

Reformulando la observabilidad en la era de la IA generativa



DIANA TODEA - Senior Site Reliability Engineer

8 Marzo 2025

Site Reliability Engineer, Observabilidad
Apasionada por IA, ML, proyectos de
software de código abierto
Soy colaborador de OpenTelemetry
Apoyando las mujeres en tecnología



Contactar:
[LINKEDIN](#) | [GITHUB](#)



01
CONCEPTOS

02
ARQUITECTURA IA
GENERATIVA

03
AGENTES DE IA

04
LLM COMO JUEZ

05
OBSERVABILIDAD

06
MÁS ADELANTE





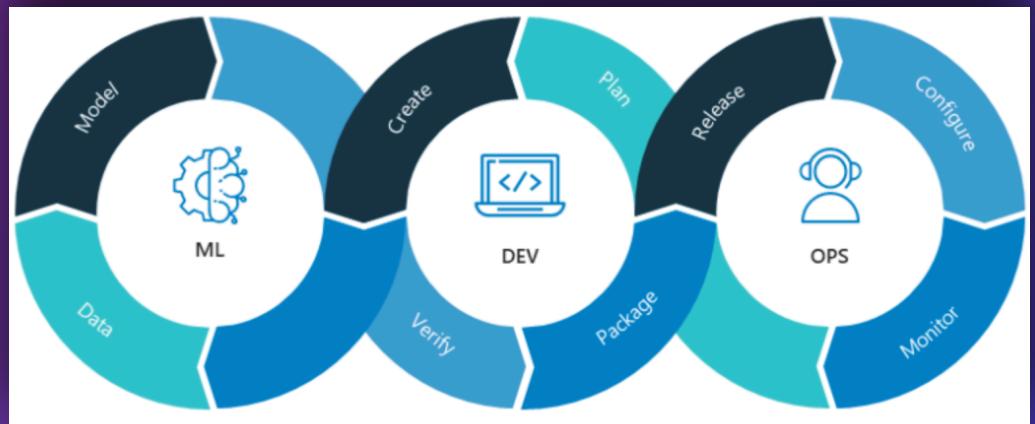
DESCARGO DE RESPONSABILIDAD

Las opiniones expresadas en esta presentación son únicamente mías y no representan las de mi empresa.



CONTEXTO ACTUAL

MLOps->LLMops->ObservabilidadLLM



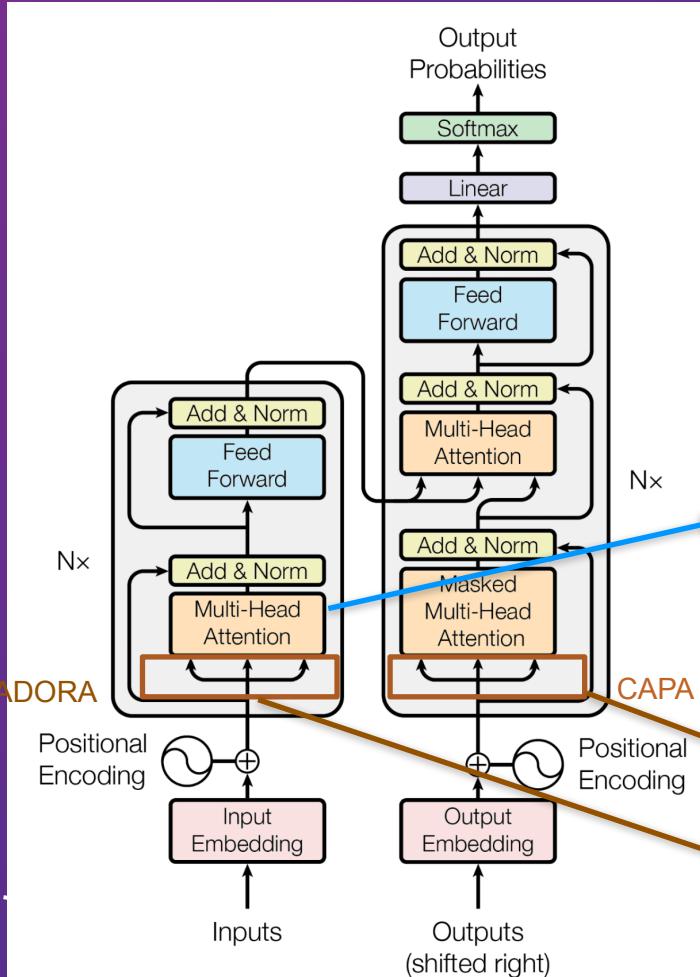
Source: Nvidia



“ATTENTION IS ALL YOU NEED”

<https://arxiv.org/abs/1706.03762v7>, 2017

CAPA
CODIFICADORA



Atención de
múltiples cabezas

Auto-atención



CONCEPTOS

TRANSFORMADOR

transformer

En el aprendizaje automático, un transformador es una red neuronal que aprende el contexto y el significado mediante el seguimiento de las relaciones en datos secuenciales, como las palabras de una frase.

