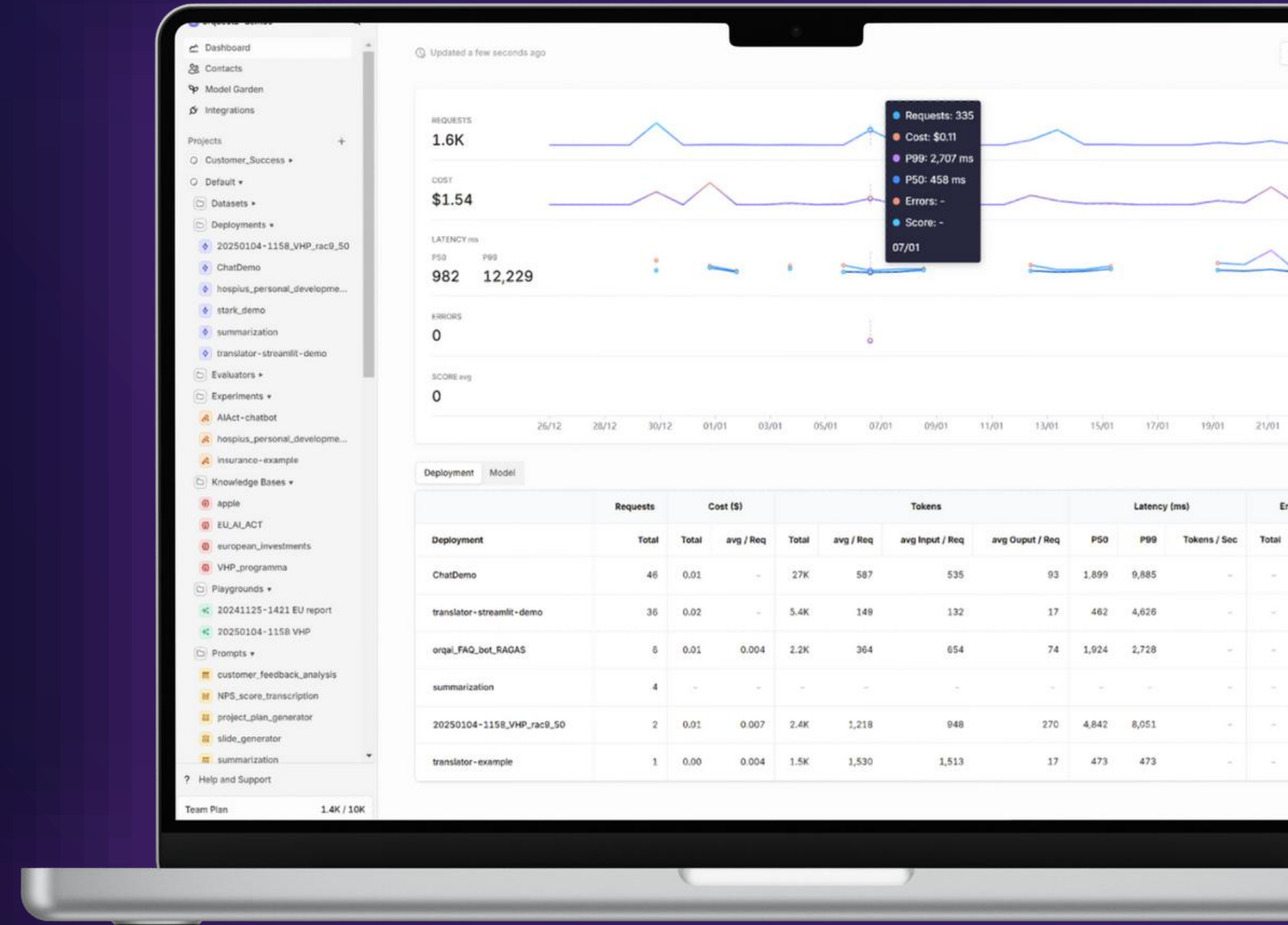
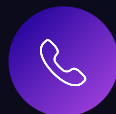




GenAI Lifecycle Management for AI product teams



orq.ai



+31 6 48 01 55 37



sohrab@orq.ai



18 years of experience building and operating engineering organizations
Founder, COO, and CTO roles within B2B SaaS

Education

Executive MBA – Erasmus University

BSc. Computer Science – Utrecht University

Born in Iran, raised in the Netherlands

Experience

COO/CTO @ Neocles

CTO Future Technologies @ Transdev

EU Tech Transformation @ Ahold-Delhaize

Associate @ McKinsey & Company

VP of Operations @ Platform161

The use of Generative AI and an AI transformations are imminent...

