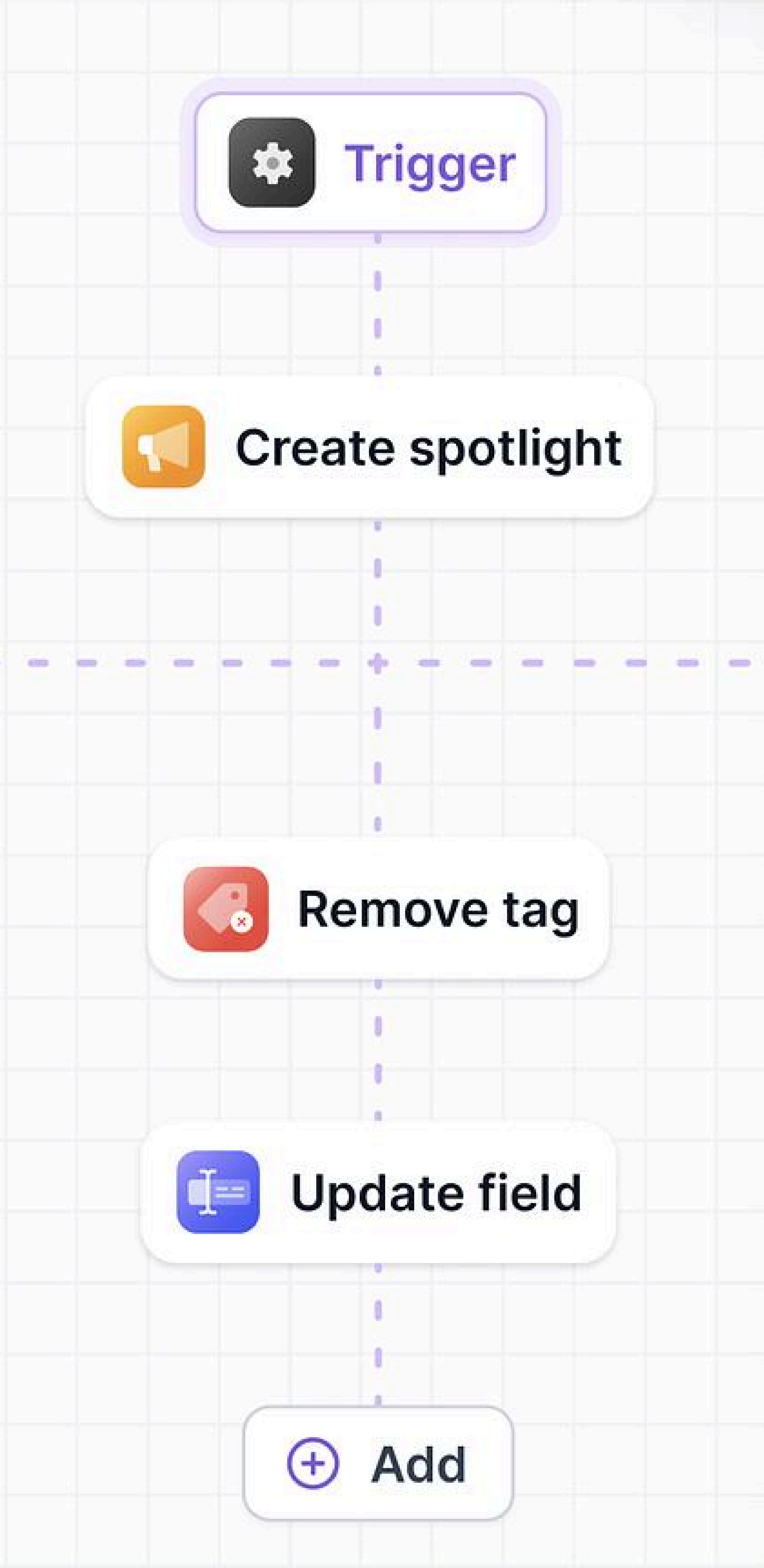
CASE STUDY

Evaluating an AI Agent's Abilities

How to assess if an AI agent can add value to your business



The world of agentic AI is now accessible and customizable.