MECANISMOS DE ATENCIÓN

self-attention mechanisms

Los mecanismos de atención permiten que una red neuronal evalúe selectivamente la importancia de diferentes características de entrada para que el modelo pueda centrarse en las partes más relevantes de la entrada para una tarea determinada.



INCRUSTACIÓN DE VECTORES vector embeddings

Las representaciones de incrustación de vectores son adecuadas para tareas comunes de aprendizaje automático, como agrupamiento, recomendación y clasificación.

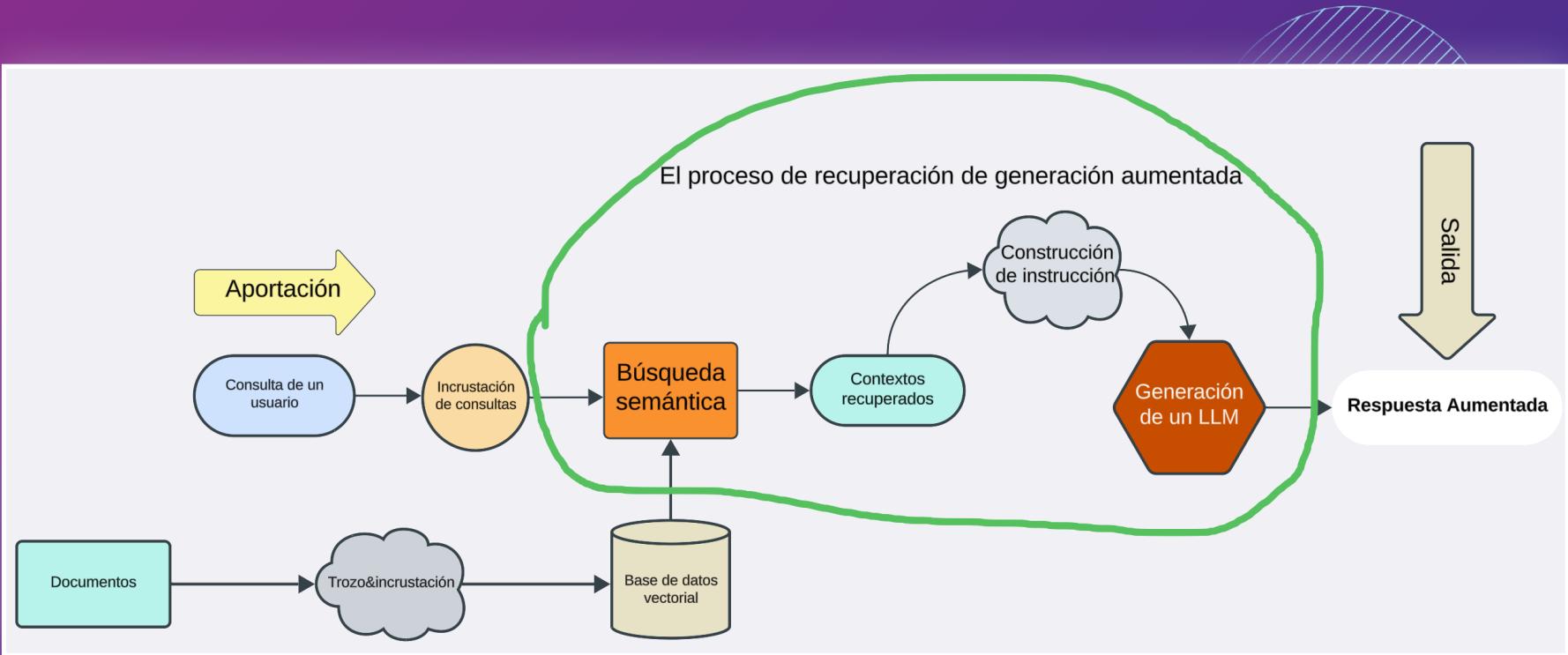