...which requires a wide array of mechanisms to be successful

- Vision & Roadmap
- Operating Model
- Talent Strategy
- Change Management
- Enabling Technology

PC

Digital

Cloud

Data

AI

Orq | Maturity levels of applied Gen AI

As organizations take on more ambitious goals, risk and complexity increases



Using Generative AI in product development surfaces new problems

The paradigm shift in lifecycle management requires entirely new capabilities



- AI is probabilistic and hard to control
- Lifecycle management is non-existent
- No collaboration with domain experts
- AI Talent Shortage



Teams need to get in control of the behavior and compliance of AI products



Lifecycle management for Gen AI is needed, involving domain experts

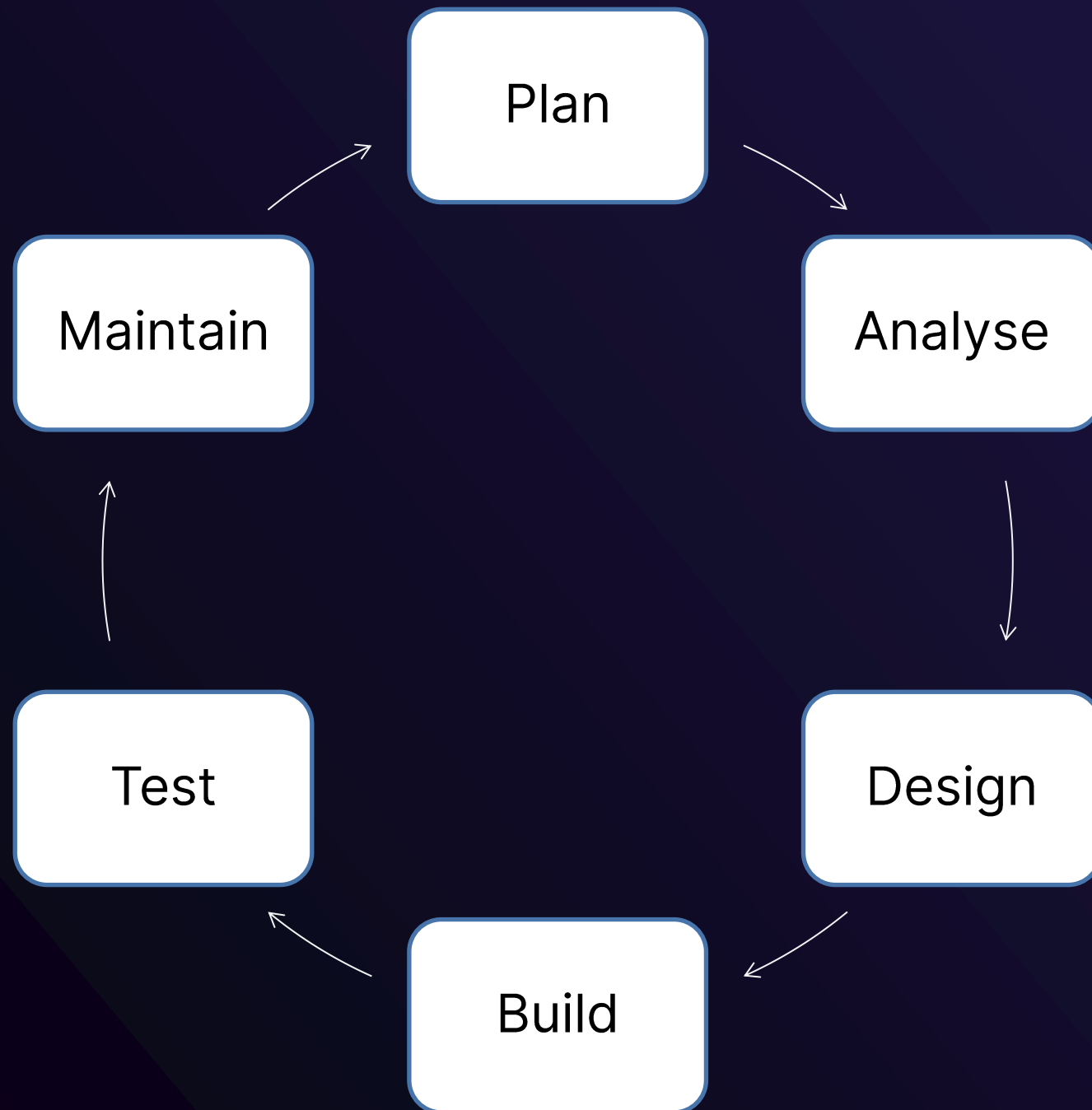
Based on ambition levels, tooling with different abstraction levels is needed

		Vertical solution	No/low code	Code-based
5	Business Model Innovation			✓
4	Services as Software (SaaS 2.0)			✓
3	Internal Process Reengineering (RPA 2.0)		✓	✓
2	Chatbot modernization	✓	?	?
1	Knowledge retrieval	✓	?	?

