

BUILDING ENTERPRISE-SCALE

RAG Chatbots

Using Azure AI Foundry

2026 Iowa AI Summit for Industry · CIRAS, Iowa State University · May 6, 2026

PRESENTER

Mehul Bhuva

Data & AI Platform Engineer · QCI · 22+ years enterprise experience

MS Computer Science · Georgia Tech · Azure AI Foundry · Databricks · RAG systems

 [linkedin.com/in/mehulbhuva](https://www.linkedin.com/in/mehulbhuva)

Production-Grounded

Architecture-Focused

Actionable for Teams

Build · Deploy · Govern

Mehul K Bhuva

Data & AI Platform Engineer · QCI · More Than Technology. Understanding.

- ▶ 22+ years building enterprise data & AI platforms
- ▶ MS Computer Science — Georgia Tech
- ▶ Designed & shipped Sprout-AI: production RAG serving 22,000+ global users
- ▶ Specializes in enterprise RAG, Azure AI Foundry & Databricks
- ▶ Experience across Azure AI Search, Azure OpenAI, Functions, Logic Apps, SharePoint & SQL
- ▶ Conference speaker: Iowa AI Summit · Databricks Data+AI · Enterprise AI Symposia

Enterprise RAG

Azure AI Foundry

Databricks

Vector Search

LLM Grounding

LET'S CONNECT

 linkedin.com/in/mehulbhuva

 learn.microsoft.com/azure/ai-foundry

 github.com/azure-samples

 CIRAS Iowa AI Summit · May 6, 2026

Discussion Starters

- Where are you today — pilot, early prod, or scaling?
- What is the biggest RAG challenge in your org?
- Which industry use case resonates most?

Agenda

Architecture · Implementation · Governance · Measurable Outcomes

01

Why RAG Now

Enterprise problem & the architectural shift

02

RAG Architecture

Ingestion → retrieval → generation pipeline

03

Azure AI Foundry

Platform, governance & key components

04

Build Path

Portal walkthrough — 6 practical steps

05

Production Design

Reference architecture & security model

06

Case Study

Sprout-AI: 22,000+ users & real outcomes

07

Best Practices

Benchmarks, observability & emerging roadmap

Why Retrieval-Augmented Generation?

The architectural shift driving enterprise AI adoption in 2026

✗ Traditional LLM Gaps

- Knowledge frozen at training cutoff
- Hallucinations in enterprise context
- No access to proprietary/restricted content
- Retraining costs \$1M–\$10M+ per cycle

VS

✓ RAG: The Enterprise Fix

- Real-time retrieval from your own knowledge
- Auditable, source-cited answers
- No retraining needed for knowledge updates
- Days to production — not months

~80%

Hallucination
Reduction

10–100×

Cheaper than
Retraining

Days

Time to
Production

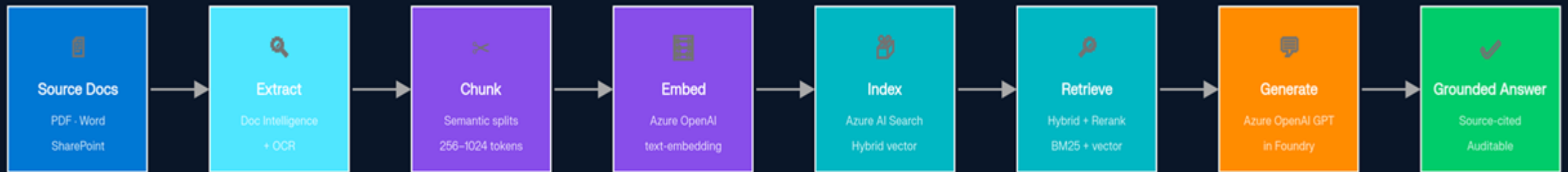
Full

Audit
Trail

RAG Architecture — Core Pipeline

From enterprise documents to grounded, cited AI answers

RAG Core Pipeline · Source Docs → Grounded Answer



Ingestion quality = accuracy

Chunking, OCR fidelity & metadata richness all determine retrieval correctness before any prompt is written.

Hybrid retrieval + reranking

Combining vector similarity, BM25, and semantic ranker achieves 95%+ relevant context delivery.