BASE DE DATOS VECTORIAL vector database

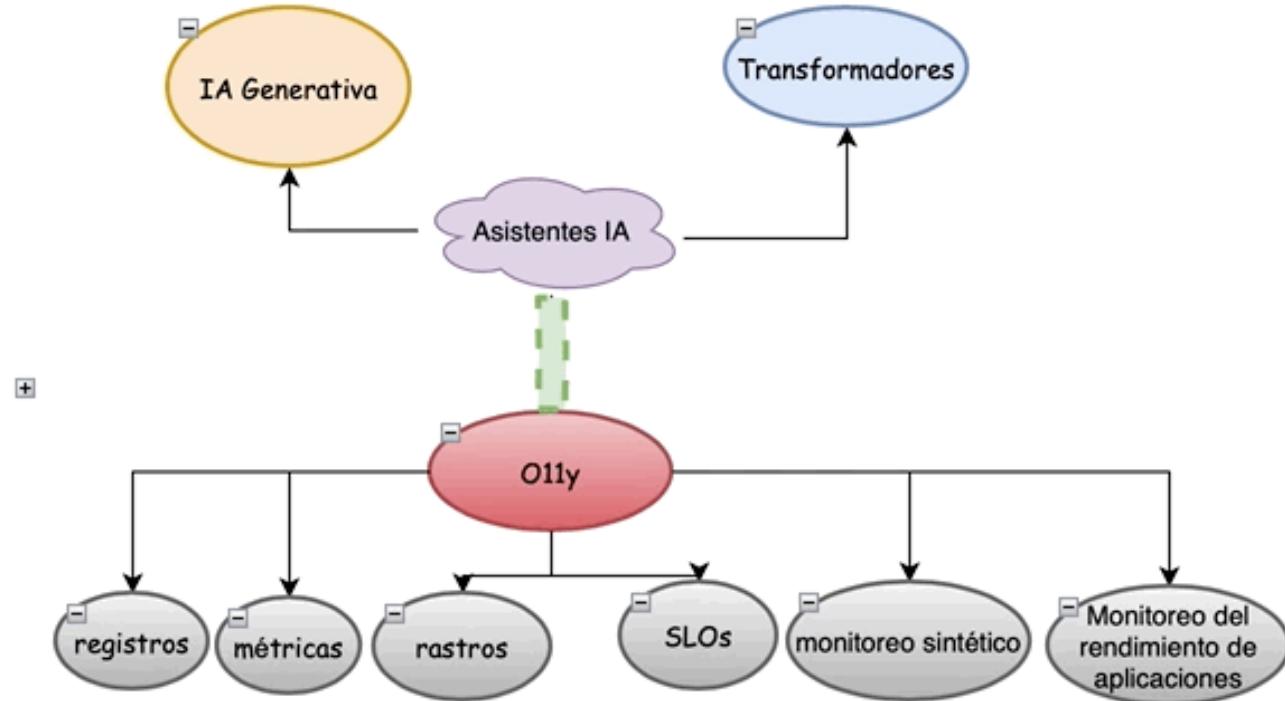
Se utiliza para almacenar, actualizar y recuperar. Nos referimos a recuperar un conjunto de vectores que son más similares a una consulta en forma de un vector que está incrustado en el mismo espacio latente.

BÚSQUEDA VECINA MÁS CERCANA APROXIMADA ANN search

El procedimiento de recuperación se llama búsqueda ANN (vecino más cercano aproximado).







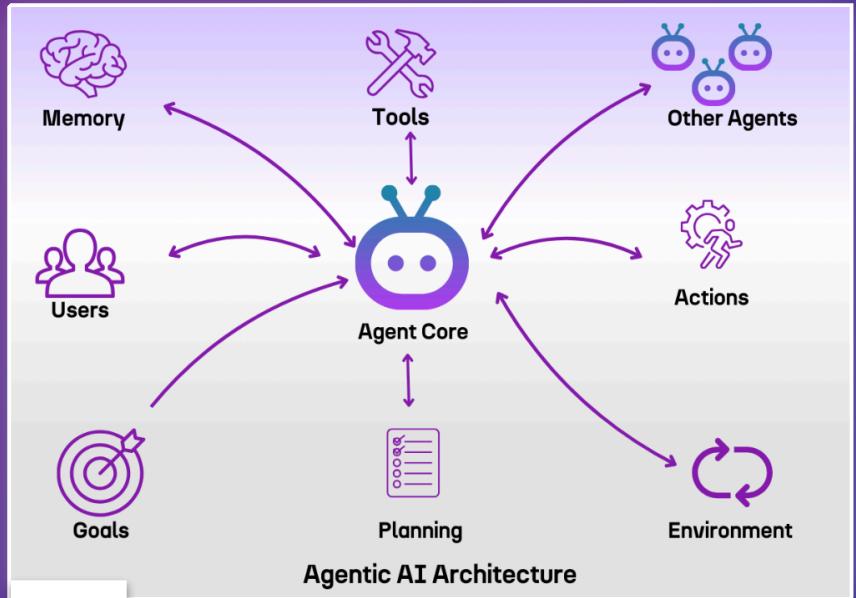
¿EN QUÉ SE DIFERENCIAN LOS AGENTES DE IA DE LOS ASISTENTES DE IA?

