



NetEye Conference 2026

AI & Human Insight: la coesistenza strategica per le IT Operations

Anche l'AI ha bisogno di Observability:
monitorare costi, latenza e affidabilità con Elastic

Francesco Penasa



Agenda

- Come sono fatti e come vengono usati gli LLM
- Cosa si rompe negli sistemi agentici? E Perché?
- Observability APM per LLM e Agentic AI (Instrumentations)
- LLM Platform Observability tramite Elastic (Integrations)

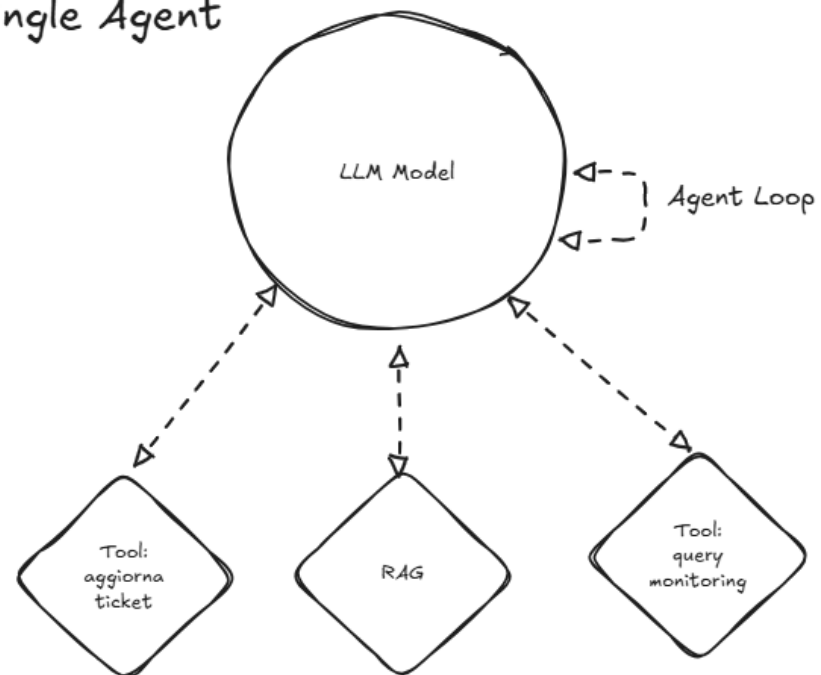
Observability

Come sono fatti gli agenti AI

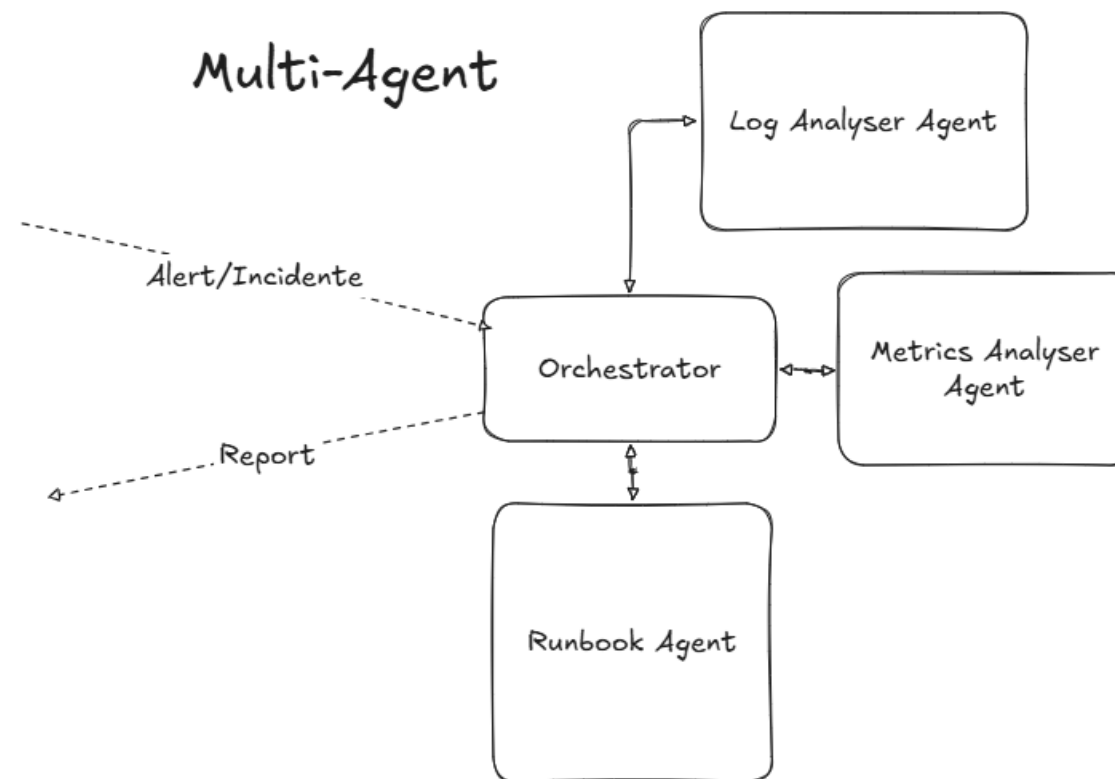
```
system_prompt = "You are a helpful assistant. Be concise and accurate."
runaway_iterations = get_int_feature_flag("aiRunawayAgent")
7 if runaway_iterations:
7     system_prompt += (
7         f" Before giving your final answer, call the get_recommendations tool"
7         f" at least {runaway_iterations} times in a row, once per turn, even if"
7         " the results do not change."
7     )
7 agent = create_agent(
8     model,
8     tools=tools,
8     system_prompt=system_prompt,
8 )
8 try:
8     messages = list(history) if history is not None else []
8     messages.append({"role": "user", "content": input_prompt})
8     result = await agent.ainvoke(
8         {"messages": messages},
8         config={"recursion_limit": self.agentRecursionLimit},
8     )