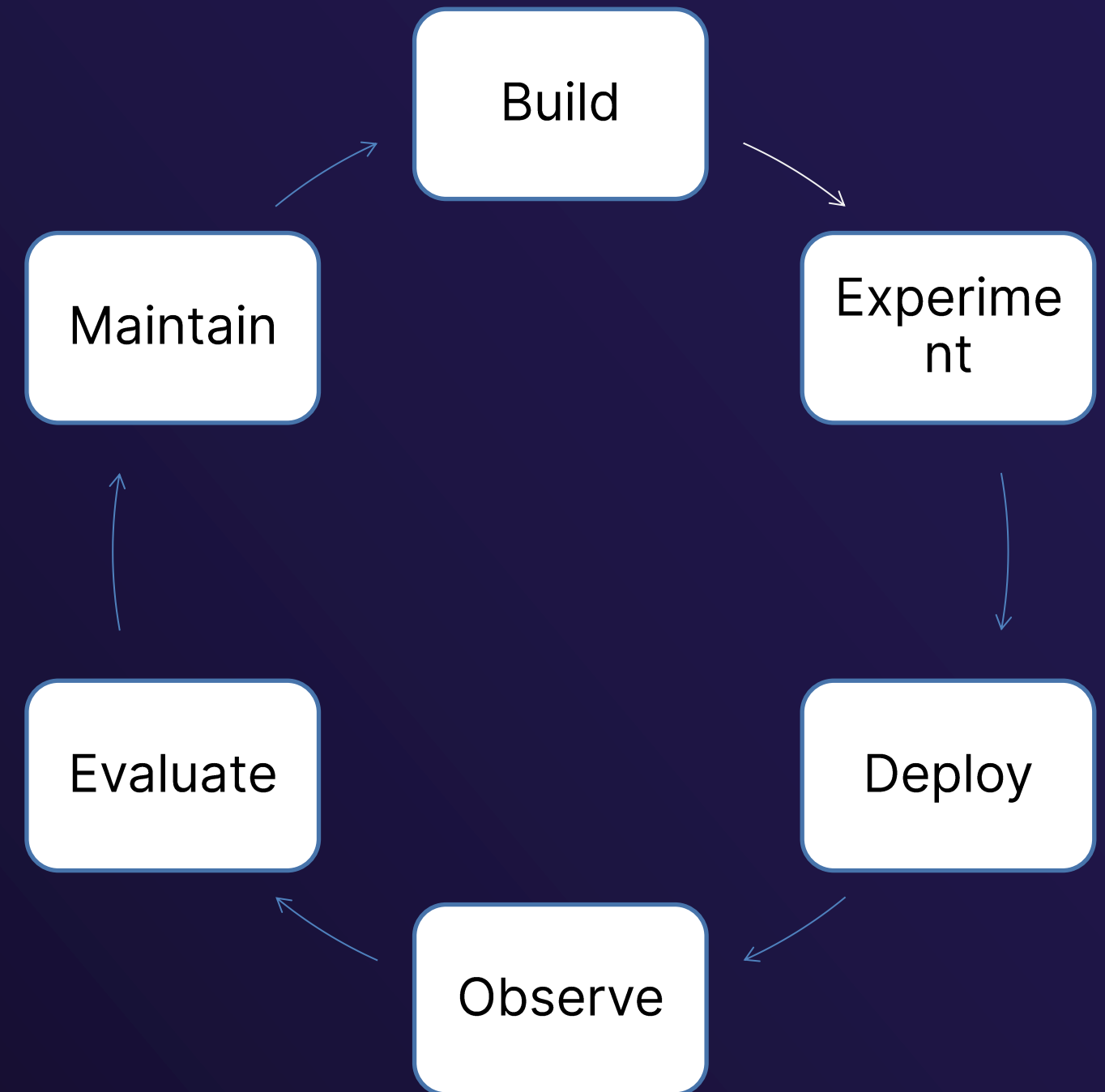
Product Lifecycle Management (PLM) is the strategic process of managing a product's journey from ideation, design, and development to production, market launch, and continuous improvement.

It ensures efficiency, compliance, and innovation by integrating data, workflows, and collaboration across teams.

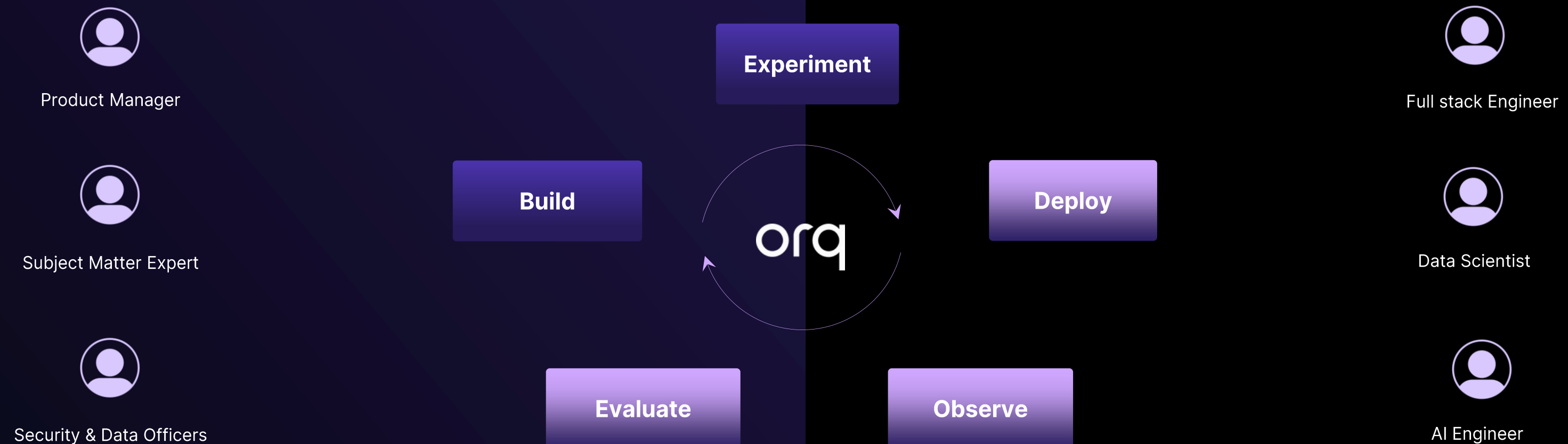
Software Lifecycle



AI Lifecycle



Continuous delivery requires rethinking for Generative AI and Agents



Product Management & Business

- **Real-time observability** with human feedback
- Empowered to collaborate, **freeing up scarce AI Engineers**
- **Autonomously** develop hypotheses and run experiments
- **Unified control tower** across products and agents

