Microsoft Foundry

From AI experiments to Enterprise-grade AI Agents

OCT 10, 2026

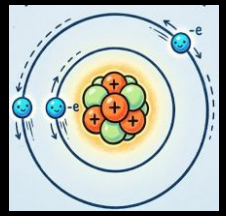


Sarang Bhrame
Principal Architect · Tech Leader

About Me

Sarang Brahme

linkedin.com/in/sbrahme





👋 Almost 2 Decades in the Tech Trenches

A Principal AI Architect, Technical Product Owner/Leader who turns big ideas into real, working AI roadmaps.

👔 C-Suite and VC Translator

I speak fluent "Business Value," "Data Science," and "Software Engineer"—and help them all get along.

🤖 Current Obsession

Agentic AI, MLOps, building/advising startups. If it makes life easier intelligently, I'm building it.

☁️ Tech Stack of Choice

Azure at the moment, but also - Databricks, Fabric, GCP, Open Source, NetApp

🎤 Speaker & Practitioner

I do talk about AI on stages; and I actually build the tech when I get off stage.

🗣️ Warning

Once started, I can't stop talking about agentic AI, quantum computing, heavy metal, and philosophy.

AGENDA

THE NEXT 45-60 MINUTES

00–05	05–13	13–33	33–55	55–60
Why agents stall	Foundry, agents and the Microsoft stack	Five pillars, incl. LLMOps and AIOps	Live demo: build, ship, operate	Takeaways and Q&A

Ask questions in the chat any time. I will pick them up at the end of each block.

THE PROBLEM

MOST AGENTS NEVER LEAVE THE NOTEBOOK

40%+

of agentic AI projects will be cancelled by the end of 2027, Gartner forecasts.

THE THREE REASONS GIVEN

- Escalating cost
- Unclear business value
- Inadequate risk controls

Notice what is missing: **"the model wasn't smart enough."** All three are platform and architecture decisions.

Source: Gartner press release, 25 June 2025

FOUNDATIONS

MICROSOFT FOUNDRY AT A GLANCE

Distribute	Teams and Microsoft 365 Copilot · Agent 365 inventory · A2A and MCP endpoints
Govern	Foundry Control Plane · Entra Agent ID · guardrails · evals · tracing · AI gateway
Run	**Foundry Agent Service: prompt agents, hosted agents, routines, memory, Voice Live**
Connect	Foundry IQ knowledge bases · Toolboxes (one MCP endpoint) · skills · tool search
Models	11,000+ models · Model Router · GPT-5.6, Claude, MAI, open models · fine-tuning

Build from anywhere: Foundry portal · Foundry Toolkit for VS Code · azd ai agent · Microsoft Agent Framework · LangGraph, OpenAI Agents SDK, Copilot SDK, Claude Agent SDK

Azure AI Foundry

AI Agent Ecosystem

Authoring Tools

- Visual Studio
- GitHub

Knowledge

- Microsoft Fabric
- Azure Cosmos DB
- SharePoint
- Microsoft Graph
- Azure AI Search
- Microsoft Bing

Enterprise Trust

- VNet deployments
- OBO Auth
- BYO-resources
- Managed Identities

Agent catalog

Ready-made agents to kickstart your agent workforce

Agent frameworks

Client SDKs for enterprise and production agentic systems

- Agent tools
- Orchestration
- Channels

Foundry Agent Service

Deploy and manage agents with fully-managed runtime

- Multi-agent workflows
- Built-in threads
- Long-term memory
- Evaluation
- Tracing & Monitoring
- Governance + Safety

Foundry Models

Azure Direct Models

OpenAI, Llama, Grok, Flux, Mistral

Azure Ecosystem Models

Phi, Cohere, Paige, Stability, Hugging Face, Industry Models

Foundry Labs

- Muse
- Magentic-One
- Aurora
- OmniParser

Open Ecosystem

Protocols

A2A, MCP, OpenAPI

Connectors

Google VertexAI, Amazon Bedrock, CrewAI, LangChain, ElasticSearch, Pinecone, Logic Apps, Azure Functions