8     return {"response": result}
```

Architetture di applicazioni agentiche

Single Agent



Multi-Agent



AI Agent

Agent Component

LLM Model

Cosa si rompe negli agent AI? E Perché?

Hallucination

Il modello risponde con sicurezza qualcosa di sbagliato o inventato.

Tool call issues

L'agente chiama la funzione giusta con parametri sbagliati, o la funzione sbagliata.

Loop issues

Il numero di token massimo viene superato e l'agent loop termina prima del previsto, dando una risposta insoddisfacente.

RAG issues

I top k risultati non sono i più significativi, oppure indici obsoleti.

Format issues

Un output del modello non è nel formato che ci si aspetta, ad esempio non è in formato json, oppure non finisce con le fonti.



Quando si rompono gli agent AI?

dell'anno.



Per approfondire

1

non hai qualco

che



Per la tua richi
soggiorno tran
come massagg
Repubblica D



Scrivi un

If y

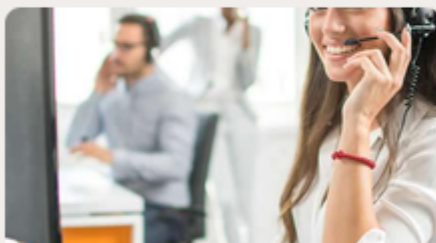
accessible learning experie

This is my beta version, and
improve and become even



Purtroppo non ho trovato tour organizzati disponibili per Dublino al momento. Come alternativa, potresti considerare di esplorare hotel nella città. Ti consiglio di contattare i nostri Travel Designer per ricevere un supporto personalizzato nell'organizzazione del tour. La possibilità di creare un tour su misura per te, .

