do I send in extra... To send in extra... 647848c... 83caa80
CHAIN query What is the current... The current retention... 647848c... 83caa80
CHAIN query Does Arize store the... Arize stores individual... 647848c... 83caa80
CHAIN query What happens if I... If you upload actuals... 647848c... 83caa80
Spans
pans 6.16k
Total Tokens 778.3k
Latency P50 3.709s
Latency P99 49.37s
Statu: Kind Name Input Output Session ID User ID Evaluations An
RETRIEVER retrieve Can you configure the... e9742e4... b7cc36c... --
CHAIN query Can you configure the... Yes, you can configure... e9742e4... b7cc36c... Hallucinati... hallucinated +1 more
EMBEDDING embedd e9742e4... b7cc36c... --
CHAIN synthesi Can you configure the... Yes, you can configure... e9742e4... b7cc36c... --
LLM llm Yes, you can configure... e9742e4... b7cc36c... --
EMBEDDING embedd e9742e4... b7cc36c... --
RETRIEVER retrieve Do you need to have a... e9742e4... b7cc36c... --
CHAIN query Do you need to have a... Yes, classification... e9742e4... b7cc36c... Hallucinati... factual +1 more
CHAIN synthesi Do you need to have a... Yes, classification... e9742e4... b7cc36c... --


SessionID: e9742e40-50e9-48aa
Total Duration: 28m:13s
Total Tokens: 3063
Total Traces: 4
Human
Pretty
Raw
Do you need a prediction ID for the training set?
Pretty
Raw
Yes, a prediction ID is required for the training set to connect predictions with
Human
Pretty
Raw
Do you need to have a prediction label for classification models?
Pretty
Raw
AI
Yes, classification models require a prediction label for their output. These models are designed to assign a class label to the input data, categorizing it into one or more predefined groups or class
Human
Pretty
Raw
Can you configure the query interval for table import jobs?
Pretty
Raw
AI
span_id: ea21c845-b83f-4a89-8268-75e8b3954f10
4.96s
Prompt Playground
Add to Dataset
Annotate
CHAIN
query
Input / Output
Evaluations 2
Attributes
Events 0
Annotations 0
name
label
score
explanation
Hallucination
factual
0
The reference text explicitly states that a `prediction ID` is required to connect predictions with delayed actuals. This directly supports the answer provided, which states that a prediction ID is required for the training set to connect predictions with delayed actuals. The reference text does not specifically mention the training set, but it does imply that prediction IDs are necessary wherever predictions and actuals need to be connected, which would logically include the training set. Therefore, the answer is based on the information provided in the reference text and does not introduce any new or unrelated facts. Thus, the answer is factual.
QA_Correctness
incorrect
0
The reference text specifies that a prediction ID is required to connect predictions with delayed actuals. However, it does not explicitly state that a prediction ID is needed for the training set. The reference text mentions that training and validation records must include both prediction and actual columns, but it does not directly mention the necessity of a prediction ID for the training set. The question specifically asks about the need for a prediction ID in the training set, and the answer provided extrapolates from the reference text to affirm that a prediction ID is required for the training set to connect predictions with delayed actuals. This extrapolation is not directly supported by the reference text, as the text does not explicitly state that the training set requires a prediction ID. Therefore, the answer is making an assumption based on the information given in the reference text but does not directly answer the question based on the explicit information provided in the reference.
User_ID:
b7cc36cc-0bdc-49...
Start Time:
4/9/2025, 11:45:09 PM
Trace Latency:
11.54s
Evaluations:
QA_Correctness correct
Hallucination factual
View Trace
Trace ID:
8c3e3367-e002-46...
Token Count:
649
User_ID:
b7cc36cc-0bdc-49...
Start Time:
4/9/2025, 11:56:23 PM
Trace Latency:
15.58s
Evaluations:
QA_Correctness incorrect
Hallucination hallucinated
Arize AI
APP
6:55 PM
This is a test alert
Monitor Name: Arize Test Alert - Monitor Name
Model: test-alert-model-name
Metric Value: .26000
Notes: dummy notes