Recent research has fueled the progress of AI agents, now capable of exploring and learning in an open-ended world (Artificial Intelligence Index 2024). Companies like Microsoft, Salesforce, and Anthropic enable Enterprise companies to create their own flows for AI agents.

The question remains: how do you assess if an AI agent can add value to your business? And how can you ensure that value grows over time?

Iterative agent evaluation and improvement is the key to ensuring the AI agent operates smoothly within your team.

Refining the Problem

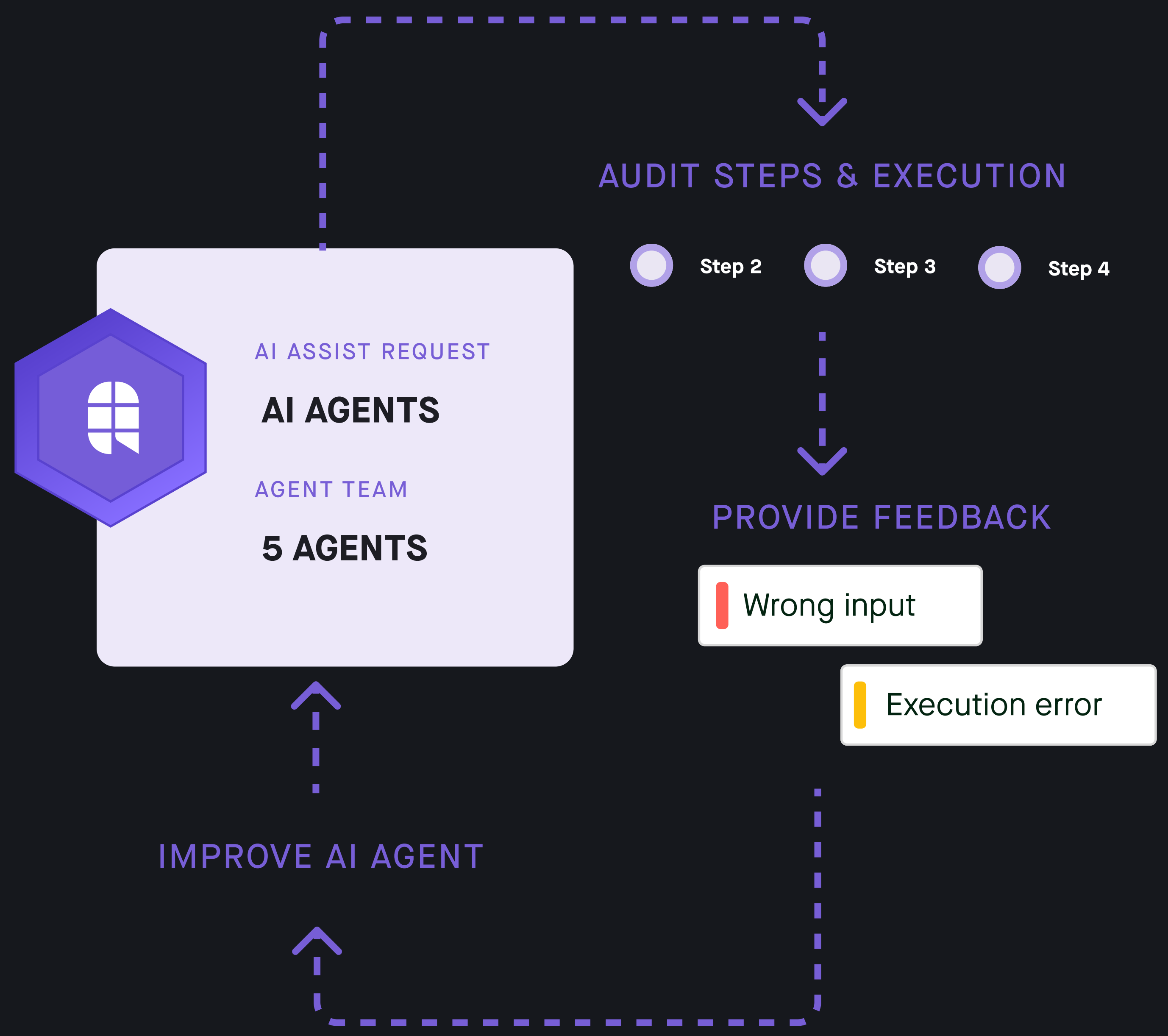
Sama has decades of experience of assessing and improving AI models for a variety of use cases. As agentic AI enters the mainstream, we wanted to learn how our teams could monitor an agent's performance, and how to leverage that feedback to improve performance over time.