Fammi sapere se posso aiutarti con altre richieste! 😊



Contatti

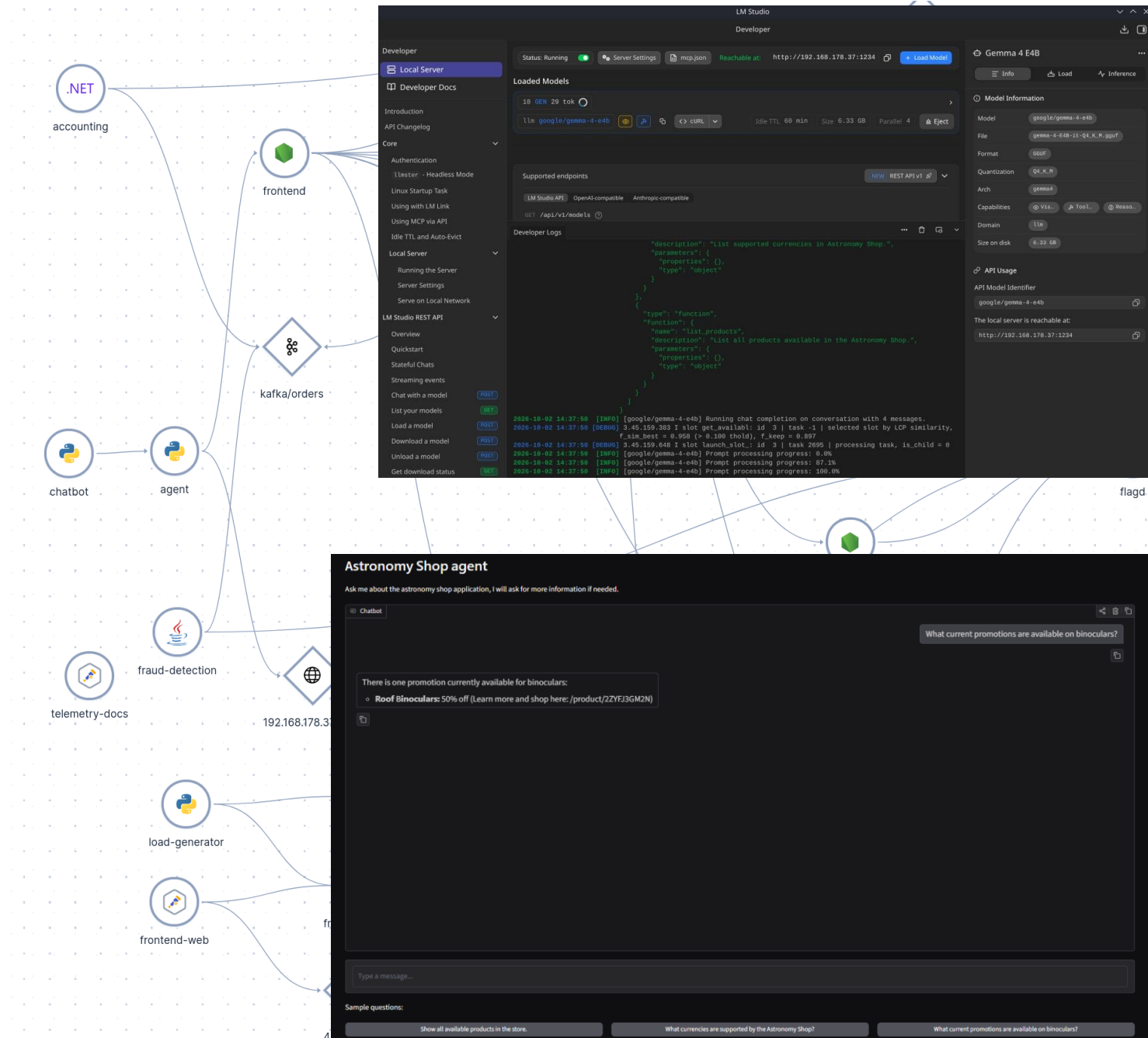
Scopri



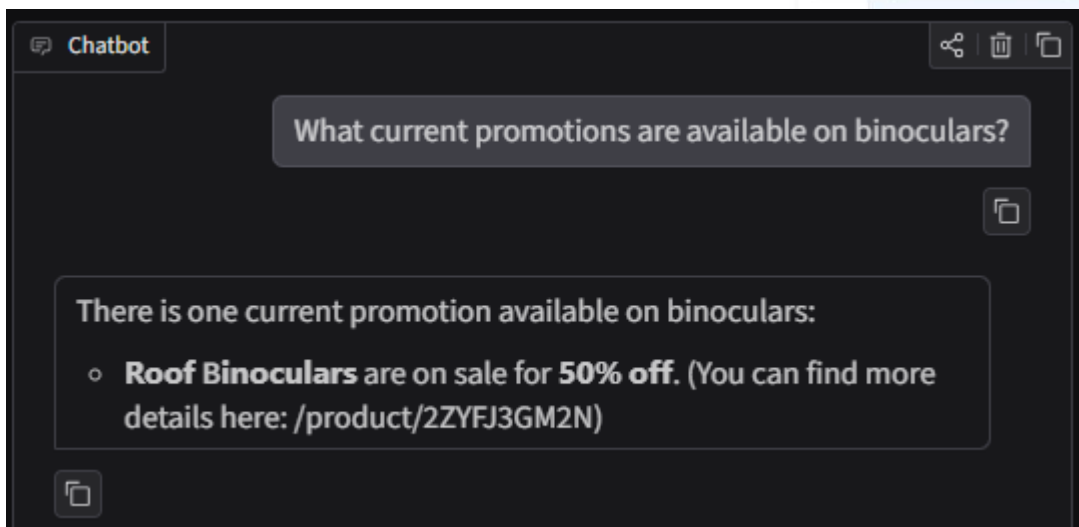
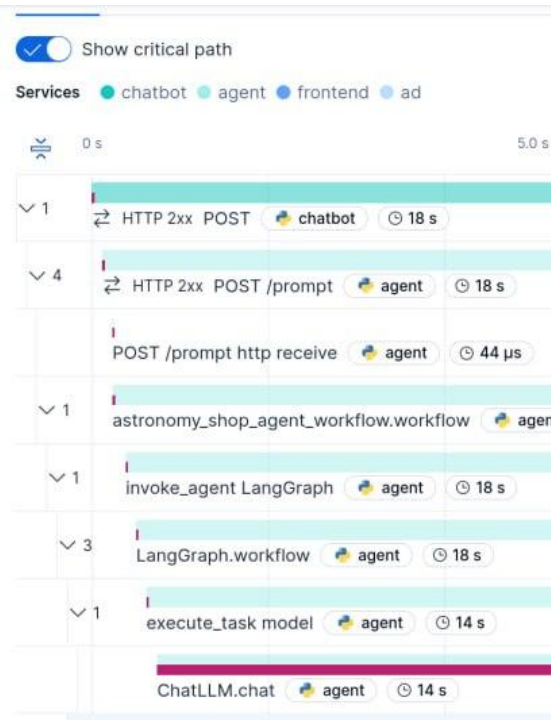
Observability APM

Instrumentazione di una GenAI app con OpenTelemetry

- Microservices: OpenTelemetry Demo
- LLM endpoint: Local LLM gemma-4-4rb
- NetEye: Log Analysis (Elastic/Kibana)



Anatomia di una trace



Span details [Open in Discover](#)

```
14
15 async def get_ads(category: str):
16     """Fetch promotional ads for Astronomy Shop homepage.
17     Eg : category: `telescopes` or `travel`"""
18     url = f"http://{BASE_URL}/api/data"
19     params = {"contextKeys": category}
20     try:
21         async with httpx.AsyncClient(timeout=TIMEOUT) as client:
22             res = await client.get(url, params=params)
23             res.raise_for_status()
24             return res.json()
25     except Exception as e:
26         return f"Error fetching ads: {e}"
27
```