ENGINEERING & DATA

- * **SDKs & APIs**
- * **Unified across all model providers**
- * **Advanced tracing and observability**
- * **CI/CD**

cs_call_transcription
Playground
Logs
\$ 0.047
Add model
{ } Variables (2)
Blocks (3)

Parameters: vision gpt-4o max tokens 500 temperature 0.1

Prompt template: Import prompt

System

Summarize and categorize the transcription with call_id `{{call_id}}` in the following 5 bulletpoints:

1. Situation - what product/ service is the client calling about?
2. Problem - what is the root cause of the problem?
3. Sentiment - assess the emotional tone of the conversation and give a net promotor score from 0-10 (NPS).
4. solution- document immediate fixes, advice given, escalations, or follow-up commitments made during the call.
5. Summarize the outcome of the call, including whether the customer's issue was resolved to their satisfaction, any unresolved concerns, and any feedback or suggestions provided by the customer.

The use of a system message is optional; it can improve response quality but is not required for the model to function.

User

transcription: `{{transcriptions}}`

Add message

Knowledge Bases

Tools

System

Summarize and categorize the transcription with call_id 101 in the following 5 bulletpoints:

1. Situation - what product/ service is the client calling about?
2. Problem - what is the root cause of the problem?
3. Sentiment - assess the emotional tone of the conversation and give a net promotor score from 0-10 (NPS).
4. solution- document immediate fixes, advice given, escalations, or follow-up commitments made during the call.
5. Summarize the outcome of the call, including whether the customer's issue was resolved to their satisfaction, any unresolved concerns, and any feedback or suggestions provided by the customer.

User

transcription: CSR: Good afternoon, thank you for calling Dutch- Telecom. My name is CSR. How may I assist you today?

Client: Hi, I've been having issues with my internet connection. It's been intermittent for the past few days.

CSR: I'm sorry to hear that you're experiencing issues with your internet connection. I'll do my best to help you resolve this. Can you please tell me more about the problem?

Client: Sure, the internet just drops out randomly. It's been happening for a few days now. I've tried resetting the router but it doesn't seem to help.

CSR: I see, thank you for the information. Let's try to troubleshoot this together. Can you please tell me the model of your router?

Client: Yes, it's a Dutch- Telecom DT-200.

CSR: Thank you. Let's first check if there's any known issue in your area. Can you please provide me with your postcode?

Client: It's 1234 AB.

Assistant

1. Situation: The client is calling about issues with their internet connection provided by Dutch- Telecom.
2. Problem: The root cause of the problem is not immediately clear, but the client is experiencing intermittent internet connectivity, which persists despite resetting the router.
3. Sentiment: The emotional tone of the conversation is calm and cooperative. The client is seeking assistance and the CSR is responsive and helpful. Net Promoter Score (NPS): 7.
4. Solution: The CSR checks for known issues in the client's area and finds none. They then guide the client through troubleshooting steps, specifically changing the channel on the router, to potentially resolve the issue.
5. Outcome: The call ends with the CSR providing step-by-step guidance to change the router channel, which may resolve the issue. The customer's issue is not immediately resolved during the call, but they are given a potential solution to try. There are no unresolved concerns or additional feedback from the customer at this time.