FOUNDATIONS

WHAT MAKES IT AN AGENT, NOT A CHATBOT

Instructions
The job, the boundaries, and when to stop and escalate

Model
Reasoning, routed by difficulty and cost, not by habit

Knowledge
Grounded, cited, permission-aware enterprise context

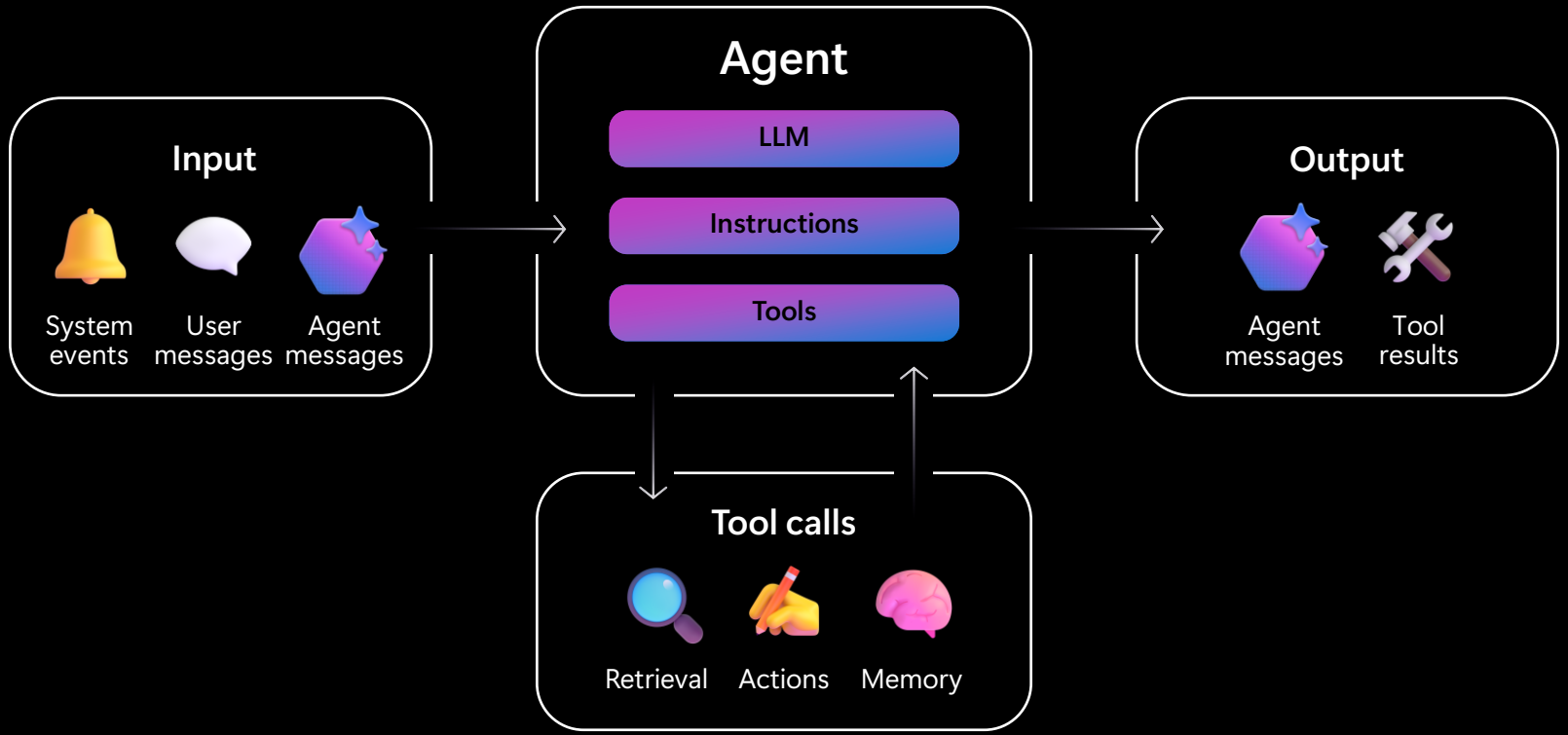
Tools
Actions on real systems through MCP and OpenAPI

Memory
Session, user and procedural: what happened and how to do the work

Identity
Who the agent is, and exactly what it is allowed to touch

THE LOOP plan → call a tool → observe → decide → … → done or escalate

Agent Flow


Input
System events
User messages
Agent messages
Agent
LLM
Instructions
Tools
Output
Agent messages
Tool results
Tool calls
Retrieval
Actions
Memory

FOUNDATIONS

THE MICROSOFT ENTERPRISE AI STACK

Where people work	Microsoft 365 Copilot · Teams · Agent Builder
Low-code build	Copilot Studio: makers own business-process agents
Pro-code build	Microsoft Foundry: custom orchestration, model control, evals
Context	Microsoft IQ: Work IQ · Fabric IQ · **Foundry IQ** · Web IQ
Trust and Azure	Entra · Purview · Defender · Agent 365 · VNet, data zones, API Management

Same models, same knowledge, same identity across every layer.