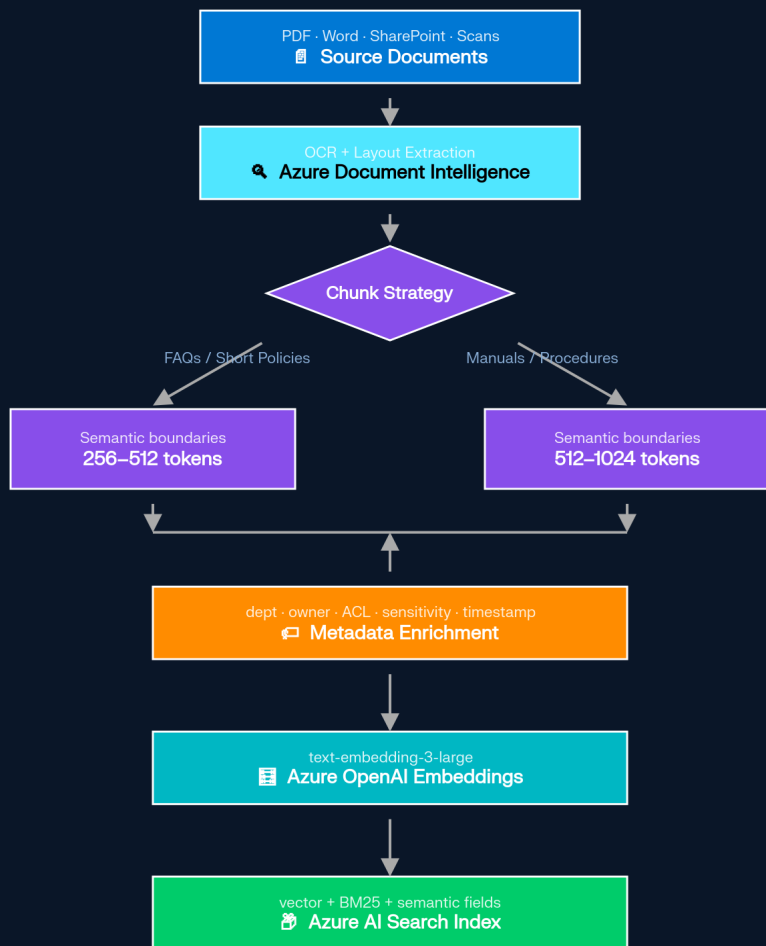
Grounding discipline

Prompt design enforces citation of retrieved chunks — the model answers from evidence, not memory.

Document Ingestion Pipeline

Enterprise content → searchable, permission-aware knowledge

Document Ingestion Pipeline · Source Docs → Searchable Index



📏 Chunk Size Strategy

256–512 tokens → FAQs & short policies
 512–1024 tokens → manuals & procedures
 Always split on semantic boundaries

🏷️ Metadata to Carry

Department · Owner · Product · Region
 Sensitivity · Classification · ACL tags
 Required for RBAC-aware retrieval

📄 Index Schema Essentials

Chunk text · Title · Source URL
 Vector field · Freshness date
 Filterable facets per field

Unified Enterprise AI Platform

Microsoft's control plane for model choice, governance, evaluation & deployment



Model Catalog

11,000+ models — OpenAI, open-source, domain-specific



Chat Playground

No-code prototyping, model comparison, prompt iteration



AI Hub & Projects

Governed workspaces, centralized compliance & RBAC



Foundry IQ / RAG

Built-in grounded experiences with access control & citations



Agent Service

Managed orchestration — multi-agent, multi-step, MCP



MCP Connectors

50+ integrations: SAP, Salesforce, Databricks, custom

Azure AI Foundry Resource Hierarchy

How central IT and delivery teams divide responsibilities — without blocking each other

Foundry Resource — Central IT

-  Network isolation & security governance
-  Managed identities & RBAC enforcement
-  Private endpoints & VNET configuration
-  Compliance policies, audit logs & BYOK



Separation
of Concerns

Projects — Dev Teams

-  Isolated dev / test / prod workspaces
-  Scoped models, agents, files & evaluations
-  Least-privilege RBAC per project
-  Dev team autonomy inside safe boundaries

 **Key Insight:** Central IT owns the guardrails (network, encryption, logging, policy). Dev teams own prompts, retrieval, evaluation & release. Neither blocks the other.