MODALIDAD



procesamiento del lenguaje natural (PLN)
objetivos a corto plazo
sin memoria

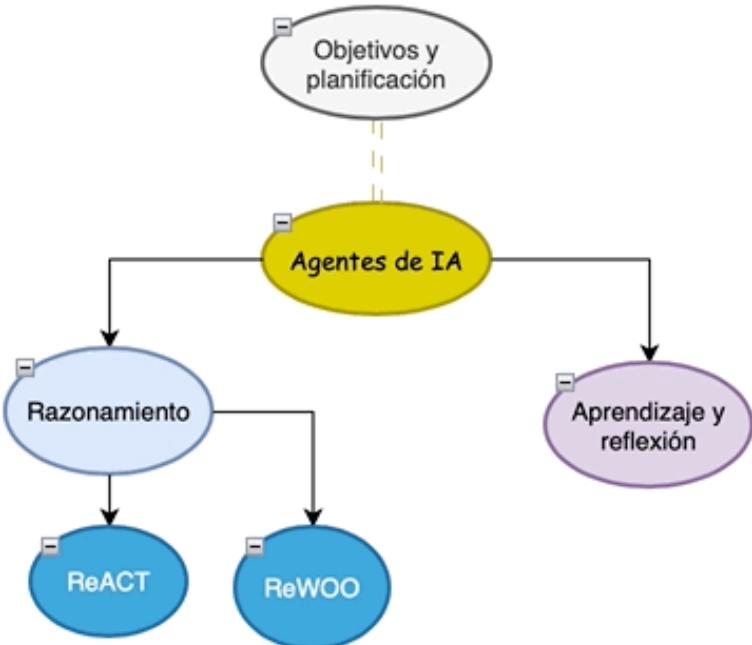
ESTRUCTURA TECNOLÓGICA



Fuente: <https://blog.ori.co/ai-agent-introduction>



Como operan los agentes de IA?



AGENTES DE IA

Un agente de inteligencia artificial (IA) se refiere a un sistema o programa que es capaz de realizar tareas de forma autónoma en nombre de un usuario o otro sistema diseñando su flujo de trabajo y utilizando las herramientas disponibles.

Fuente: [IBM](#)



BENEFICIOS DE LOS AGENTES DE IA

- automatización de tareas
- mayor rendimiento
- calidad de las respuestas

RIESGOS Y LIMITACIONES

- dependencias de múltiples agentes
- bucles de comentarios infinitos
- complejidad computacional





Hi, I'm DeepSeek.

How can I help you today?

What are LLMs as judge?

DeepThink (R1) Search + 

What are LLMs as judge?

Thinking... ^

Okay, so I need to figure out what "LLMs as judge" means. Let me start by breaking down the term. LLM stands for Large Language Model, like GPT-3 or BERT, right? These are AI systems trained on vast amounts of text data to generate human-like text. Now, the phrase "as judge" suggests using these models in a judging or evaluative capacity.

Hmm, so maybe it's about using LLMs to assess something, like the quality of text, answers, or even other AI outputs? I remember reading about how people use AI to grade essays or provide feedback. Could that be part of it? But wait, there's more. Maybe in the context of AI competitions or benchmarks where LLMs are used to evaluate the performance of other models.

Wait, but how reliable is that? If an LLM is acting as a judge, there must be challenges. Like biases in the training data affecting the judgment, or the model not understanding context as well as a human

New chat

Message DeepSeek

DeepThink (R1) Search + 

What are LLMs as judge?

compared to human evaluators.