Add message
Clear chat
Generate
\$ 0.00556
6,540

American Airlines Experi... Report Logs Experiment Run #1 + New run

Status Completed Cost \$0.073 Runtime 00:01:07 Add Row 🔍 📄

Cost				Latency				AA - JSON Schema Evaluator				American Airlines Friendliness eval				Cosine Similarity	
A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D	A	B
0.0001	0.0001	0.0016	0.0016	1,773	2,040	3,160	3,054	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	0.89	
0.0002	0.0002	0.0018	0.0018	1,693	1,786	2,182	2,968	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	0.89	
0.0001	0.0001	0.0015	0.0015	1,444	1,497	2,898	3,252	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	0.96	
0.0001	0.0001	0.0015	0.0015	1,184	1,345	2,221	2,253	PASSED	PASSED	PASSED	PASSED	FAILED	FAILED	PASSED	PASSED	0.87	
0.0001	0.0001	0.0016	0.0016	2,256	2,065	3,101	2,804	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	0.89	
0.0002	0.0002	0.0025	0.0025	3,357	6,078	4,588	4,435	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	0.93	
0.0001	0.0002	0.0018	0.0019	1,816	2,195	2,904	3,220	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	0.88	
0.0002	0.0002	0.0019	0.0020	1,788	2,057	3,088	3,592	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED		
0.0002	0.0002	0.0019	0.0020	2,070	2,146	2,734	3,446	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED		
0.0002	0.0002	0.0021	0.0021	2,860	2,453	3,205	3,269	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED		
0.0001	0.0001	0.0018	0.0017	1,384	1,507	3,700	3,360	PASSED	PASSED	PASSED	PASSED	FAILED	FAILED	PASSED	PASSED		
0.0001	0.0001	0.0018	0.0016	1,361	5,191	5,184	3,855	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED		
0.0001	0.0001	0.0020	0.0020	4,948	2,104	4,585	4,434	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED		
0.0001	0.0001	0.0017	0.0016	1,230	1,352	2,740	2,195	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED		
0.0002	0.0002	0.0019	0.0019	1,828	1,573	2,927	2,864	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED		
0.0001	0.0001	0.0016	0.0016	1,564	1,775	2,562	3,388	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED		
0.0001	0.0001	0.0020	0.0016	3,893	1,401	5,952	3,989	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED		
0.0001	0.0001	0.0012	0.0012	2,319	2,292	2,325	2,409	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	FAILED	FAILED		
0.0002	0.0002	0.0022	0.0021	3,156	2,796	3,521	4,229	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED	PASSED		

Create Evaluator

GET API Call
Evaluate the quality of generated text using HTTP requests

{ } JSON Schema
Validate API responses or outputs against a JSON schema

LLM-as-a-judge
Use LLM's to evaluate the quality of generated text

Python
Evaluate the quality of generated text using Python code

Typescript COMING SOON
Evaluate the quality of generated text using Typescript code

Library LLM Ragas Function

Age Appropriate
Assesses if text is appropriate for specific age groups based on language, content, tone, and cultural...

LLM-as-a-judge +

BERT Score
Computes the similarity of two sentences as a sum of cosine similarities between their tokens...

</> Function +

BLEU Score
Compares the machine-generated text to reference translations, focusing on the presence of n-grams (sequence...

</> Function +

Bot Detection
Detects whether a text is AI-generated by looking for common phrases and patterns.

LLM-as-a-judge +

Contains
Check if the text contains a specific string

</> Function +

Contains All
Check if the text contains all the specific keywords

</> Function +

Contains Any
Contains Email
Contains None

Cancel Next

The screenshot displays the Orq Deploy interface for configuring a chatbot deployment. The main window is titled "gpt-4o deployment" and includes tabs for Variant, Routing, Settings, Analytics, and Logs. The interface is divided into several sections:

- Primary Model:** Configured with "gpt-4o-mini", Max Tokens 256, and Temperature 0.2. Parameters include Top P 0, Frequency Penalty 0, Presence Penalty 0, and Retries 1. Response Format is set to "None".
- Fallback Model:** Configured with "gpt-4o", Max Tokens 256, and Temperature 1. Parameters include Top P 0.7, Frequency Penalty 0, and Presence Penalty 0. Response Format is set to "None".
- Knowledge Bases (1):** A single knowledge base named "EU_AI_ACT" is listed.
- Tools:** A section for adding tools, currently empty.
- Variables (1):** A section for adding variables, currently empty.
- Notes:** A section for adding notes, currently empty.
- Preview:** A section for previewing the deployment configuration.

The **Prompt Template** section shows a system message: "You are a legal expert for the EU AI Act that answers user's question based on the knowledge base. {{EU_AI_ACT}}". Below it, a user message field contains the placeholder ">{{userquery}}".

On the right side, a sidebar contains configuration options:

- Output Evaluators (2):** Includes "Bot Detection" (Sample Rate 100) and "Ragas Response Relevancy" (Sample Rate 100), both with toggle switches turned on.
- Output Guardrails (1):** Includes "Age Appropriate - Professionals" with a toggle switch turned on.
- Guardrail Behavior:** Includes "Retry" (Recommended) and "Fallback".
- Cache (Beta):** Includes an "Invalidate" button and a toggle for "Enable Caching" (turned off). A note states: "Note: Caching is not supported for image models at this time."
- Security:** Includes a toggle for "Output Masking" (turned off) with the note: "Prevent the response to be logged".

Timestamp	Name
25/03/2025, 14:41:30	translation_en_fr
25/03/2025, 14:41:30	translation_en_fr
25/03/2025, 14:41:29	translation_en_fr
25/03/2025, 14:41:28	translation_en_fr
25/03/2025, 14:41:27	translation_en_fr
25/03/2025, 14:41:26	translation_en_fr
25/03/2025, 14:41:25	translation_en_fr
25/03/2025, 14:41:24	translation_en_fr
25/03/2025, 14:41:22	translation_en_fr
25/03/2025, 14:41:21	translation_en_fr
25/03/2025, 14:41:20	translation_en_fr
25/03/2025, 14:41:19	translation_en_fr
25/03/2025, 14:39:45	translation_en_fr
25/03/2025, 14:39:44	translation_en_fr
25/03/2025, 14:39:42	translation_en_fr
25/03/2025, 14:39:40	translation_en_fr
25/03/2025, 14:39:38	translation_en_fr
25/03/2025, 14:39:36	translation_en_fr
25/03/2025, 14:39:36	translation_en_fr
25/03/2025, 14:39:35	translation_en_fr
21/03/2025, 09:11:49	orqai_FAQ_bot_RAGAS
20/03/2025, 15:33:49	orqai_FAQ_bot_RAGAS
20/03/2025, 13:44:00	orqai_FAQ_bot_RAGAS
20/03/2025, 10:56:54	orqai_FAQ_bot_RAGAS
19/03/2025, 23:01:22	ChatDemo