As we know, the underlying Generative AI models are probabilistic and can go haywire in unexpected ways. The variety of use cases we expect them to cover can be virtually infinite, and so it is hard to be confident before deploying them that they will satisfy all performance expectations.

Additionally, feedback on model behaviour often focuses on the output only. For example, it is common to show the user two potential results and ask them to select their preferred option out of two potential model outputs.

For a user to analyze if the result matches their needs can be quite intuitive, more so than having to look at every step in the model reasoning in detail. However some research has shown how providing feedback on the process steps could help improve model performance beyond what outcome-only supervision can achieve.

But allowing users to dig into the inner workings of an AI agent is not simple.

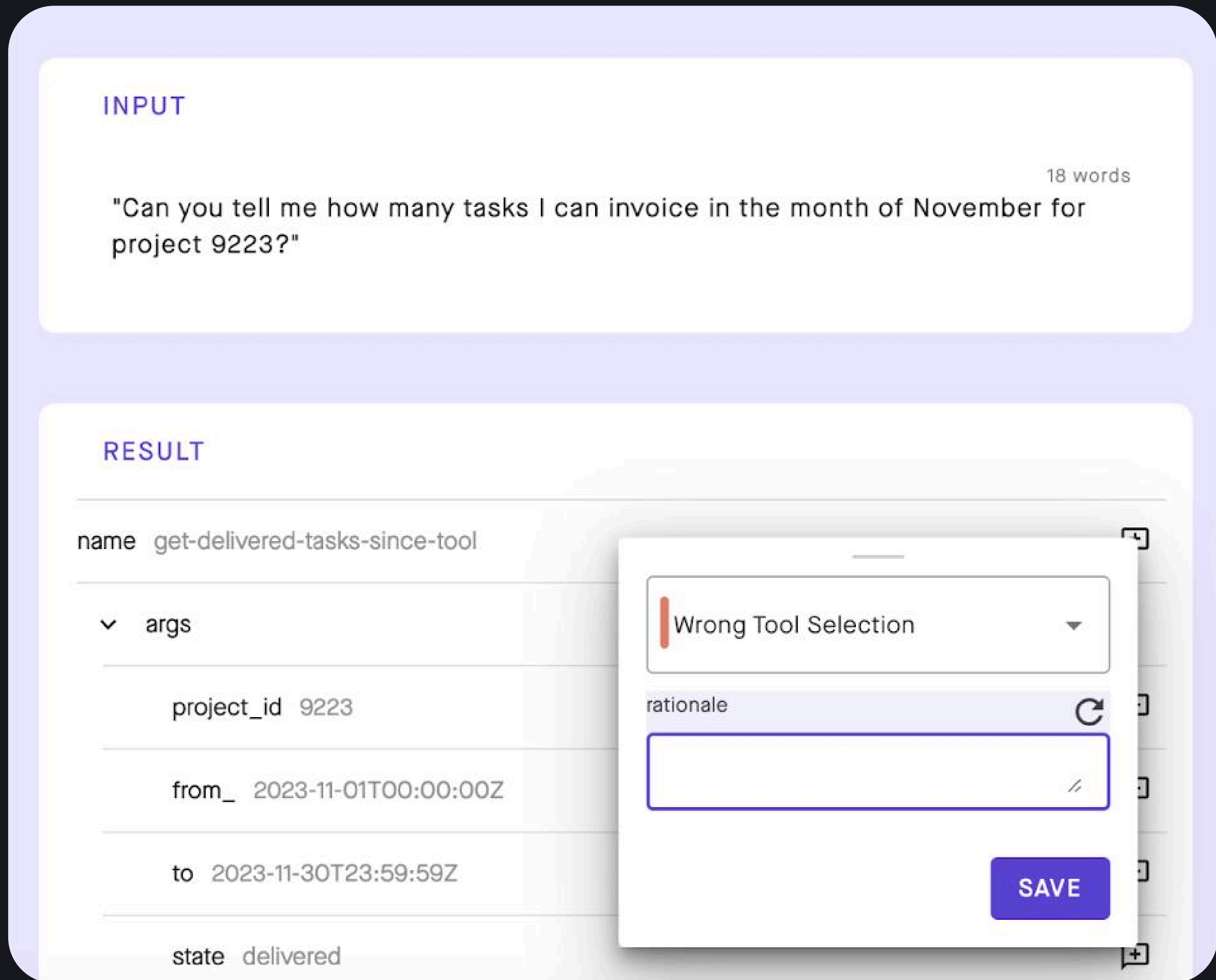


Study Overview

We wanted to understand the types of feedback that would be helpful to both understand and improve an AI agent's performance over time. In the development of AI Agent frameworks, we find that this critical piece is often overlooked.