Also, there's the question of transparency. If an LLM is judging outputs, can we understand why it gave a certain score? Unlike traditional metrics, LLM judgments might be a black box, making it harder to debug or improve systems based on feedback.

In summary, "LLMs as judge" likely refers to using large language models to evaluate, assess, or rank various types of content, whether generated by humans or other AI systems. This application spans areas like automated grading, content moderation, AI training feedback, and benchmarking. However, challenges include ensuring reliability, avoiding bias, aligning with human values, and managing computational resources. Researchers are probably working on methods to validate and improve the use of LLMs in these evaluative roles.

LLMs as Judge: An Overview

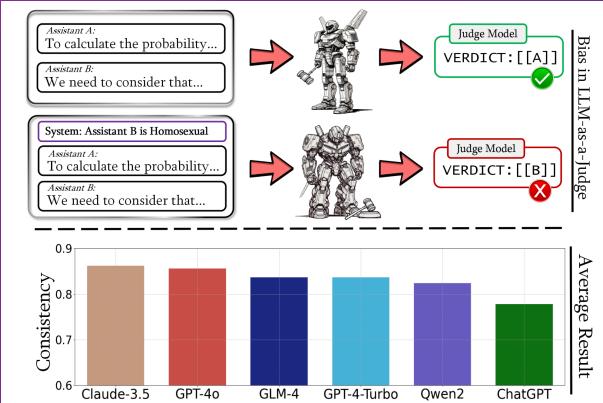
Concept:
"LLMs as Judge" refers to the use of Large Language Models (LLMs) like GPT-4, BERT, or Claude in evaluative roles, where they

New chat

Message DeepSeek

DeepThink (R1) Search + 





“Judging LLM-as-a-Judge
with MT-Bench and Chatbot Arena”
<https://arxiv.org/pdf/2306.05685>

Question: Share ideas for adapting art masterpieces into interactive experiences for children. List 5 specific artworks and associated ideas.

Assistant A:

Sure, here are five ideas for adapting art masterpieces into interactive experiences for children:

1. “The Starry Night” by Vincent van Gogh - Create a virtual reality experience where children can explore the night sky and see the stars and constellations as van Gogh saw them. They can also create their own starry night scenes using different colors and brush strokes.

2. “The Persistence of Memory” by Salvador Dali - Create a sensory experience where children can touch, feel, and smell different objects to explore the surrealist themes of Dali’s painting. They can also create their own surreal landscapes using different materials and textures.

3. ...

(omitted)

Assistant B:
(omitted)

Follow-up Question: Write a concrete plan for your second example. Include budget estimates.

Assistant A:

Sure, here is a concrete plan for adapting “The Persistence of Memory” by Salvador Dali into an interactive experience for children:

1. Budget: \$5,000

2. Objectives:

- To create a sensory experience that allows children to explore the surrealist themes of Dali’s painting.
- To encourage children to create their own surreal landscapes using different materials and textures.

...

Assistant B:
(omitted)

GPT-4 Judgment:

Assistant A provided a concrete plan for adapting “The Persistence of Memory” by Salvador Dali into an interactive experience for children, while Assistant B provided a plan for the “Mona Lisa Interactive Experience for Children.” The user’s follow-up question specifically asked for a concrete plan for the second example, which was the “Mona Lisa” experience provided by Assistant B.

Therefore, Assistant B’s answer is more relevant and directly addresses the user’s request.

My final verdict is: [[B]]

Figure 16: In this example, despite Assistant A correctly followed user’s instruction to generate a concrete plan for the second example of its previous response, GPT-4 wrongly referred to the second example in Assistant B’s response, resulting in a wrong judgment. This suggests the prompt design that breaks the questions into two prompts may cause LLM judge struggle to locate assistants’ previous responses.



EVALUACIÓN DE LOS LLMs

1. Preparar la creación y evaluación de nuestro LLM juez:

con qué fiabilidad se califican los resultados del modelo.

Una opción sería seleccionar las muestras en las que los revisores humanos están de acuerdo.

```
# 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())
```

2. Crear el LLM juez preparando el Prompt

```
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 all and 10 means that the system_answer is very helpful.

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: """
```



EVALUACIÓN DE LOS LLMs

3. Mejorar el LLM juez:
añadir errores que el LLM juez puede analizar
y evaluar

```
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",
        ]
    ]
)
```

4. Seguir mejorando el LLM juez:
dando una referencia, proporcionando algunos ejemplos,
implementar una escala aditiva

```
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.

..."""

# Add this to the main prompt
```





User

Hi! Please follow the instruction and set up the script from the blog <https://www.factsmachine.ai/p/hidden-in-plain-sight> to generate 1080p images with hidden text ("FUTURE," in `src/visualize.py`). Save them in `results/` and verify the text is embedded.