What Powers Your RAG System

Azure services & Foundry capabilities mapped to the RAG pipeline



AI Hub

Control plane for governance, policy & enterprise resource provisioning



Vector Stores

Azure AI Search + pgvector. Hybrid retrieval and semantic indexing built-in.



Playground

Rapid prototyping, model comparison & prompt engineering — no code needed



Projects

Isolated dev, test & prod workspaces with scoped RBAC per team



MCP Connectors

Standard interface for SAP, Salesforce, Databricks & custom Azure Functions



Observability

Azure Monitor + App Insights + OpenTelemetry. Track latency, tokens & drift

Building a RAG Chatbot — Step by Step

A portal-driven implementation path any engineering team can follow today

1

Create Resource & Project

Provision AI Foundry resource.
Create a project for the use case.

2

Select Model

Choose from 11,000+ models by
cost, latency & domain fit.

3

Connect Knowledge

Upload docs from Blob Storage
or SharePoint connector.

4

Configure Index

Review chunking, embedding
& vector index settings.

5

Tune & Evaluate

Set system prompt, retrieval filters,
run Playground evaluation.

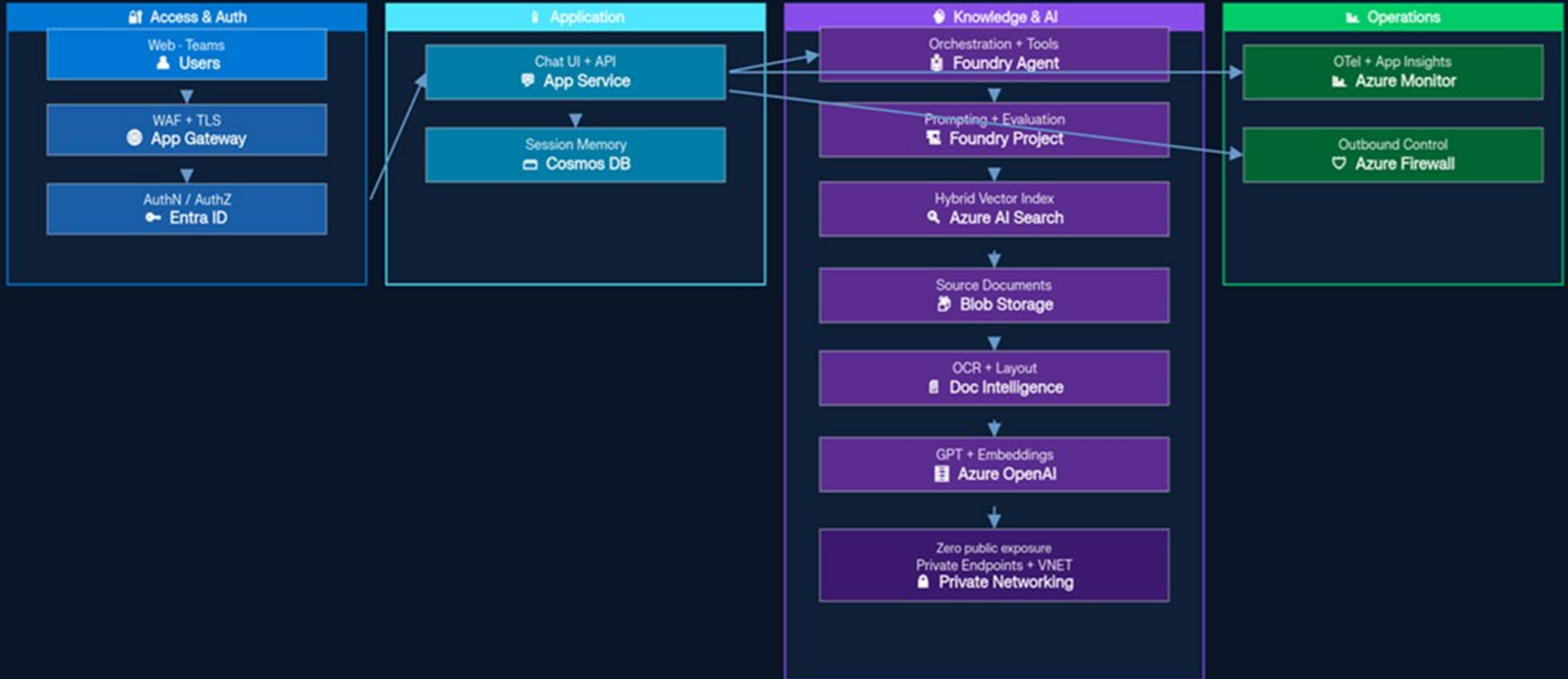
6

Publish & Integrate

Expose via SDK or REST API
into Teams, web or enterprise apps.