Choose the build surface by who must own the change in six months.

EXPERIMENTATION

EXPERIMENT FAST: THE INNER LOOP

Pick a model

- Catalog and playground
- Side-by-side compare
- Model Router

Prototype

- Prompt agent in the portal
- Foundry Toolkit for VS Code
- azd ai agent init, run

Add tools, data

- File search, code interpreter
- Public MCP servers
- A sample knowledge base

Vibe check

- A dozen hand-typed prompts
- Local trace viewer
- "Looks right to me"

Cheap to change and cheap to throw away. That is the point.

Also here: dev-tier fine-tuning with no hosting fees, and Foundry Local for offline work.

EXPERIMENT VS. ENTERPRISE

SAME AGENT, TWO VERY DIFFERENT SYSTEMS

Dimension	Experiment	Enterprise agent
Identity	API key in a .env file	Entra Agent ID, least privilege, Conditional Access
Knowledge	Files uploaded to a thread	Foundry IQ knowledge base: cited, permission-aware
Tools	Keys and endpoints in code	Toolbox: one MCP endpoint, managed auth, versioned
Runtime	Laptop or playground	Hosted agents: sandboxed sessions, VNet, scale
Quality	"Looks right to me"	Eval sets, Rubric, ASSERT, in CI on every change
Release	Edit it in the portal	Pull request, eval gate, then dev, QA, prod
Safety	Default content filter	Guardrail policy, ACS checkpoints, red teaming
Operations	print() and hope	Traces, SLO alerts, runbook, rollback
Cost	Whatever the bill says	Model Router, caching, cost per task, ROI
Reach	A demo link	Teams and M365 Copilot, listed in Agent 365

THE PRODUCTION PILLARS

PRODUCTION IS A PROPERTY OF THE SYSTEM AROUND THE MODEL, NOT OF THE MODEL.

Ground

Shared, cited, permission-aware knowledge

Act

A real runtime and governed tools

Trust

Evaluated, guarded, red-teamed

Govern

Identity, inventory, cost, control

Operate

LLMOps to ship, AIOps to run

PILLAR 1 · GROUND

FOUNDRY IQ: ONE KNOWLEDGE LAYER, MANY AGENTS



Knowledge bases · GA | In Copilot Studio · GA | Serverless · preview | New sources · preview | Permission sync, labels · preview | Web IQ · limited access

PILLAR 2 · ACT

RUN AGENTS LIKE SERVICES, NOT SCRIPTS

Hosted agents — GA

Each session in its own sandbox. Bring Agent Framework, LangGraph or any SDK.

Toolboxes — GA

One MCP endpoint per agent. Foundry holds tool credentials, versioned and shared.

Model Router — 28 regions

One deployment picks the model per request by your quality and cost preference.

Routines — Preview

Run any agent on a schedule: overnight triage, daily reports, reconciliations.

Memory — Preview

Session, user and procedural. Procedural: +7 to 14 pts task success on Tau-bench.

Skills, search — Preview

Versioned skills in a project catalog. Agents load only the tools a task needs.

INSIDE YOUR NETWORK VNet integration · private endpoints to storage, SQL and Key Vault

PILLAR 3 · TRUST

EVALUATE BEFORE YOU SHIP, AND AFTER

Evaluate

- **Built-in evaluators:** quality, safety, task adherence, tool-call accuracy
- **Rubric** (preview): weighted criteria generated for your agent
- **ASSERT** (open source): written policy into test scenarios
- **Any framework** (preview): tracing and evals for LangChain and custom code
- **Agent Optimizer** (coming): traces into ranked fixes

Control

- **Guided Guardrail Setup** (preview): PII, jailbreak, task adherence
- **Agent Control Specification** (open spec): checks at input, LLM, state, tool, output
- **Human approval** on any action that moves money or data
- **Automated red teaming** before every release

RULE No eval set, no deploy. Re-run on every prompt, model, tool or knowledge change.