Component	Latency	Count
ChatDemo	0.0696	13,045
Evaluation_engine		2
Retrieval		8,551
Embedding	0.0	954
Rerank		7,568
Vision	0.0001	4,469
Evaluator		6,504
ChatBedrock		4,018
Chat	0.0061	4,013
ChatBedrock		6,430
Chat	0.0073	6,424
ChatBedrock		5,689
Chat	0.0071	5,684
ChatBedrock		4,928
Chat	0.0061	4,923
ChatBedrock		2,939
Chat	0.0059	2,934
ChatBedrock		3,449
Chat	0.006	3,442
ChatBedrock		4,525
Chat	0.0061	4,521
ChatBedrock		6,115
Chat	0.0059	6,112
ChatBedrock		5,612
Chat	0.0061	5,609
ChatBedrock		5,048
Chat	0.0061	5,043


```

Input
1 {
2   "context": {},
3   "inputs": {},
4   "invoke_options": {
5     "include_retrievals": true
6   },
7   "key": "ChatDemo",
8   "messages": [
9     {
10      "role": "user",
11      "content": [
12        {
13          "type": "text",
14          "text": "say hi"
15        }
16      ]
17    },
18    {
19      "content": [
20        {
21          "type": "text",
22          "text": "Hello! How can I assist you today?"
23        }
24      ]
25    }
26  ]
27 }
Assistant
The EU AI Act is a legislative proposal by the European Commission, which is part of the European Union's institutional framework. It involves contributions from various stakeholders, including policymakers, legal experts, and public consultations.
  