Production Enterprise RAG Architecture

Azure-native · Private-first · Observable — built for scale and governance



Production RAG at Global Scale

Sprout-AI — a real-world enterprise pattern serving 22,000+ users worldwide

22,000+

Global Users

95%

Retrieval Accuracy

<2 sec

Response Time

40–60%

Cost Reduction



Business Challenge

22,000+ global staff needed trusted, fast answers from fast-changing enterprise knowledge — without searching fragmented systems or opening support tickets.



Technical Pattern

Azure Document Intelligence ingestion · Adaptive semantic chunking · Hybrid retrieval + reranking · Azure OpenAI GPT in Foundry · Cosmos DB chat history · Private networking.



Business Impact

Higher answer trust · Measurable support-ticket reduction · Faster knowledge access at scale · Governance model that passed enterprise security review.

Enterprise RAG Across Industries

Same architecture pattern — domain content and governance layers adapt per vertical



Agriculture

Crop advisory tools, agronomic knowledge assistants, product guidance & field support



Healthcare

Policy & workflow retrieval with HIPAA-aware RBAC and permission-gated content



Manufacturing

Maintenance copilots, SOP assistants & troubleshooting guides over work instructions



Logistics

Route planning, customs documentation & regulatory lookup for global operations



Financial Services

Compliance bots, audit support, policy lookup & risk guidance with source citations



Public Sector

Policy Q&A, citizen-service assistants & permitting guidance grounded in regulations