PILLAR 4 · GOVERN

GOVERN THE FLEET, NOT JUST THE AGENT

Foundry Control Plane

For builders and AI engineers

- Fleet inventory across projects
- Traces, metrics, eval results
- Guardrail and model policies
- Defender and Purview alerts
- AI gateway brings in non-Foundry agents

Entra Agent ID

For identity teams

- Every Foundry agent gets its own identity
- Conditional Access, least privilege
- One registry across Copilot Studio and Foundry
- Revoke an agent without touching people

Agent 365

For IT and security · GA May 2026

- One inventory: M365 Copilot, Copilot Studio, Foundry, partners
- Purview data protection, Defender threat protection
- $15 per user per month, or in Microsoft 365 E7

> Plus the basics that never went away: private networking, data zones, and one named cost owner per agent.

PILLAR 5 · OPERATE · LLMOPS

LLMOPS: EVERY CHANGE SHIPS THROUGH A GATE

01 Version

Instructions, tools, knowledge base config and model pin live in Git. Agents deploy as name:version.

02 Evaluate

Every pull request runs the eval set and a Rubric against the last good baseline. Red-team scan on release.

03 Gate

The build fails if groundedness, task adherence or cost per task regress past a threshold.

04 Promote

azd and Bicep deploy the same agent to dev, QA and prod projects. Prod requires an approval.

05 Observe

Sampled production traces are evaluated continuously, not just at release time.

06 Learn

Bad traces become regression cases. Agent Optimizer proposes ranked, reviewable fixes.

TRIGGERS prompt · model or router pool · tool schema · knowledge source · guardrail

PILLAR 5 · OPERATE · AIOPS

AIOPS: RUN AGENTS AGAINST SLOS

Signal	Where it comes from
Availability, p95 latency	Azure Monitor, Application Insights
Groundedness, adherence	Evals on sampled production traces
Escalation rate	Trace attributes, human handoffs
Tool and retrieval errors	OpenTelemetry spans per call
Cost per task	Tokens per run, ROI view
Knowledge freshness	Indexer runs, source sync status
Capacity and models	Quota, PTU use, router pool changes

When an alert fires

1. Azure Monitor alert reaches on-call
2. Open the trace, find the failing span
3. Contain: roll back to the last good version, tighten a guardrail, or disable the agent
4. Fix forward: the trace becomes a test
5. Review quality, cost and ROI weekly

THE PATH

FOUR STAGES, AND THE GATE OUT OF EACH

Gate: a named business outcome and an eval set built from real cases

Experiment

Gate: grounded on Foundry IQ, own Entra Agent ID, guardrails on, every run traced

Pilot

Gate: hosted in your network, eval-gated releases, SLO alerts and a tested rollback, cost per task reported

Production

Ongoing: shared knowledge bases and Toolboxes, Control Plane and Agent 365, ROI reviewed

Fleet

REALITY CHECK

WHAT IS GA AND WHAT IS PREVIEW, TODAY

Generally available

Hosted agents	Jul 2026
Toolboxes	Jul–Aug 2026
Publish to Teams, M365 Copilot	Jun 2026
Foundry IQ knowledge bases	Jun 2026
Foundry IQ in Copilot Studio	Oct 2026
Foundry Toolkit for VS Code	Jun 2026
Agent 365	May 2026
Tracing and evaluations	Spring 2026

Preview or limited

Routines	Preview
Memory (incl. procedural)	Preview
Skills, tool search	Preview
Foundry IQ Serverless, new sources	Preview
Evals on production traces	Preview
Rubric, Guided Guardrail Setup	Preview
Web IQ in Foundry IQ	Limited
Agent Optimizer · Agent ROI	Soon · Private

Checked 9 October 2026. Previews move monthly: re-check the week you build and put production on the left column.

THE TRIAGE PROBLEM

A claim that should take days takes weeks

CAUSE 1

Intake is unstructured

Photos, a police report PDF, two repair estimates in different formats, an FNOL call, free-text notes.

CAUSE 2

Policy reading is judgment

Coverage, deductible, exclusions, endorsements. Forty pages that must be read whole, not in chunks.

CAUSE 3

No API on the system of record

In most carriers the adjuster is the integration layer — reading one screen, typing into another.

CAUSE 4

Exceptions escalate to a queue

And the queue is a person. This is where the days actually disappear, undocumented.

11 days expected vs 23.9 days actual (J.D. Power 2024) — none of it a model problem, all of it context and orchestration.

THE TRIAGE PROBLEM

From Collision to Clarity

STEP 1

Claim arrives

A form lands in the queue with photos and free text attached.