```

Output Guardrails

Age Appropriate - Professionals ⚙️ PASSED Yes

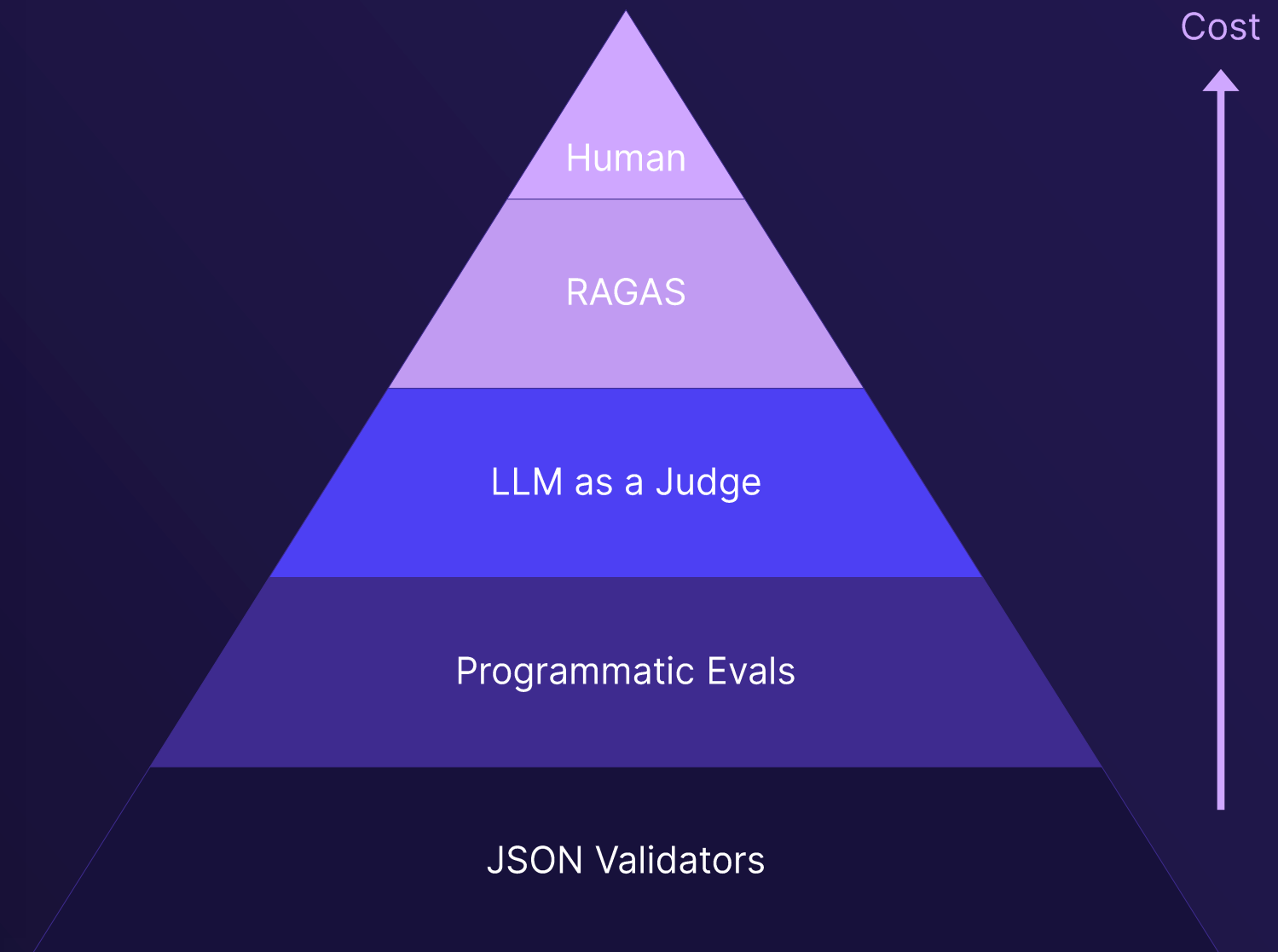
Output Evaluators

Bot Detection ⚙️ No

Ragas Response Relevancy 0.325

The screenshot displays the Orq Evaluate interface for a chat session with a `gpt-4o-mini` model. The system prompt is: "You are a legal expert for the EU AI Act that answers user's question based on the knowledge ...". The user asks: "When will the AI Act go into effect?". The assistant's response is: "The EU AI Act is set to go into effect two years from the date of its entry into force. However, certain provisions, including prohibitions related to unacceptable risks associated with AI, will apply six months after the entry into force. Additionally, obligations for providers of general purpose AI models will begin 12 months after the entry into force. Codes of practice are expected to be ready nine months after the entry into force to help providers demonstrate compliance." A correction box shows the revised response: "The EU AI Act is set to go into effect three years from the date of its entry into force. However, certain provisions, including prohibitions related to unacceptable risks associated with AI, will apply six months after the entry into force. Additionally, obligations for providers of general purpose AI models will begin 24 months after the entry into force. Codes of practice are expected to be ready nine months after the entry into force to help providers demonstrate compliance." The feedback panel on the right includes a "Rating" section with "good" and "bad" buttons, a "Defects" section with buttons for "grammatical", "spelling", "hallucination", "repetition", "inappropriate", "off_topic", "incompleteness", and "ambiguity", an "Interactions" section with buttons for "saved", "selected", "deleted", "shared", "copied", and "reported", and "Output Guardrails" and "Output Evaluators" sections. The "Ragas Response Relevancy" score is 0.824.