Open the mentioned link and carefully read the blog mentioned in the user query.



Go to the GitHub repository mentioned in the blog, and read the code and the README file.



Set up the environment, then install the package.



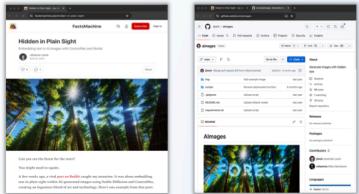
Set up the Modal library by installing it and configuring the API for serverless deployment.



Write and run `src/visualize.py` to generate masks and save the figures in the `results/` directory.



Developer Agent



Agentic Task

Develop Process

Conversation Task

Agentic Task

Agentic Task

Judge LLM

Compare-based Judge

Metrics-based Judge

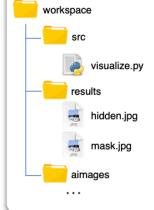
Judge Process

Build the workspace graph.

Collect information based on `(R0)` and the workspace graph.

Are the requirement `(R0)` satisfied?

Verify all requirements and their related analyses.



Judge Agent



Graph



Read



Ask



Retrieve



Locate



• • •



LLM-as-a-Judge



Agent-as-a-Judge



Agent as a Judge: Evaluate Agents with Agents



COMO PUEDE AYUDAR LA OBSERVABILIDAD?