projects > **tracing-agent**

Total Traces	Total Tokens	Latency P50	Latency P99
1	1,115	0.52s	0.52s

Stream Last 7 Days

Spans Traces Sessions Config

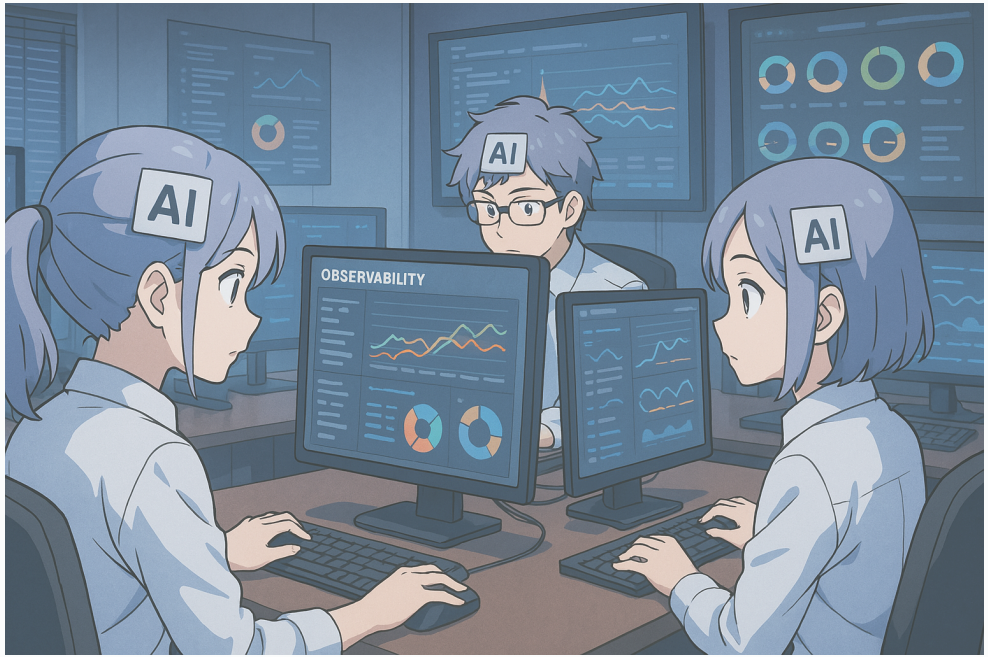
filter condition (e.x. span_kind == 'LLM')

Root Spans All Columns

	sta...	kind	name	input	output	annotations	start time	latency	cumulative
☐	✓	agent	AgentRun	[{"role": "user", "content...	The stores that perfor...		5/15/2025, 10:22 AM	0.51s	1115

the impact of AI Agents & LLMs as Judge on olly platforms

- enrich olly platforms by automating the incident remediation cycle
- add visibility & communication between all infrastructure systems
- reduce MTTR & bring down incident costs
- self-healing systems
- integrate AI agents workflows with IDEs directly in the olly platforms



resources

https://devm.io/devops/generativeai-observability-serverless
https://devm.io/machine-learning/generative-ai-transformer-architecture
https://devm.io/devops/beyond-rag-fine-tuning-machine-learning-devops
JuStRank: Benchmarking LLM Judges for System Ranking: https://arxiv.org/pdf/2412.09569
Agent-as-a-Judge: Evaluate Agents with Agents: https://arxiv.org/pdf/2410.10934
https://malywut.github.io/ai-engineering-landscape/
https://arize.com/blog/the-role-of-opentelemetry-in-llm-observability/
https://opentelemetry.io/blog/2025/ai-agent-observability/
CNCF Slack #otel-genai-instrumentation

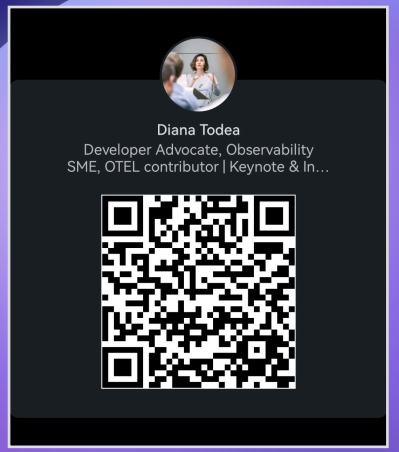
Platmosphere

A MIA-PLATFORM INVITATION

Thanks
Mille grazie

Q&A Session

SCAN & SEND YOUR QUESTION →



AN EVENT BY mia Platform