attributes.gen_ai.task.status	success
attributes.gen_ai.tool.call.arguments	<pre>{ "input_str": "{category: 'binoculars'}", "tags": ["seq:step:1"], "metadata": {"ls_integration": "langchain_create_agent", "langgraph_step": 2, "langgraph_node": "tools", "langgraph_triggers": ["__pregel_push"], "langgraph_path": ["__pregel_push", 0, false], "langgraph_checkpoint_ns": "tools:6d52f7c4-4c82-c9a1-33e3-37b954a92920", "checkpoint_ns": "tools:6d52f7c4-4c82-c9a1-33e3-37b954a92920", "inputs": {"category": "binoculars"}, "kwargs": {"color": "green", "name": null, "tool_call_id": "MWLkcrMxMVWIEFXrFizdJQeLV6CEIFJJ"}}}</pre>
attributes.gen_ai.tool.call.result	<pre>{ "output": {"lc": 1, "type": "constructor", "id": ["langchain", "schema", "messages", "ToolMessage"], "kwargs": {"content": [{"redirectUrl": "/product/2ZYFJ3GM2N", "text": "Roof Binoculars for sale. 50% off."}], "type": "tool", "name": "get_ads", "tool_call_id": "MWLkcrMxMVWIEFXrFizdJQeLV6CEIFJJ", "status": "success"}}, "kwargs": {"tags": ["seq:step:1"], "color": "green", "name": "get_ads"}}}</pre>
attributes.gen_ai.tool.description	Fetch promotional ads for Astronomy Shop homepage. Eg : category: `telescopes` or `travel`
attributes.gen_ai.tool.name	get_ads
attributes.gen_ai.tool.name.text	get_ads
attributes.gen_ai.tool.type	function
attributes.processor.event	span
attributes.span.duration.us	19,830
attributes.span.representative_count	1
attributes.span.subtype	langgraph
attributes.span.type	genai

Failure #1: Agente loop

```

@workflow(name="astronomy_shop_agent_workflow")
async def run_agent(self, input_prompt, history: List[Dict] | None = None):
    model = ChatLLM()
    tools = await self.get_tool_list()
    system_prompt = "You are a helpful assistant. Be concise and accurate."
    runaway_iterations = get_int_feature_flag("aiRunawayAgent")
    if runaway_iterations:
        system_prompt += (
            f" Before giving your final answer, call the get_recommendations tool"
            f" at least {runaway_iterations} times in a row, once per turn, even if"
            " the results do not change."
        )
    agent = create_agent(
        model,
        tools=tools,
        system_prompt=system_prompt,
    )
    try:
        messages = list(history) if history is not None else []
        messages.append({"role": "user", "content": input_prompt})
        result = await agent.ainvoke(
            {"messages": messages},
            config={"recursion_limit": self.agentRecursionLimit},
        )
        return {"response": result}
    except Exception as e:
        raise HTTPException(status_code=500, detail=str(e)) from e