One of the key challenges with assessing AI agents is the opaqueness of the actions taken by the agent. Even in the most popular agentic frameworks, the traces of steps taken by agents are cryptic and full of superfluous information, making them very hard to interpret.

We built a framework to load the detailed logs of an AI agent's behavior and serve them to users who have the required knowledge to supervise the agent's actions. Note that for many cases, it will not be developers, skilled at reading massive amounts of system logs, but rather domain experts who may be less technically savvy.



The types of requests that could be made to the agent include:

- Find specific tasks based on search criteria.
- Perform certain actions on tasks such as delivering or rejecting them.
- Update task information.
- A combination of the options above.

Methodology & Tooling

For our case study, we chose an internal problem we are familiar with: task management, where a "task" is the unit worked on by the annotation team.

Managing thousands of tasks as they move through our workflow is a very time consuming process. With an AI agent that has access to the Sama API, we can automate many of the manual steps involved.

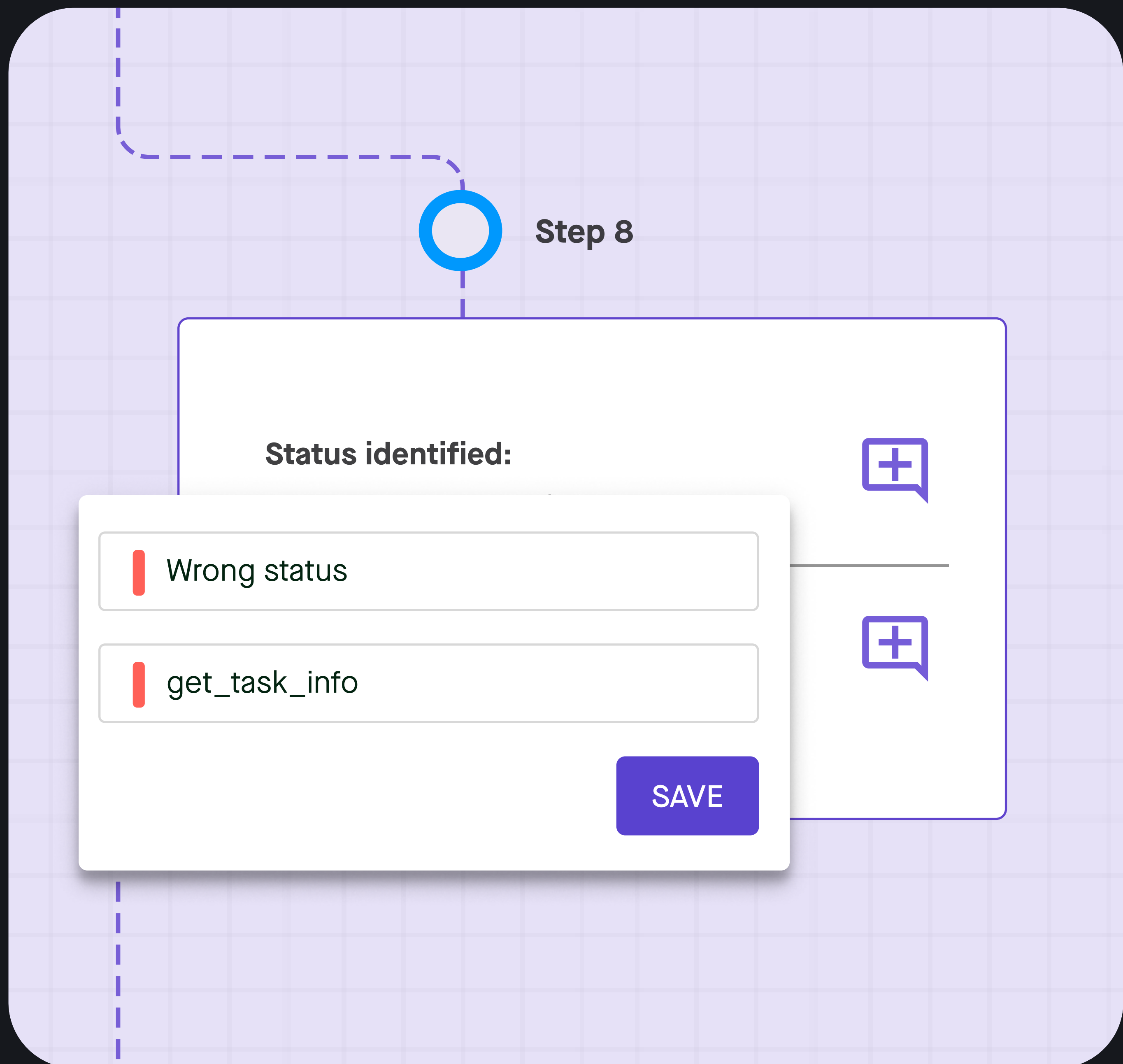
For our use case, we built our agent using LangTrace, so our framework focuses on ingesting that framework's logs. However it can easily be extended to cover more of the popular AI agent frameworks.