STEP 2

Adjuster opens two systems

The document in one window, policy and history in another.

STEP 3

Manual cross check

Coverage, limits, deductible, and past claims are compared by eye.

STEP 4

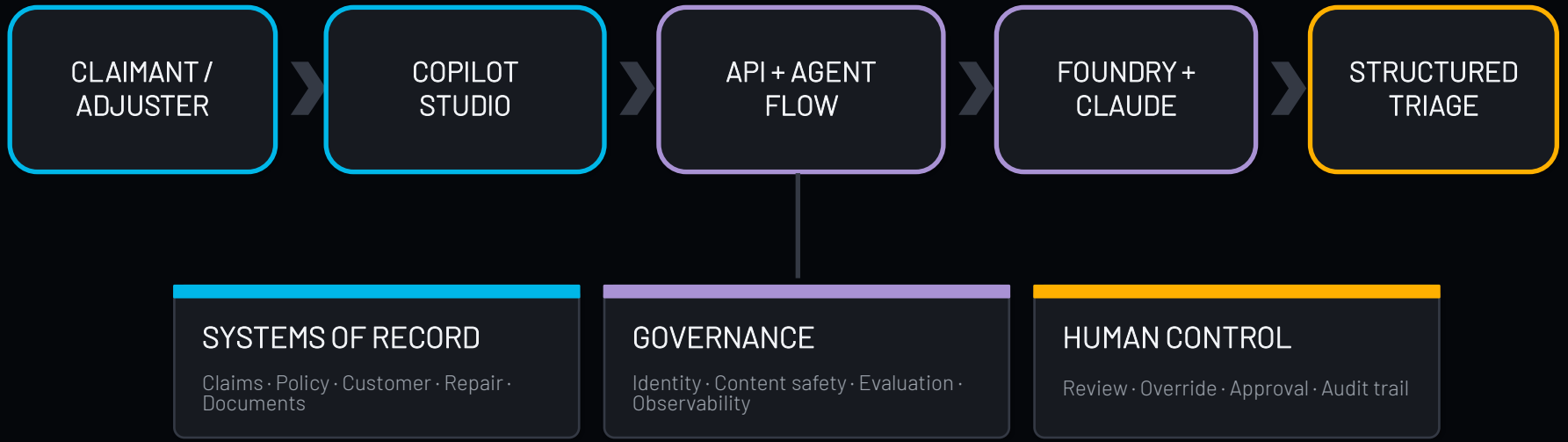
Judgement call

Approve, deny, or escalate, with the reasoning undocumented.

Every claim repeats the same manual comparison, and no two adjusters weigh it the same way.

REFERENCE PROCESS

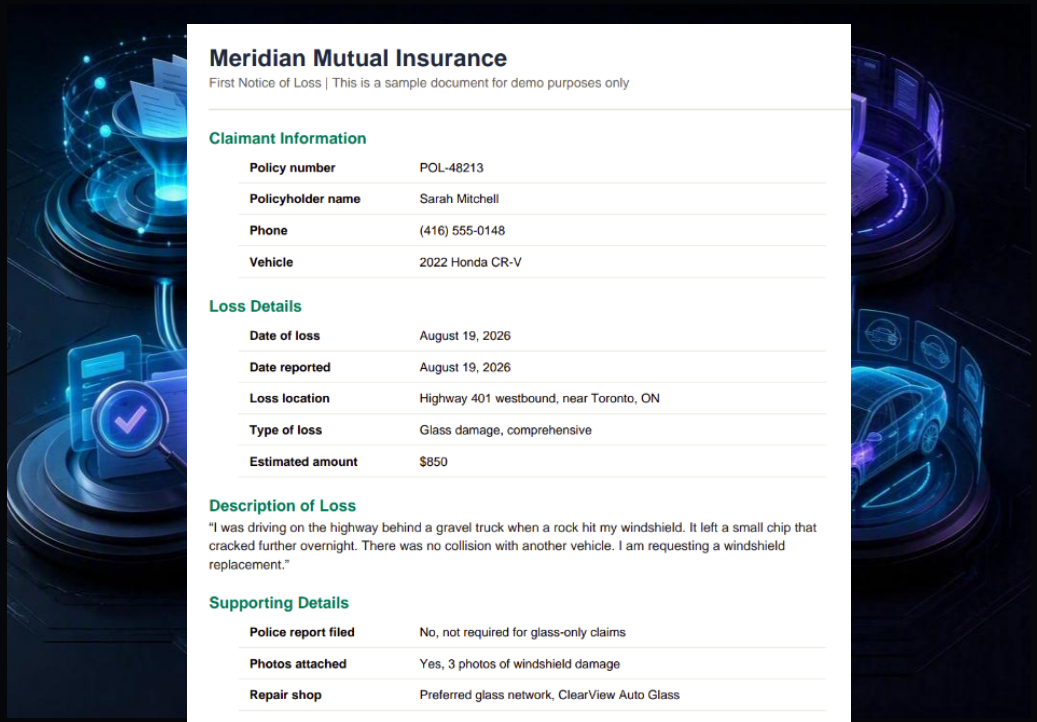
From conversation to governed action.