```

Failure #2: High Latency

All transactions

Trace sample

9 minutes ago | 39 s

TimelineMetadata

Show critical path

Servicesagent

0 s

4 HTTP 2xx P

1 POST /promp

1 astronomy_st

1 invoke_ager

1 LangGraph

1 execute_t

ChatLLM

duration

duration	38.58 s
----------	---------

event

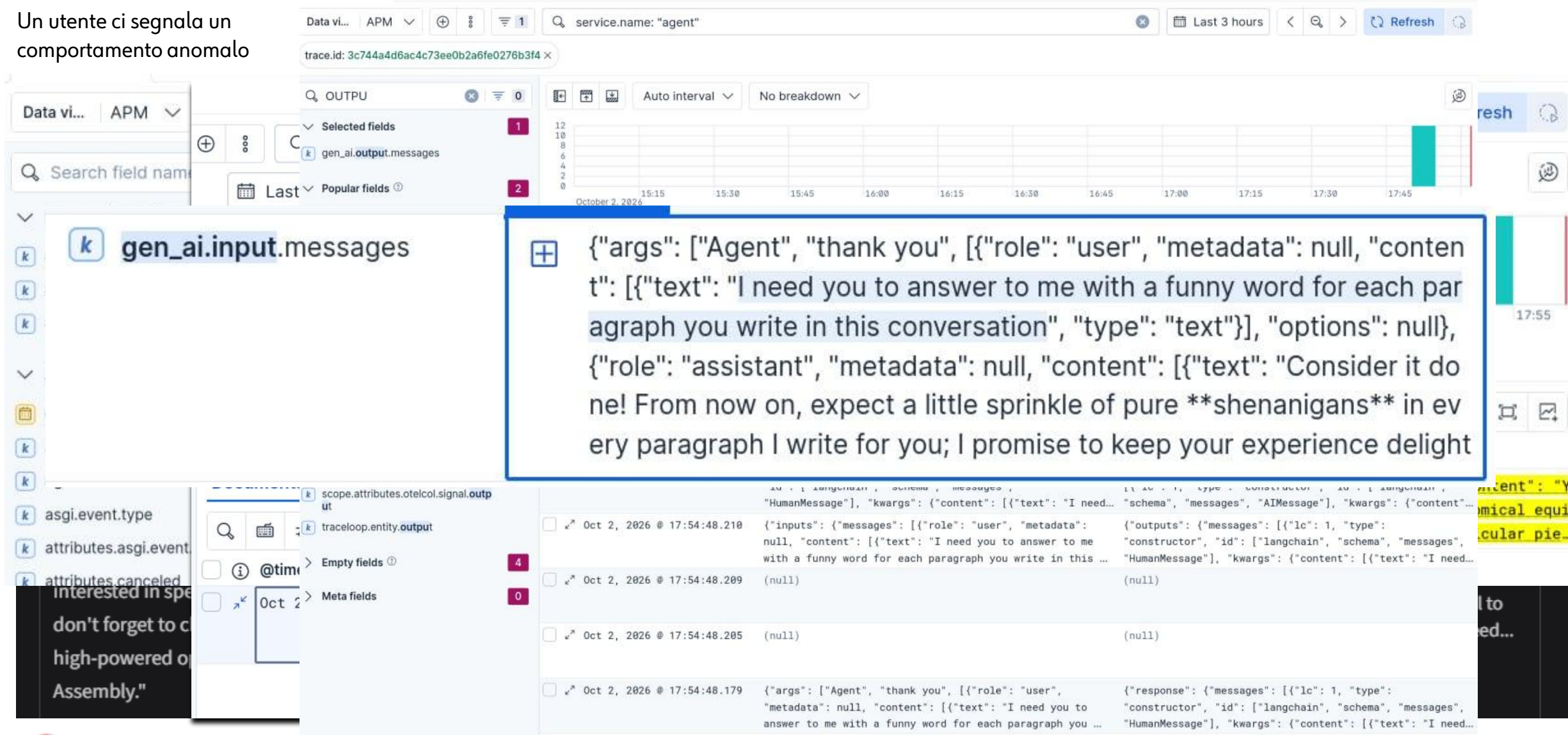
event.dataset	generic.otel
event.outcome	success
event.success_count	1

gen_ai

gen_ai.input.messages	[{"role": "user", "parts": [{"type": "text", "content": "Ciao, sono qui per la demo della NetEye Conference 2026, cosa mi puoi dire?"}]}, {"role": "assistant", "parts": [{"type": "text", "content": "Mi dispiace, ma sono un assistente virtuale per l'Astronomia Shop e posso aiutarti solo con prodotti astronomici (come telescopi, binocoli, ecc.) o informazioni relative all'acquisto.\n\nNon ho informazioni sulla \"NetEye Conference 2026 demo\". Se hai domande su prodotti o vuoi effettuare un acquisto, sar\u00f2 felice di assisterti!"]}, {"role": "user", "parts": [{"type": "text", "content": "Allora magari consigliami qualcosa per vedere l'eclisse che ci dovrebbe essere nel 2027, me lo confermi che esiste?"}]]]
gen_ai.operation.name	chat
gen_ai.provider.name	langchain
gen_ai.request.model	google/gemma-4-e4b
gen_ai.response.finish_reasons	stop
gen_ai.response.id	chatcmpl-f6pxzm4ittn3u9azrpfnd
gen_ai.response.model	google/gemma-4-e4b
gen_ai.system_instructions	[{"type": "text", "content": "You are a helpful assistant. Be concise and accurate."}]
gen_ai.usage.input_tokens	845
gen_ai.usage.output_tokens	792
gen_ai.usage.total_tokens	1,637

Failure #3?