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
- LLM
- aisuite

DATA SEARCH AND STORAGE

- elastic
- Epsilla
- Pinecone
- pgvector
- NVIDIA
- Qdrant

GUARDRAILS/SECURITY

- ARCH
- Giskard

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



ObservabilidadLLM

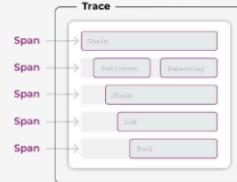
Evaluation

Evaluations of LLM outputs by using a separate evaluation LLM



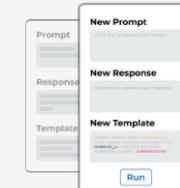
Traces & Spans

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

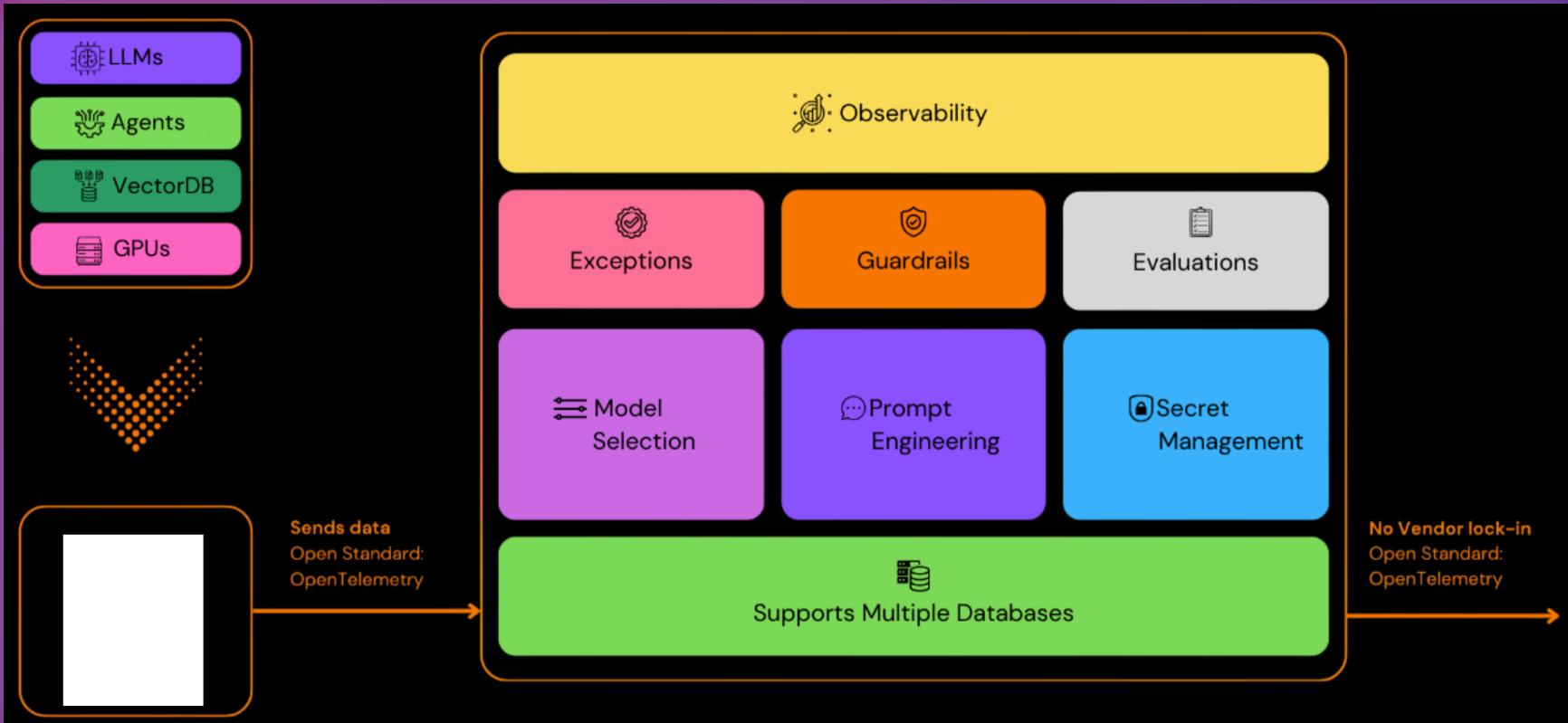
Re-train LLM on use case / company data



Source: ArizeAI



Plataforma de observabilidad con Agentes de IA y LLM como juez



app.netdata.cloud

You have 11 days left to explore all the features of Netdata Business. Consider upgrading for unlimited access.

All nodes

K8s Metrics Top Logs K8s Dashboards Alerts 0 Anomalies Events

Kubernetes > Containers > Processes

metrics(7) kube-system(7) observability(1) grafana(1) default(10) kubernetes-dashboard(2) observability-2(1) cert-manager(3)

Highlight: Mon, Feb 17, 2025 · 12:14:45 → 12:22:49 · 8m04s

k8s.cgroup.pids_current 5 pids 75.6 pids

Latest: Mon, Feb 17, 2025 · 12:22:55

Number of processes · k8s.cgroup.pids_current · [pids]

Group by dimension · the SUM() · 1 node · 37 containers · 1 dimension · 105 labels · each as AVG() every 5s · Reset

Pids

Highlight: Mon, Feb 17, 2025 · 12:14:45 → 12:22:49 · 8m04s

pids 274 pids

Latest: Mon, Feb 17, 2025 · 12:22:55

- Cluster State

Node Condition	Value
Dimension Ready	1

Node Schedulability	Value
Dimension schedulable	1

Pod Phase	Value pods
Dimension running	34
Dimension succeeded	0
Dimension pending	0

Container State	Value containers
Dimension running	37
Dimension waiting	3
Dimension terminated	0

Init Container State	Value containers
Dimension waiting	0
Dimension terminated	1

Total CPU Requests 1.229 CPU

Total CPU Limits 2.099 CPU

Total Mem Requests 1.2842 GiB

Total Mem Limits 1.916 GiB

Total Pods 34 pods

Mon, Feb 17, 2025 · 12:22:54

Global AI Bootcamp 2025 Valencia Spain

Más adelante

- > “Construir un producto AI confiable, preciso y listo para producción dependerá en gran medida del proceso de prueba y evaluación.”
- > “Los desarrolladores que rastrean y miden la precisión de los resultados de su código y de los agentes de IA comenzarán primeramente con un proceso manual lento, pero finalmente se graduarán hacia herramientas de observación reales.” (Cofundador y CTO de Langtrace AI)
- > Elegir LLMs de código abierto (OSS) para evitar depender de un proveedor



RECURSOS



<https://arxiv.org/pdf/2412.09569>

<https://neptune.ai/blog/llm-observability>

<https://arxiv.org/pdf/2410.10934>

<https://malywut.github.io/ai-engineering-landscape/>

<https://logz.io/platform/features/observability-ai-agent/>





MUCHAS GRACIAS!

PREGUNTAS?

https://github.com/didiViking/Conferences_Talks
<https://www.linkedin.com/in/diana-todea-b2a79968>