As part of our tooling, we made the AI agent logs interpretable by a non-technical user, and we designed a structured feedback format which could be re-ingested by the agent in order to improve its performance going forward.

For our case study, we simply inserted this structured feedback back into the prompt. That way, for future queries, the model can see examples of queries it has failed at in the past and correct itself.

As the amount of feedback grows, you could imagine setting up a RAG system to fetch previous relevant feedback or using the feedback to retrain the generative model underpinning the agent.

Using this approach, we were able to extend the capabilities of the AI agent responsible for executing task management operations.



Example: Task Invoicing

We invoice customers based on tasks which they *Acknowledge*, meaning that they confirm it has passed the project's quality Service Level Agreement.

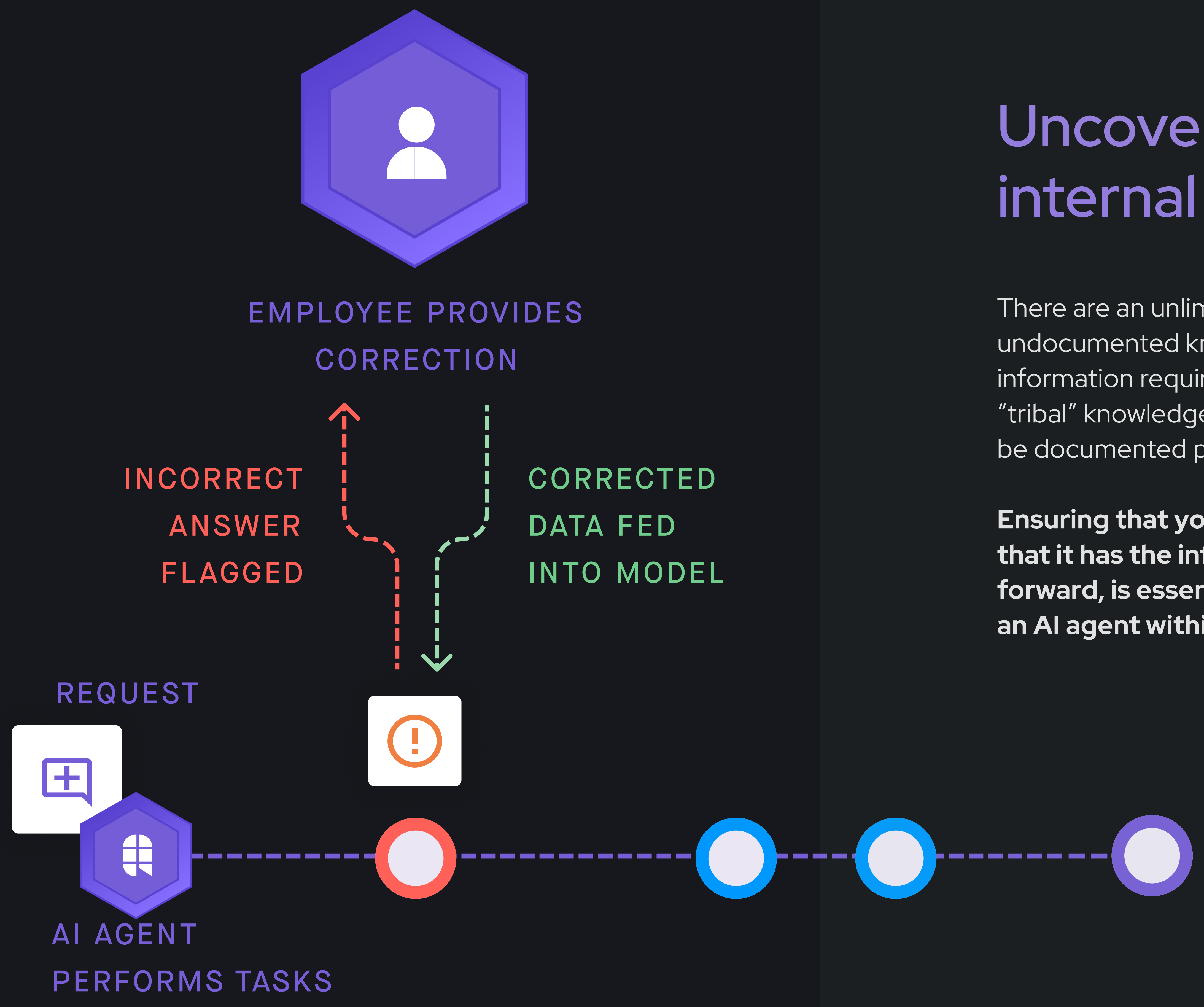
The simplified task lifecycle is that a task is...

1. Created
2. Worked on
3. Reviewed
4. Delivered to the customer
5. Acknowledged as accepted and correct

When we asked the AI agent questions regarding invoicing—for example, "How many tasks can be invoiced for last month's work?"—it would often get confused. The invoicing relationship to task state is not very explicit in the API documentation, and the agent would sometimes query for tasks in the *Delivered* state instead of the *Acknowledged* one.

Our annotators provided feedback on invoicing related queries. We then fed this data back into the model as a simple form of retraining.