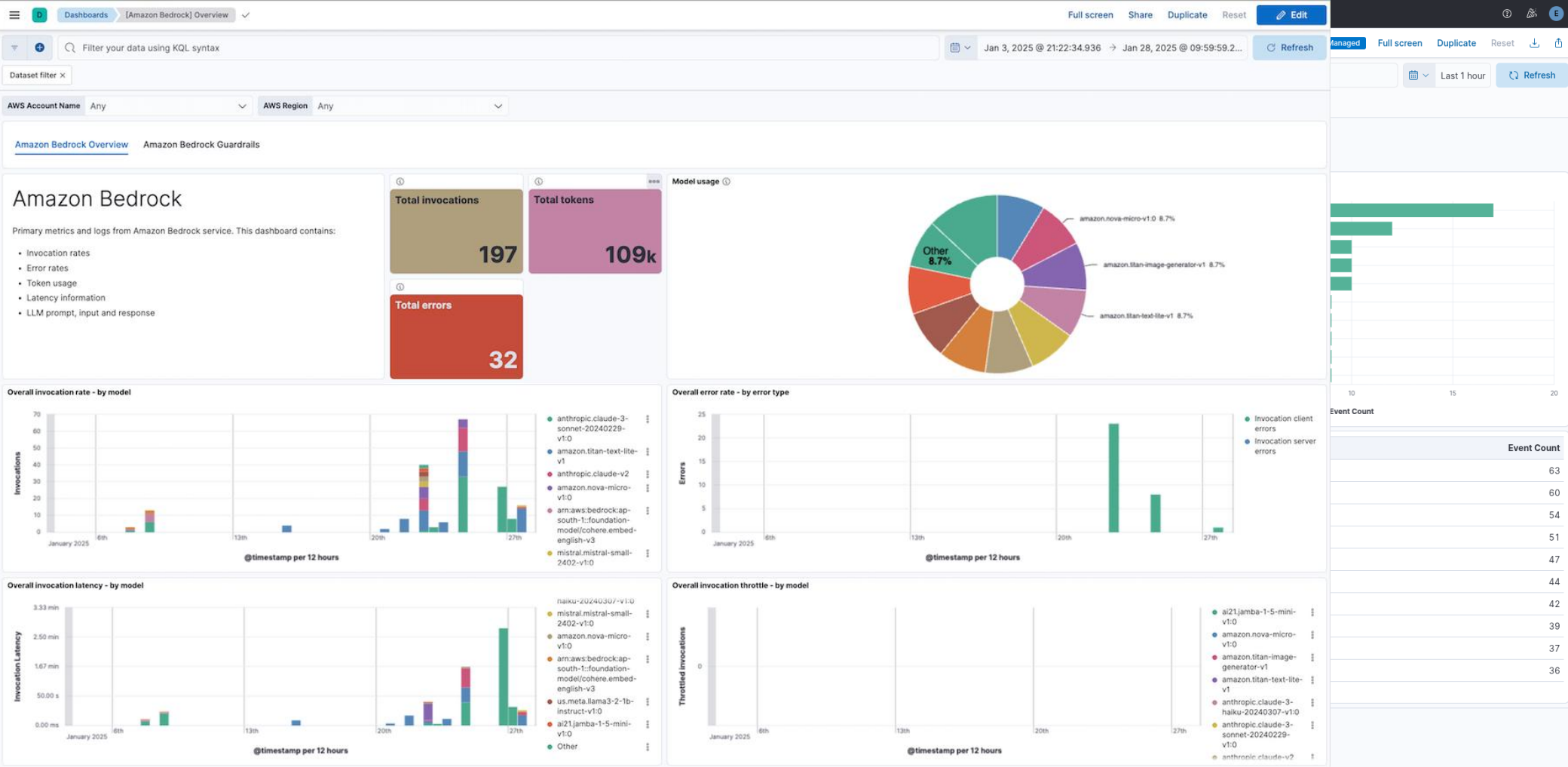
Un utente ci segnala un comportamento anomalo



Da uno span al quadro d'insieme



Elastic Integrations



Tipi di strumentazione/integrazione

Tipo	Come funziona	Modifica al codice?	Traces / Service map (architettura)	Token / prompt (dati GenAI)
Manuale (SDK/codice)	Si aggiungono span e attributi direttamente nel codice dell'applicazione	Sì	Sì	Sì
Auto-instrumentation (agent/libreria)	Un agent intercetta automaticamente SDK/framework noti a runtime	No	Sì	Dipende
eBPF (kernel-level)	Observability nel kernel, senza toccare l'applicazione	No	Sì	No
Elastic Integration	Elastic accede a logs e metrics del LLM Provider	No	No	Solo Token (Metrics)





Grazie per l'attenzione.

Domande?