Defense Layers

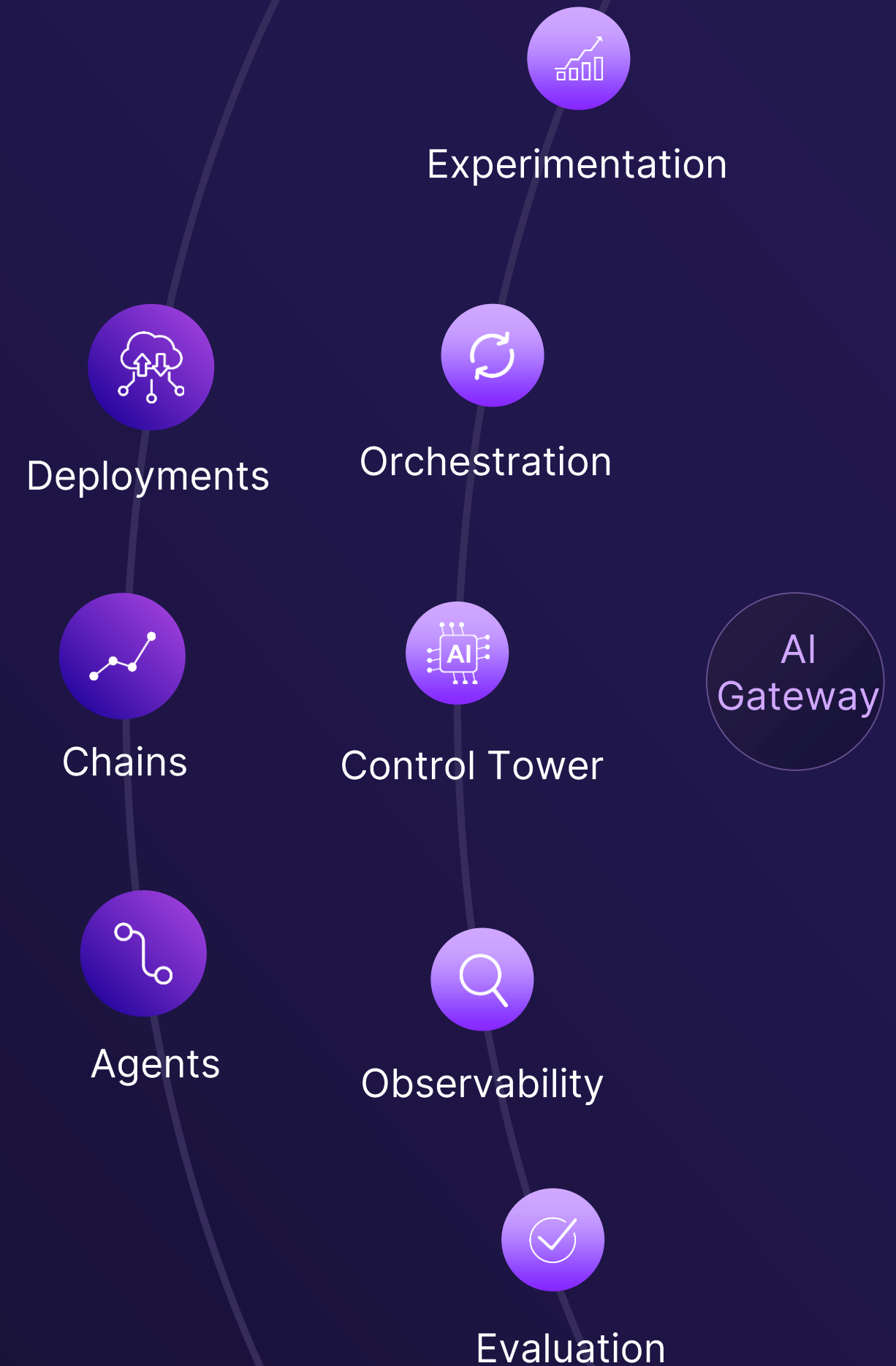


GenAI OS for AI product teams to operate at scale

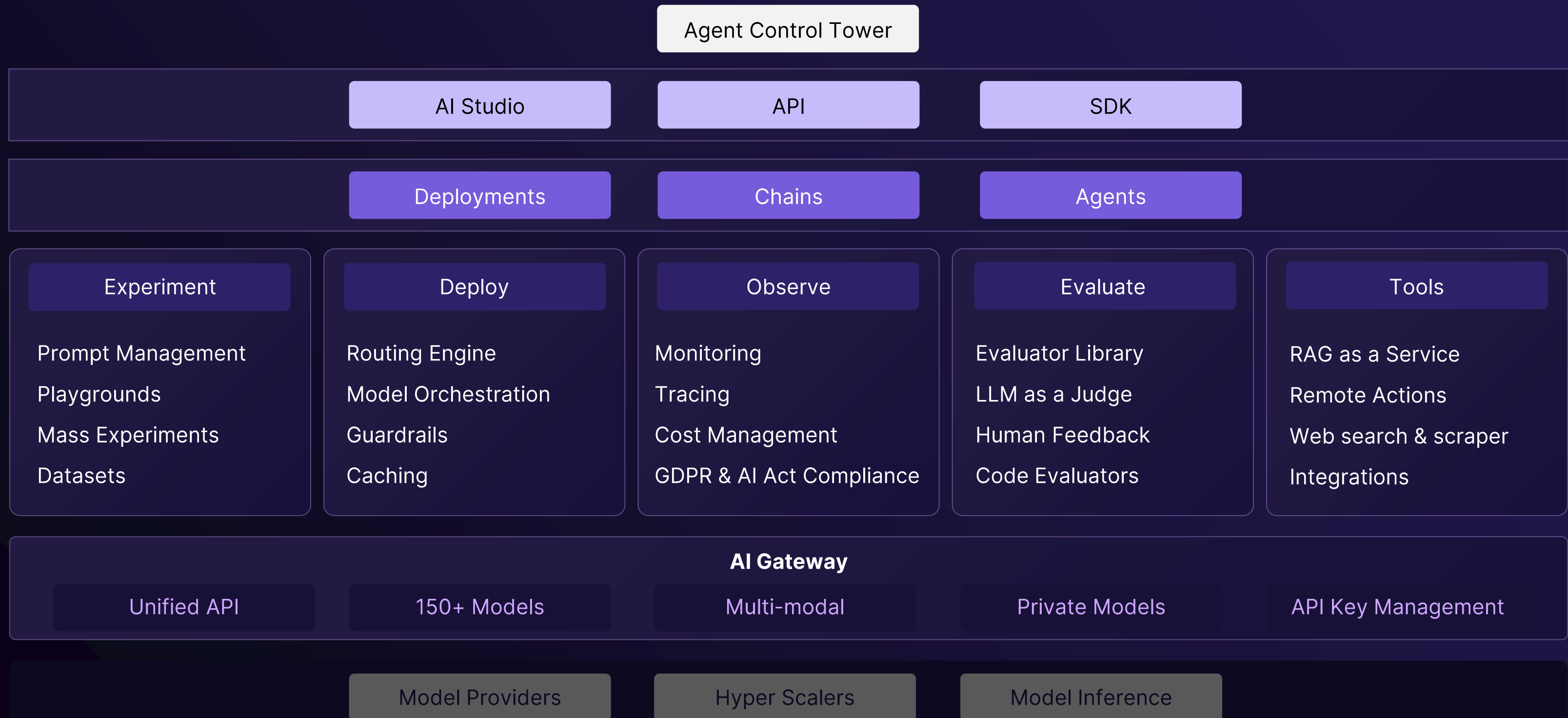
AI product development platform unifying classical software engineering and AI engineering with lifecycle management

Central control tower with unified observability, quality, and privacy controls across all products, agents and models

Streamlining **end-to-end collaboration and providing CI/CD** capabilities



Unified workflow where product teams build, deploy and monitor products and agents



Connect Me



Get Onboarded