Production Performance Benchmarks

Representative outcomes from enterprise RAG deployments on Azure

95%

Retrieval Accuracy

hybrid search + reranking

<2 sec

Response Time

p95 for common queries

~80%

Hallucination Cut

grounding at retrieval layer

40–60%

Cost Reduction

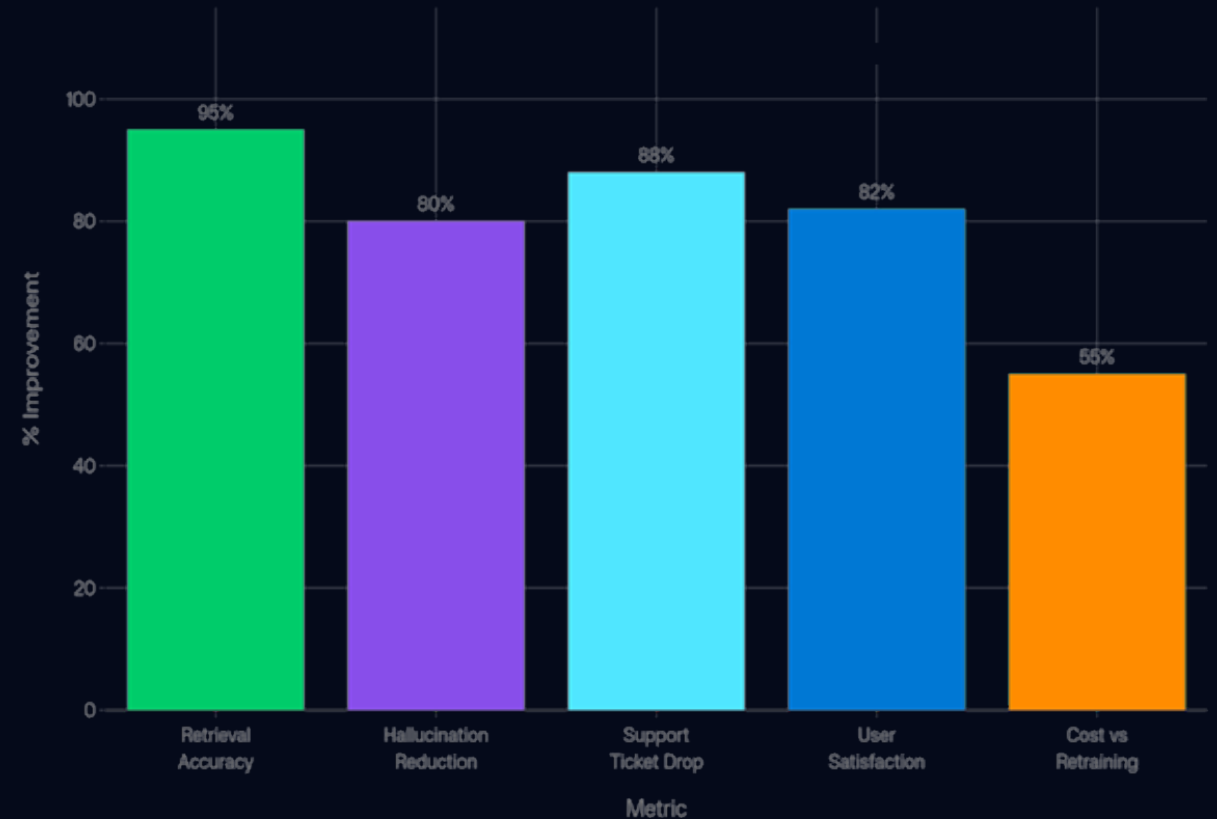
vs fine-tuning or retraining

What drives these outcomes:

Hybrid retrieval + reranking · Adaptive chunking · Prompt discipline · Right-sized model selection

Enterprise RAG: Production Performance Benchmarks

Source: Sprout-AI Production | Azure AI Foundry deployment



Scalability & Security at Enterprise Scale

The controls that transform RAG from a pilot into a production platform

Scalability Architecture

- Provisioned throughput + serverless elasticity for variable load
- High-throughput vector search — millions of docs, <100ms p95
- Multi-region & availability zone deployments
- Horizontal scale for 10,000+ concurrent users

10,000+

Concurrent users

<100ms

Vector search p95

Security Stack

- Private endpoints + managed VNET — zero public exposure
- Microsoft Entra ID authentication & authorization
- RBAC at resource & project level — least privilege
- BYOK encryption · audit logging · AI Red Teaming · content safety

BYOK

Encryption

Zero

Public Exposure

Best Practices for Enterprise RAG

What consistently improves accuracy, reliability & cost in production



Data Preparation

Semantic chunking · Metadata enrichment
Deduplication · Freshness scheduling



Retrieval Design

Hybrid search · Reranking · Permission filtering
Confidence-aware answers



Observability

Track precision · Latency · Hallucination rate
Index drift — continuously



Security

Disable public access · Filter by user permissions
BYOK · Regular red teaming



Cost Optimization

Smaller models for simpler tasks
Provisioned throughput · Event-driven scaling



Automation

Bicep IaC · azd CLI · GitHub Actions CI/CD
Repeatable evaluation pipelines

Emerging Trends & Azure AI Foundry Roadmap

Where enterprise RAG is heading — and how Azure AI Foundry supports it



Multimodal RAG

Retrieve across text, tables, images & diagrams.
GPT-4o vision extends grounding beyond text-only indexes.



Agentic RAG

Foundry Agent Service orchestrates multi-step retrieval with tool calls — the new standard for complex workflows.



Foundry IQ

Foundry IQ is a managed knowledge layer for agents that connects structured and unstructured data.



Efficient Models

Phi-4 and smaller domain models enable cost-conscious and edge/offline RAG scenarios.



Real-time Streaming

Event-driven index updates via Azure Functions & Event Hubs keep knowledge fresh, reducing stale answers.

Build · Deploy · Govern

Enterprise AI grounded in your real business knowledge.

Production-ready. Secure. Scalable. Measurable.

Azure AI Foundry

Enterprise RAG platform

Enterprise RAG

Scalable grounded AI

Databricks · Azure

Modern AI data stack

Thank You · CIRAS Iowa AI Summit 2026 · Mehul Bhuvra · [linkedin.com/in/mehulbhuvra](https://www.linkedin.com/in/mehulbhuvra)

DEMO - Azure AI Foundry

Q & A

LET'S CONNECT

 linkedin.com/in/mehulbhuva

 learn.microsoft.com/azure/ai-foundry

 github.com/azure-samples

 CIRAS Iowa AI Summit · May 6, 2026

Discussion Starters

- Where are you today — pilot, early prod, or scaling?
- What is the biggest RAG challenge in your org?
- Which industry use case resonates most?