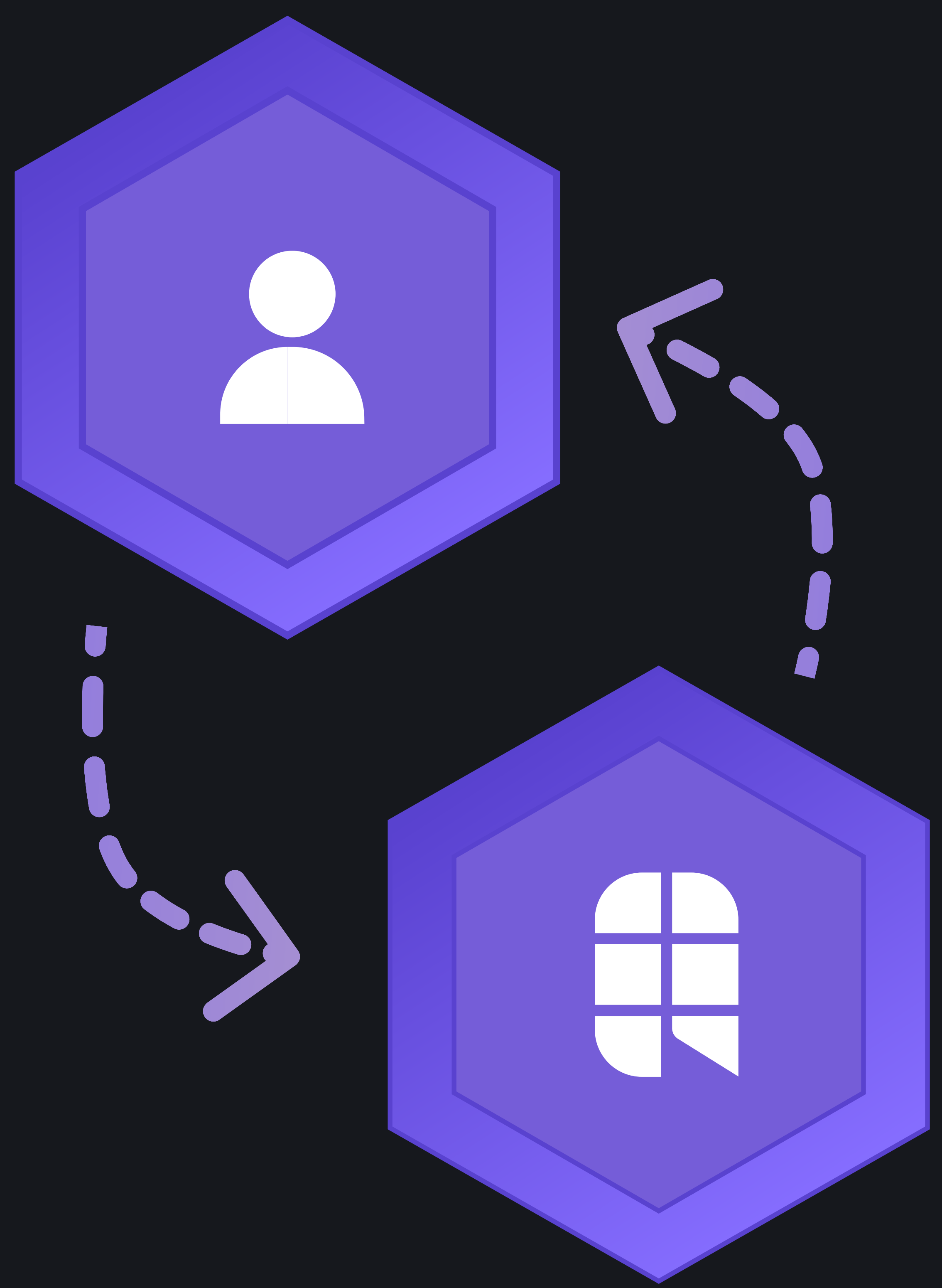
After ingesting the new data, the agent was reliably able to correctly answer queries.



Uncovering undocumented internal knowledge

There are an unlimited number of corner cases which require undocumented knowledge to address properly. A lot of the information required to perform tasks properly live in a company's "tribal" knowledge, and it is unreasonable to expect that everything be documented properly from the get go.

Ensuring that your AI agent gets feedback on its mistakes so that it has the information to tackle these tricky cases going forward, is essential to a successful long-term deployment of an AI agent within your company.



The Solution

Investing in the communication loop between AI agents and domain experts is the key to unlocking value from the technology. Employees might spend less time executing tasks, but they will remain critical to guide and improve the AI agents actually executing the mundane tasks.

Ensuring that your domain experts can transparently understand the actions taken by AI agents and that the AI agents can ingest feedback to improve the system's handling of certain requests will help guarantee an efficient and high performing hybrid team.

In other words, communication and collaboration at your company must consider the AI agents you deploy. There needs to be open and transparent dialog between humans and agents on the team.

Learn more about AI Agent evaluation at Sama

CONTACT US NOW

www.sama.com/ai-agent-evaluation

How Sama Improves AI Agents

At Sama, we can help scale your team of domain experts with proficient data annotators to ensure you are reviewing enough of your agent's work and covering all the use cases you need.

Our teams are dedicated to each workflow and work closely with clients on a weekly basis to ensure they continuously meet project needs. This is key to ensure AI agent behavior is monitored with a thorough understanding of your business context.

Our business is to help you build enterprise AI—responsibly. We help get models into production 3x faster through a scalable platform and an in-house, never-crowdsourced workforce of over 5,000 data experts.

sama

Certified B Corporation