Enterprise APIs remain the systems of record.

Meet Claim A-1047

Incomplete, multimodal, and slightly inconsistent by design.



Meridian Mutual Insurance

First Notice of Loss | This is a sample document for demo purposes only

Claimant Information

Policy number	POL-48213
Policyholder name	Sarah Mitchell
Phone	(416) 555-0148
Vehicle	2022 Honda CR-V

Loss Details

Date of loss	August 19, 2026
Date reported	August 19, 2026
Loss location	Highway 401 westbound, near Toronto, ON
Type of loss	Glass damage, comprehensive
Estimated amount	$850

Description of Loss

"I was driving on the highway behind a gravel truck when a rock hit my windshield. It left a small chip that cracked further overnight. There was no collision with another vehicle. I am requesting a windshield replacement."

Supporting Details

Police report filed	No, not required for glass-only claims
Photos attached	Yes, 3 photos of windshield damage
Repair shop	Preferred glass network, ClearView Auto Glass

This First Notice of Loss form is a fictional sample created for a conference demo. Meridian Mutual Insurance is not a real company.

WHAT ARRIVED

- Claim submission
- Claimant Information
- Loss Details
- Description of Loss
- Policy and vehicle details
- Prior claims summary

THE COMPLICATION

The narrative describes a rear impact. The submitted photos primarily show front-left damage.

EXPLAINABILITY

A recommendation is not enough. Show the evidence.

RECOMMENDED ROUTE

Complex damage review

CONFIDENCE · MEDIUM

1. Coverage appears active for the incident date.
2. Front-left damage is visible across submitted photographs.
3. The narrative describes a rear impact.
4. Suspension work in the estimate is not clearly supported by available images.
5. A required wide-angle scene photograph is missing.

HUMAN REVIEW REQUIRED

What this workload needs from the solution?

Model selection: six reasoning demands this investigation places on the model.

01	02	03
Long-context analysis	Reasoning across multiple pieces of evidence	Handling detailed operational instructions
04	**05**	**06**
Analyzing supporting and contradicting evidence	Generating structured hypotheses	Working through ambiguous investigation scenarios

The model should be chosen based on the reasoning workload — not because one model is always best.

LIVE DEMO

BUILD, SHIP, OPERATE: SIX ACTS

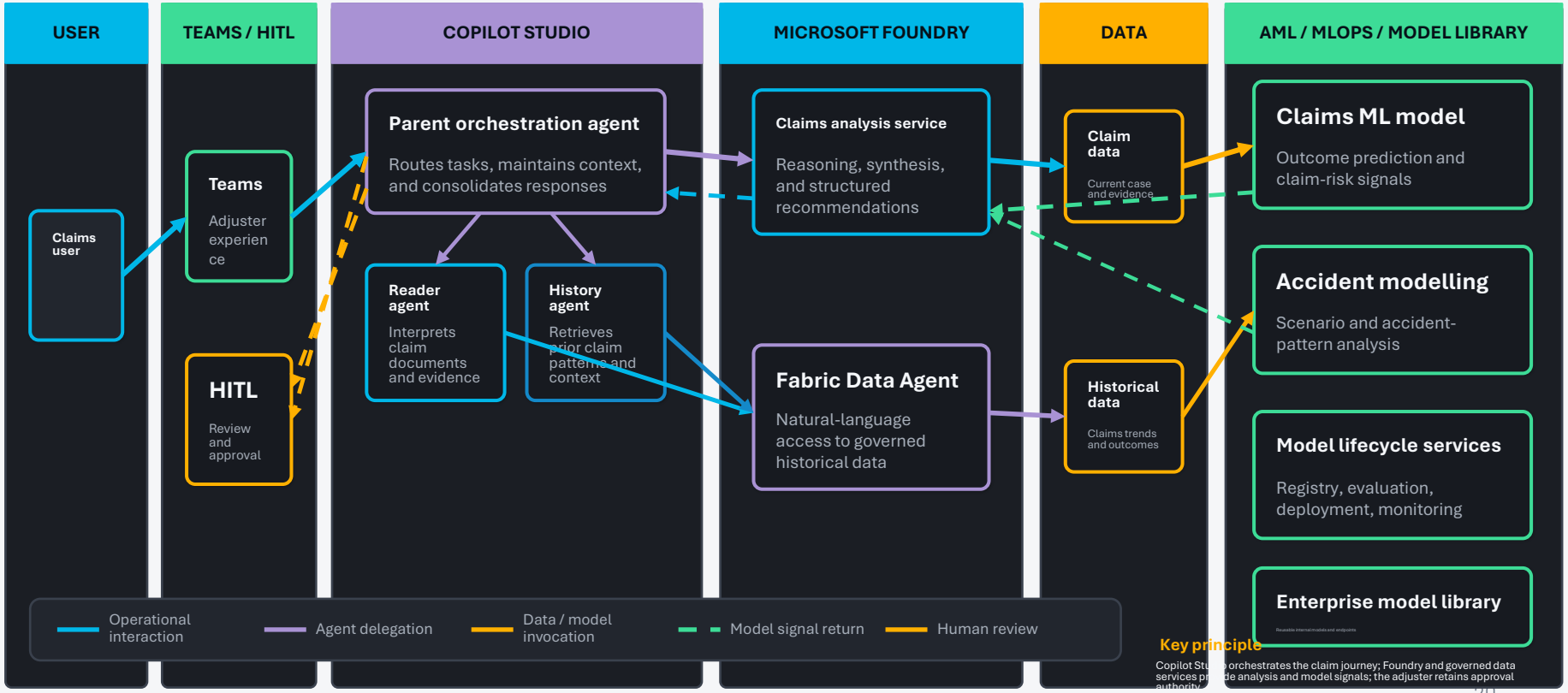
ACT 1 · 2 MIN **The experiment**	Playground agent with a policy PDF. It answers from the wrong policy version, with no citation, running as you.
ACT 2 · 5 MIN · GROUND **Foundry IQ**	claims-kb over the policy library, procedures and claims table. Minimal vs. medium effort, citations, permission trimming.
ACT 3 · 3 MIN · ACT **Productionize**	Toolbox with claims-kb and the claims API. Hosted coverage agent via azd, with its Entra Agent ID and guardrail policy.
ACT 4 · 4 MIN · LLMOPS **Ship a change**	A prompt change opens a PR. The pipeline evaluates it against the baseline, the gate blocks the regression, the fix promotes to prod.
ACT 5 · 4 MIN · AIOPS **Operate it**	SLO dashboard for latency, groundedness, escalations and cost per claim. An alert fires: trace, roll back, turn the trace into a test.
ACT 6 · 3 MIN · GOVERN **Scale the workforce**	Two more agents reuse the knowledge base and Toolbox. Intake runs as a Routine, the assistant in Teams, all three in Agent 365.

Difinity

Demo Time!



THE ENTERPRISE ARCHITECTURE, END TO END



TAKEAWAYS

Three things to take into Monday

1

Separate the knowledge from the agent.

One Foundry IQ knowledge base, many agents, permissions enforced at retrieval.

2

Production is a checklist, not a feeling.

Identity, guardrails, an eval gate, SLO alerts and a rollback, before the first user.

3

Build production on GA, pilot on preview.

Check status the week you build, and keep preview features behind an interface.

Key takeaways

What to remember about bringing Claude into Copilot Studio.

01

Not just a model swap

Claude in Copilot Studio is a step-change in what your agents can understand and deliver.

02

Use it where it counts

Document analysis, long-form summarization, and multi-step problem solving are its sweet spots.

03

Choose per task

Multi-model means picking Claude or GPT per agent or skill — balanced, not absolute.

04

Design decides value

Thoughtful prompts and well-structured knowledge sources matter more than flipping a selector.

Q&A

Please share Your Feedback

LinkedIn

linkedin.com/in/sbrahme

Slides and Playbook

https://gospinup.dev/speaking/2026-10-ai-agents-that-ship

My Work

https://www.longviewsystems.com/

Thank You!